The State of the Art Ten Years After a State of the Art

Future Research in Music Information Retrieval

Sturm, Bob L.

Published in:
Journal of New Music Research

DOI (link to publication from Publisher):
10.1080/09298215.2014.894533

Publication date:
2014

Document Version
Early version, also known as pre-print

The State of the Art Ten Years After a State of the Art: Future Research in Music Information Retrieval

Bob L. Sturm*

Abstract

A decade has passed since the first review of research on a "flagship application" of music information retrieval (MIR): the problem of music genre recognition (MGR). During this time, about 500 works addressing MGR have been published, and at least 10 campaigns have been run to evaluate MGR systems, which makes MGR one of the most researched areas of MIR. So, where does MGR lie now? We show that in spite of this massive amount of work, MGR does not lie far from where it began, and the paramount reason for this is that most evaluation in MGR lacks validity. We perform a case study of all published research using the most-used benchmark dataset in MGR during the past decade: *GTZAN*. We show that none of the evaluations in these many works is valid to produce conclusions with respect to *recognizing genre*, i.e., that a system is using criteria relevant to recognize genre. In fact, the problems of validity in evaluation also affect research in music emotion recognition and autotagging. We conclude by discussing the implications of our work for MGR and MIR in the next ten years.

1 Introduction

"Representing Musical Genre: A State of the Art" (Aucouturier and Pachet, 2003) was published a decade ago now, a few years after the problem of music genre recognition (MGR) was designated a "flagship application" of music information retrieval (MIR) (Aucouturier and Pampalk, 2008). During that time, there have been at least four other reviews of research on MGR (Scaringella et al., 2006; Dannenberg, 2010; Fu et al., 2011; Sturm, 2012b), nearly 500 published works considering MGR, and at least 10 organized campaigns to evaluate systems proposed for MGR: ISMIR 2004,[1] ISMIS 2011,[2] and MIREX 2005, 2007–13.[3] Figure 1 shows the massive amount of published work on MGR since that of Matityaho and Furst (1995).[4] So, where does MGR now lie? How much progress has been made?

One might find an indication by looking at how the construction and performance of systems designed for MGR have changed during the past decade. The reviews by Aucouturier and Pachet (2003), Scaringella et al. (2006) and Fu et al. (2011) all describe a variety of approaches to feature extraction and machine learning that have been explored for MGR, and provide rough comparisons of how MGR systems perform on benchmark datasets and in evaluation campaigns. Conclusions from these evaluations as a whole have been drawn about research progress on the problem of MGR. For instance, Bergstra et al. (2006) — authors of the MGR system having the highest accuracy in MIREX 2005 — write, "Given the

*Audio Analysis Lab, AD:MT, Aalborg University Copenhagen, A.C. Meyers Vænge 15, DK-2450 Copenhagen, Denmark, (+45) 99407633, e-mail: bst@create.aau.dk. BLS is supported in part by Independent Postdoc Grant 11-105218 from Det Frie Forskningsråd.

[1] `http://ismir2004.ismir.net/genre_contest/index.htm`

[2] `http://tunedit.org/challenge/music-retrieval`

[3] `http://www.music-ir.org/mirex/wiki/MIREX_HOME`

[4] The bibliography and spreadsheet that we use to generate this figure are available here: `http://imi.aau.dk/~bst/software`.

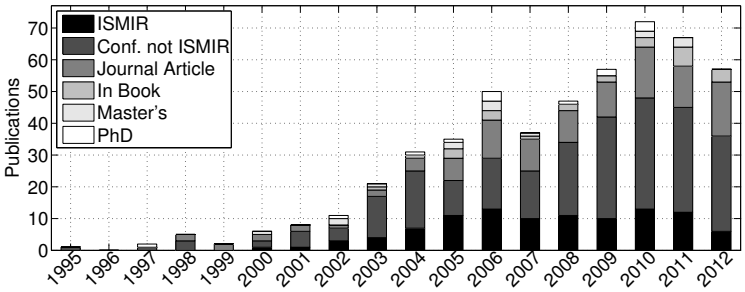

Figure 1: Annual numbers of published works in MGR with experimental components, divided into publication venues.

steady and significant improvement in classification performance ... we wonder if automatic methods are not already more efficient at learning genres than some people." In Fig. 2, we plot the highest reported classification accuracies of about 100 MGR systems all evaluated in the benchmark dataset *GTZAN* (Tzanetakis and Cook, 2002).[5] We see there to be many classification accuracies higher than the 61% reported by Tzanetakis and Cook (2002). The maximum of each year shows progress up to 2009, which then appears to stop. Similarly, Humphrey et al. (2013) look at the best accuracies in the MGR task of MIREX from 2007 to 2012 and suggest that progress "is decelerating, if not altogether stalled."

In spite of all these observations, however, might it be that the apparent progress in MGR is an illusion? Our exhaustive survey (2012b) of nearly 500 publications about MGR shows: of ten evaluation designs used in MGR, one in particular (which we call *Classify*) appears in 91% of work having an experimental component; and the most-used public dataset is *GTZAN*, appearing in the evaluations of about 100 published works (23%). These findings are not promising. First, *Classify* using a dataset having independent variables that are not controlled cannot provide *any* valid evidence to conclude upon the extent to which an MGR system is *recognizing* the genres used by music (2013a; 2013b). In other words, just because an MGR system reproduces all labels of a dataset does not then mean it is making decisions by using criteria relevant to genre (e.g., instrumentation, composition, subject matter). In fact, we have clearly shown (2012c; 2013g) that an MGR system can produce a high accuracy using confounded factors; and when these confounds break, its performance plummets. Second, we have found (2012a; 2013d) *GTZAN* has several faults, namely, repetitions, mislabelings, and distortions. Because all points in Fig. 2 use *Classify* in *GTZAN*, their meaningfulness is thus "doubly questionable." What valid conclusions, then, can one draw from all this work?

Our "state of the art" here attempts to serve a different function than all previous reviews of MGR (Aucouturier and Pachet, 2003; Scaringella et al., 2006; Dannenberg, 2010; Fu et al., 2011). First, it aims not to summarize the variety of features and machine learning approaches used in MGR systems over the past ten years, but to look closely at how to interpret Fig. 2. Indeed, we have a best case scenario for drawing conclusions: these evaluations of many different systems use the same same kind of evaluation (*Classify*) *and* the same dataset (*GTZAN*). Unfortunately, we find the faults in *GTZAN* impose an insurmountable impediment to interpreting Fig. 2. It is tempting to think that since each of these MGR systems faces the same faults in *GTZAN*, and that its faults are exemplary of

[5]The dataset can be downloaded from here: `http://marsyas.info/download/data_sets`

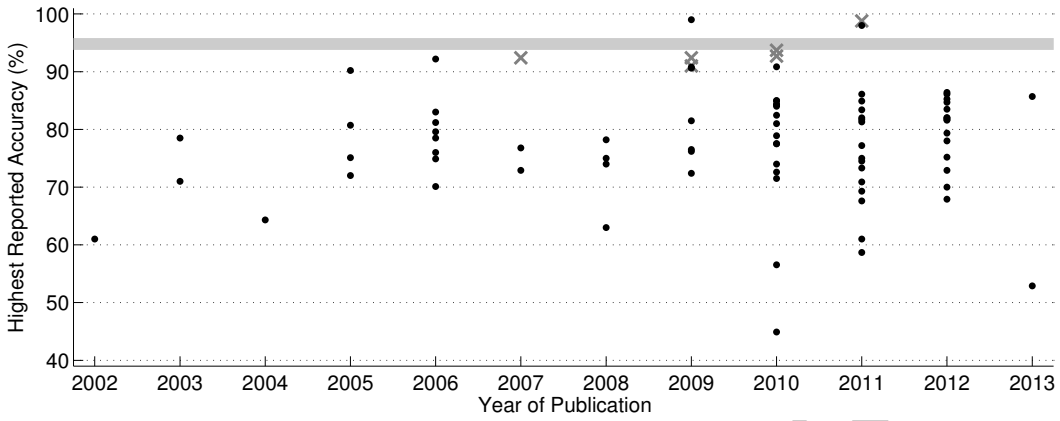

Figure 2: Highest classification accuracies (%) using all *GTZAN* as a function of publication year. Solid gray line is our estimate of the "perfect" accuracy in Table 2. Six "×" denote incorrect results, discussed in Section 4.5.

real world data, then the results in Fig. 2 are still meaningful for comparing systems. We show this argument to be wrong: the faults simply do not affect the performance of MGR systems in the same ways. Second, this article aims not to focus on MGR, but to look more broadly at how the practice of evaluation in the past decade of MGR research can inform the practice of evaluation in the next decade of MIR research. An obvious lesson from *GTZAN* is that a researcher must know their data, know real data has faults, and know faults have real impacts on evaluation; but, we also distill five other important "guidelines": 1) Define problems with use cases and formalism; 2) Design valid and relevant experiments; 3) Perform deep system analysis; 4) Acknowledge limitations and proceed with skepticism; and 5) Make reproducible work reproducible.

In the next section, we briefly review "the problem of music genre recognition." The third section provides a comprehensive survey of how *GTZAN* has been and is being used in MGR research. In the fourth section, we analyze *GTZAN*, and identify several of its faults. In the fifth section, we test the real effects of its faults on the evaluation of several categorically different MGR systems. Finally, we conclude with a discussion of the implications of this work on future work in MGR, and in MIR more broadly.

2 The Problem of Music Genre Recognition (in brief)

It is rare to find in any reference of our survey (2012b) a formal or explicit definition of MGR; and few works explicitly define "genre," instead deferring to describing why genre is useful, e.g., to explore music collections (Aucouturier and Pachet, 2003). Aucouturier and Pachet (2003) describe MGR as "extracting genre information automatically from the audio signal." Scaringella et al. (2006) and Tzanetakis and Cook (2002) both mention automatically arranging music titles in genre taxonomies. Since 91% of MGR evaluation employs *Classify* in datasets with uncontrolled independent variables (Sturm, 2012b), the majority of MGR research implicitly interprets MGR as reproducing *by any means possible* the "ground truth" genre labels of a music dataset (Sturm, 2013b). Almost all work in our survey (2012b) thus treats genre in an *Aristotelean* way, assuming music *belongs* to categories like a specimen belongs to a species, which belongs to a genus, and so on.

So, what is genre? The work of Fabbri (1980, 1999) — dealt with in depth by Kemp (2004), and cited by only a few MGR works, e.g., McKay and Fujinaga (2006) and Craft (2007) — essentially conceives of music genre as "a set of musical events (real or possible) whose course is governed by a definite set of socially accepted rules" (1980). This "course" applies to the "musical events (real or possible)", where a "musical event" Fabbri defines as being a "type of activity ... involving sound" (1980). Adopting the language of set theory, Fabbri speaks of "genres" and "sub-genres" as unions and subsets of genres, each having well-defined boundaries, at least for some "community." He goes as far to suggest one can construct a matrix of rules crossed with subgenres of a genre, with each entry showing the applicability of a particular rule for a particular subgenre. By consequence, producing and using such a master list amounts to nothing more than Aristotelean categorization: to which set does a piece of music belong?

A view of genre alternative to that of Fabbri is provided by Frow (2005), even though his perspective is of literature and not music. Frow argues that "genre" is a dynamic collection of rules and criteria specifying how a person approaches, interprets, describes, uses, judges, and so on, forms of human communication *in* particular contexts. These rules and criteria consequently precipitate from human communication as an aid for people to interact with information and with each other in the world.[6] For instance, genre helps a person to read, use, and judge a published work as a scientific article, as opposed to a newspaper column. Genre helps an author write and sell a written work as a publish-worthy scientific article, as opposed to a newspaper column. This is not just to say that the "medium is the message" (McLuhan, 1964), but also that a person reading, using, and judging a published scientific journal article is using cues — both intrinsic and extrinsic — to justify their particular treatment of it as a scientific journal article, and not as a newspaper column (although it could be read as a newspaper column in a different context). Frow's view of genre, then, suggests that music does not *belong* to genre, but that the human creation and consumption of music in particular contexts necessarily *use* genre. Hence, instead of "to which genre does a piece of music belong," the meaningful question for Frow is, "what genres should I use to interpret, listen to, and describe a piece of music in a particular context?" In Frow's conception of genre, it becomes clear why any attempt at building a taxonomy or a list of characteristics for categorizing music into musical genres will fail: they lack the human *and* the context, both necessary aspects of genre.

Essentially unchallenged in MGR is the base assumption that the nature of "genre" is such that music *belongs to* categories, and that it makes sense to talk about "boundaries" between these categories. Perhaps it is a testament to the pervasiveness of the Aristotelean categorization of the world (Bowker and Star, 1999) that alternative conceptions of music genre like Frow's have by and large gone unnoticed in MGR. Of course, some have argued that MGR is ill-defined (Pachet and Cazaly, 2000; Aucouturier and Pachet, 2003; McKay and Fujinaga, 2006), that some genre categories are arbitrary and unique to each person (Craft, 2007; Sordo et al., 2008), that they are motivated by industry (Aucouturier and Pachet, 2003), and/or they come in large part from domains outside the purview of signal processing (Craft, 2007; Wiggins, 2009). Processing only sampled music signals necessarily ignores the "extrinsic" properties that contribute to judgements of genre (Aucouturier and Pachet, 2003;

[6]Fabbri (1999) also notes that genre helps "to speed up communication."

McKay and Fujinaga, 2006; Wiggins, 2009). Hence, it is of little surprise when subjectivity and commercial motivations of genre necessarily wreak havoc for any MGR system.

The value of MGR has also been debated. Some have argued that the value of producing genre labels for music already labeled by record companies is limited. Some have suggested that MGR is instead encompassed by other pursuits, e.g., music similarity (Pampalk, 2006), music autotagging (Aucouturier and Pampalk, 2008; Fu et al., 2011), or extracting "musically relevant data" (Serra et al., 2013). Others have argued that, in spite of its subjective nature, there is evidence that genre is not so arbitrary (Lippens et al., 2004; Gjerdingen and Perrott, 2008; Sordo et al., 2008), that some genres can be defined by specific criteria (Barbedo and Lopes, 2007), and that research on MGR is still a worthy pursuit (McKay and Fujinaga, 2006; Scaringella et al., 2006). Many works also suggest that MGR provides a convenient testbed for comparing new audio features and/or machine learning approaches, e.g., Andén and Mallat (2011); Andén and Mallat (2013).

While most of the work we survey (2012b) implicitly poses MGR as an Aristotelean categorization, the goal posed by Matityaho and Furst (1995) comes closer to the conception of Frow, and to one that is much more useful than the reproduction of "ground truth" genre labels: "[building] a phenomenological model that [imitates] the human ability to distinguish between music [genres]." Although it is missing the context essential to genre for Frow, and although the empirical work of Matityaho and Furst (1995) still treats genre in an Aristotelean manner, this goal places humans at the center, establishes the goal as *imitation*, and prescribes the use of humans for evaluation. This has led us to define the "principal goals of MGR" (2013b): *to imitate the human ability to organize, recognize, distinguish between, and imitate genres used by music.* "Organization" implies finding and expressing characteristics used by particular genres, e.g., "blues is strophic"; "recognition" implies identification of genres, e.g., "that sounds like blues because ..."; "distinguishing" implies describing why, or the extents to which, some music uses some genres but not others, e.g., "that does not sound like blues because ..."; and "imitation" implies being able to exemplify the use of particular genres, e.g., "play the Bach piece as if it is blues." Stated in such a way, MGR becomes much richer than generating genre-indicative labels for music signals.

3 A Survey and Analysis of *GTZAN*

In this section, we look closely at *GTZAN* since its creation a decade ago: how it has been used, of what it is composed, what its faults are, how its faults affect evaluation, and what this implies for MGR in the past decade, and *GTZAN* in the next decade.

3.1 How has *GTZAN* been used?

Figure 3 shows how the number of publications that use *GTZAN* has increased since its creation in 2002. We see that it has been used more in the past three years than in its first eight years. The next most-used public dataset is *ISMIR2004*,[7] which was created for the MGR task of ISMIR 2004. That dataset appears in 75 works, 30 of which use *GTZAN* as well. Of the 100 works that use *GTZAN*, 49 of them use only *GTZAN*.[8]

[7] `http://kom.aau.dk/~jhj/files/ismir2004genre/`

[8] All relevant references are available at: `http://imi.aau.dk/~bst/software`.

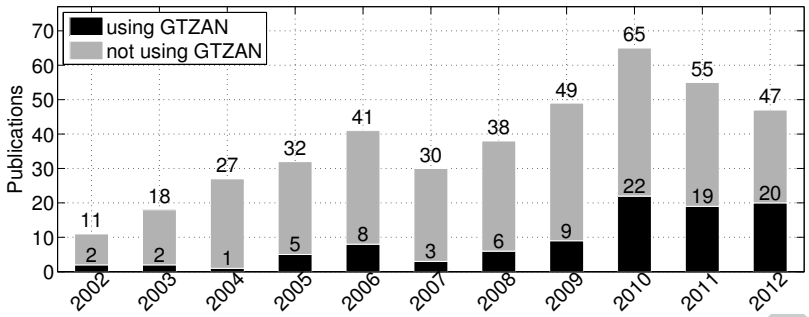

Figure 3: Annual numbers of published works in MGR with experimental components, divided into ones that use and do no use *GTZAN*.

In our review of evaluation in MGR (2012b), we delimit ten different evaluation designs. Of the 100 works using *GTZAN*, 96 employ the evaluation design *Classify* (an excerpt is assigned labels, which are compared against a "ground truth"). In seven works, *GTZAN* is used with the evaluation design *Retrieve* (a query is used to find similar music, and the labels of the retrieved items are compared); and one work (Barreira et al., 2011) uses *GTZAN* with the evaluation design *Cluster* (clustering of dataset and then a comparison of the labels in the resulting clusters). Our work (2012c) uses *Compose*, where an MGR system generates new music it scores as highly representative of each *GTZAN* category, which we then test for identifiability using a formal listening test. We find a few other uses of *GTZAN*. Markov and Matsui (2012a,b) learn bases from *GTZAN*, and then apply these codebooks to classify the genres of *ISMIR2004*. In our work (2013e), we train classifiers using *ISMIR2004*, and then attempt to detect all excerpts in *GTZAN* Classical (and two in *GTZAN* Jazz). As described above, Fig. 2 shows the highest classification accuracies reported in the papers that consider the 10-class problem of *GTZAN* (we remove duplicated experiments, e.g., Lidy (2006) contains the results reported in Lidy and Rauber (2005)).

Among the 100 published works using *GTZAN*, we find only five outside ours (2012c; 2013a; 2013f; 2013e) that indicate someone has listened to at least some of *GTZAN*. The first appears to be Li and Sleep (2005), who find that "two closely numbered files in each genre tend to sound similar than the files numbered far [apart]." Bergstra et al. (2006) note that, "To our ears, the examples are well-labeled ... our impression from listening to the music is that no artist appears twice." This is contradicted by Seyerlehner et al. (2010), who predict "an artist effect ... as listening to some of the songs reveals that some artists are represented with several songs." In his doctoral dissertation, Seyerlehner (2010) infers there to be a significant replication of artists in *GTZAN* because of how classifiers trained and tested with *GTZAN* perform as compared to other artist-filtered datasets. Furthermore, very few people have mentioned specific faults in *GTZAN*: Hartmann (2011) notes finding seven duplicates; and Li and Chan (2011) — who have manually estimated keys for all *GTZAN* excerpts[9] — remember hearing some repetitions.[10] Hence, it appears that *GTZAN* has by and large been assumed to have satisfactory integrity for MGR evaluation.

[9]Available here: `http://visal.cs.cityu.edu.hk/downloads/#gtzankeys`

[10]Personal communication.

Label	ENMFP	by self	metadata from `last.fm` # songs (# tags)	# artists (# tags)
Blues	63	100	75 (2904)	25 (2061)
Classical	63	97	12 (400)	85 (4044)
Country	54	95	81 (2475)	13 (585)
Disco	52	90	82 (5041)	8 (194)
Hip hop	64	96	90 (6289)	5 (263)
Jazz	65	80	54 (1127)	26 (2102)
Metal	65	83	73 (5579)	10 (786)
Pop	59	96	89 (7173)	7 (665)
Reggae	54	82	73 (4352)	9 (616)
Rock	67	100	99 (7227)	1 (100)
Total	60.6%	91.9%	72.8% (42567)	18.9% (11416)

Table 1: For each category of *GTZAN*: number of excerpts we identify by fingerprint (ENMFP); then searching manually (by self); number of songs tagged in `last.fm` (and number of those tags having "count" larger than 0); for songs not found, number of artists tagged in `last.fm` (and number of tags having "count" larger than 0). Retrieved Dec. 25, 2012, 21h.

3.2 What is in *GTZAN*?

Until our work (2012a), *GTZAN* has never had metadata identifying its contents because those details were not assembled during compilation. Hence, every MGR evaluation using *GTZAN*, save ours (2012c; 2013a; 2013e; 2013d), has not been able to take into account its contents. We now identify the excerpts in *GTZAN*, determine how music by specific artists compose each category, and survey the tags people have applied to the music and/or artist in order to obtain an idea of what each *GTZAN* category means.

We use the Echo Nest Musical Fingerprinter (ENMFP)[11] to generate a fingerprint of an excerpt in *GTZAN* and then to query the Echo Nest database having over 30,000,000 songs (at the time of this writing). The second column of Table 1 shows that this identifies only 606 of the 1000 excerpts. We manually correct titles and artists as much as possible, e.g., we reduce "River Rat Jimmy (Album Version)" to "River Rat Jimmy"; and "Bach - The #1 Bach Album (Disc 2) - 13 - Ich steh mit einem Fuss im Grabe, BWV 156 Sinfonia" to "Ich steh mit einem Fuss im Grabe, BWV 156 Sinfonia;" and we correct "Leonard Bernstein [Piano], Rhapsody in Blue" to "George Gershwin" and "Rhapsody in Blue." We find four misidentifications: Country 15[12] is misidentified as being by Waylon Jennings (it is by George Jones); Pop 65 is misidentified as being Mariah Carey (it is Prince); Disco 79 is misidentified as "Love Games" by Gazeebo (it is "Love Is Just The Game" by Peter Brown); and Metal 39 is Metallica playing "Star Wars Imperial March," but ENMFP identifies it as a track on a CD for improving sleep.[13] We then manually identify 313 more excerpts, but have yet to identify the remaining 81 excerpts.[14]

Figure 4 shows how each *GTZAN* category is composed of music by particular artists. We see only nine artists are represented in the *GTZAN* Blues. *GTZAN* Reggae is the category

[11] `http://developer.echonest.com`

[12] This is the file "country.00015.wav" in *GTZAN*.

[13] "Power Nap" by J. S. Epperson (Binaural Beats Entrainment), 2010.

[14] Our machine-readable index of this metadata is available at: `http://imi.aau.dk/~bst/software`.

with the most excerpts from a single artist: 35 excerpts of Bob Marley. The category with the most artist diversity appears to be Disco, where we find at least 55 different artists. From this, we can bound the number of artists in *GTZAN*, which has until this time been unknown (Seyerlehner, 2010; Seyerlehner et al., 2010). Assuming each unidentified excerpt comes from different artists than those we have already identified, the total number of artists represented in *GTZAN* cannot be larger than 329. If all unlabeled excerpts are from the artists we have already identified, then the smallest this number can be is 248.

We now wish to determine the *content* composing each *GTZAN* category. In our previous analysis of *GTZAN* (2012a), we assume that since there is a category called, e.g., "Country," then *GTZAN* Country excerpts should possess typical and distinguishing characteristics of music using the country genre (Ammer, 2004): stringed instruments such as guitar, mandolin, banjo; emphasized "twang" in playing and singing; lyrics about patriotism, hard work and hard times; and so on. This led us to the claim that at least seven *GTZAN* Country excerpts are mislabeled because they exemplify few of these characteristics. We find other work that assumes the genres of music datasets overlap because they share the same genre labels, e.g., the taxonomies of Moerchen et al. (2006) and Guaus (2009). In this work, however, we do not make the assumption that the excerpts in *GTZAN* Country possess the typical and distinguishing characteristics of music using the country genre. In other words, we now consider a *GTZAN* category name as "short hand" for the collection of consistent and/or contradictory concepts and criteria, both objective and subjective, that Tzanetakis employed in assembling music excerpts to form that *GTZAN* category.

To obtain an idea of the content composing each *GTZAN* category, we query the application programming interface provided by `last.fm`, and retrieve the "tags" that users of the service (whom we will call "taggers") have entered for each song or artist we identify in *GTZAN*. A tag is a word or phrase a tagger associates with an artist, song, album, and so on, for any number of reasons (Bertin-Mahieux et al., 2010), e.g., to make a music collection more useful to themselves, to help others discover new music, or to promote their own music or criticize music they do not like. A past analysis of the `last.fm` tags (Bertin-Mahieux et al., 2008) finds that they are most often genre labels (e.g., "blues"), but they can also be instrumentation ("female vocalists"), tempo (e.g., "120 bpm"), mood (e.g., "happy"), how they use the music (e.g., "exercise"), lyrics (e.g., "fa la la la la"), the band (e.g., "The Rolling Stones"), or something else (e.g., "favorite song of all time") (Law, 2011).

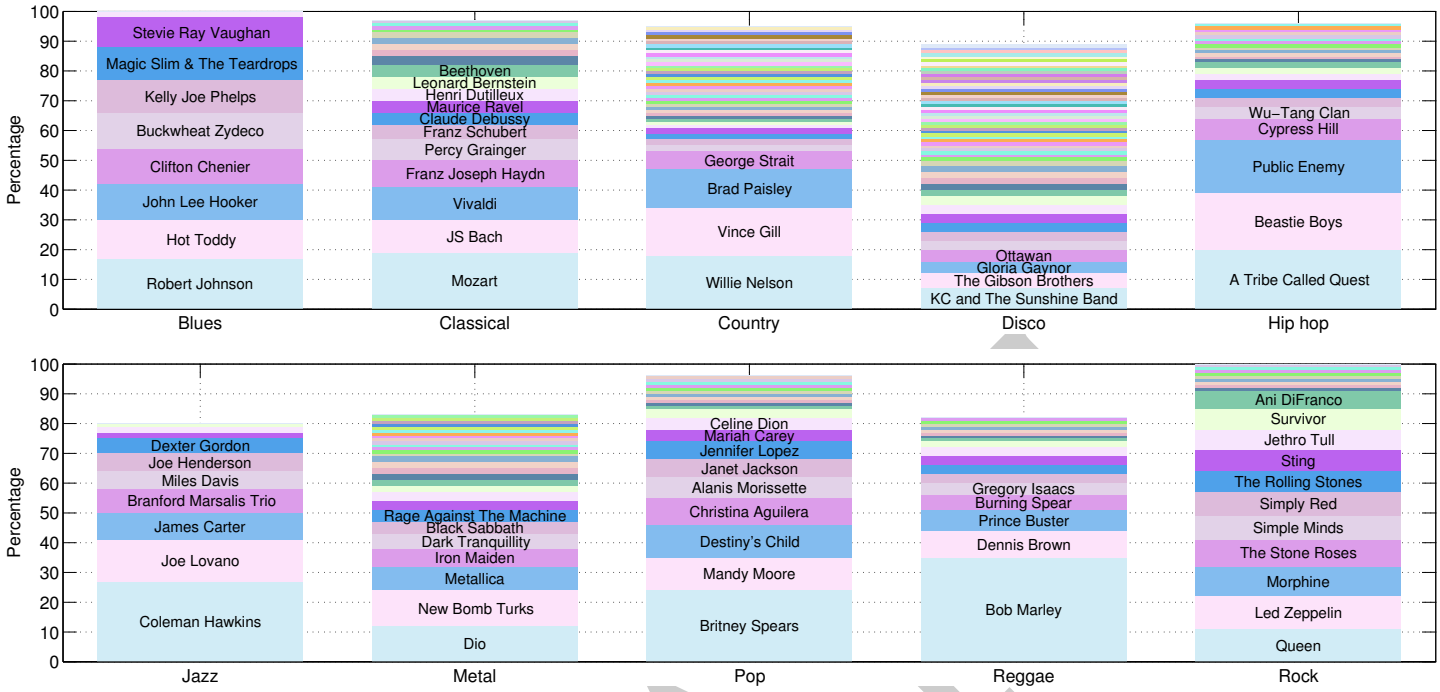

Figure 4: Artist composition of each *GTZAN* category. We do not include unidentified excerpts.

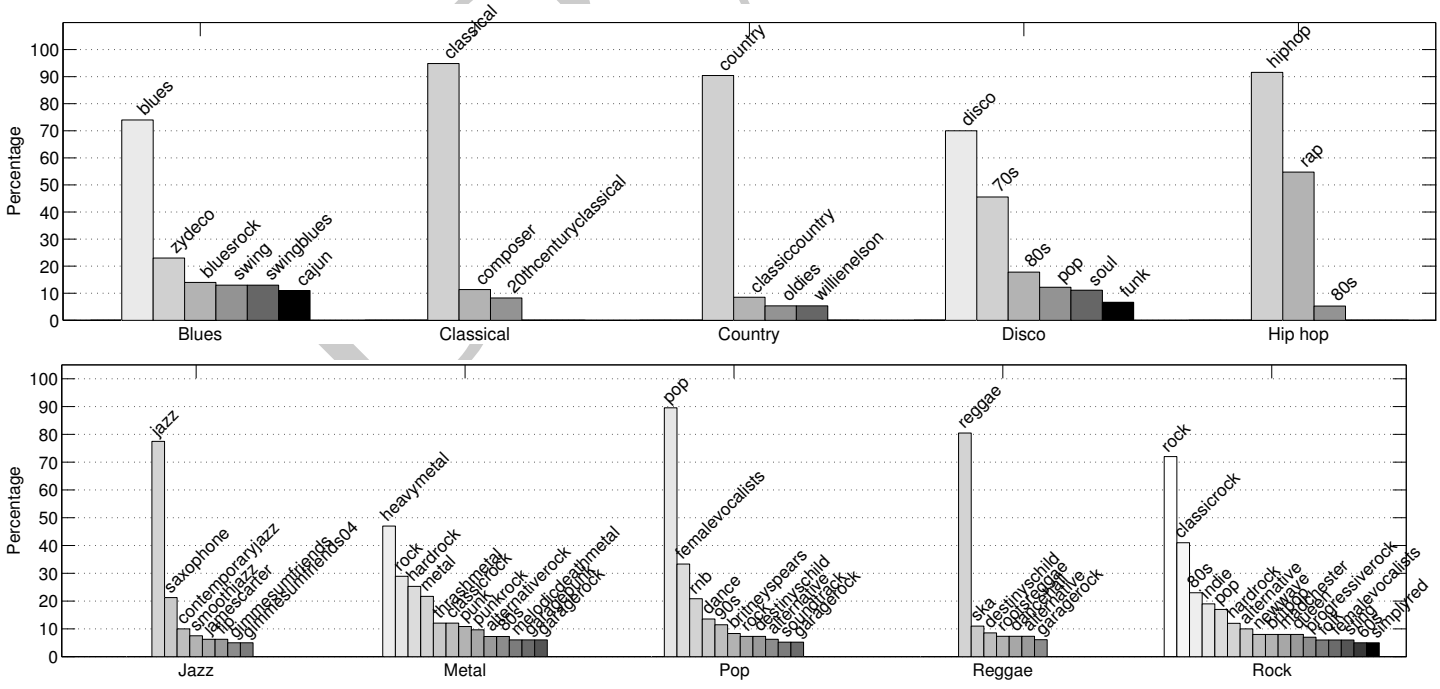

Figure 5: Category top tags for *GTZAN* from `last.fm`.

The collection of tags by `last.fm` is far from being a controlled process. Any given tagger is not necessarily well-versed in musicology, or knows the history of musical styles, or can correctly recognize particular instruments, or is even acting in a benevolent manner; and any group of taggers is not necessarily using the same criteria when they all decide on tagging a song with a particular tag appearing indicative of, e.g., genre or emotion. For these reasons, there is apprehension in using such a resource for music information research (Aucouturier, 2009). However, `last.fm` tags number in the millions and come from tens of thousands of users; and, furthermore, each tag is accompanied by a "count" parameter reflecting the percentage of taggers of a song or artist that choose that particular tag (Levy and Sandler, 2009). A tag for a song having a count of 100 means that tag is selected by all taggers of that song (even if there is only one tagger), and 0 means the tag is applied by the fewest (`last.fm` rounds down all percentages less than 1). We cannot assume that each tag selected by a tagger for a given song is done independent of those given by previous taggers, but it is not unreasonable to interpret a tag with a high count as suggesting a kind of consensus that `last.fm` taggers find the tag very relevant for that song, whatever that may mean. Though `last.fm` tags have found use in other MIR research (Bertin-Mahieux et al., 2008; Barrington et al., 2008; Levy and Sandler, 2009), we proceed cautiously about interpreting the meanings behind the `last.fm` tags retrieved for the identified contents in *GTZAN*. With respect to our goals here, we assume some tags are meant by taggers to be descriptive of a song or artist.

Using the index of *GTZAN* we create above, the fourth column of Table 1 shows the number of songs in *GTZAN* we identify that have `last.fm` tags, and the number of tags with non-zero count (we keep only those tags with counts greater than 0). When we do not find tags for a song, we request instead the tags for the artist. For instance, though we identify all 100 excerpts in Blues, only 75 of the songs are tagged on `last.fm`. Of these, we get 2,904 tags with non-zero counts. For the remaining 25 songs, we retrieve 2,061 tags from those given to the artists. We thus assume the tags that taggers give to an artist would also be given to the particular song. Furthermore, since each *GTZAN* excerpt is a 30 s excerpt of a song, we assume the tags given the song or artist would also be given to the excerpt.

With our considerations and reservations clearly stated, we now attempt to shed light on the content of each *GTZAN* category using the `last.fm` tags. First, we define the *top tags* of a song or artist as those that have counts above 50. To do this, we remove spaces, hyphens, and capitalization from all tags. For example, the tags "Hip hop", "hip-hop" and "hip hop" all become "hiphop." Then, we find for each unique top tag in a *GTZAN* category the percentage of identified excerpts in that category having that top tag. We define *category top tags* all those unique top tags of a *GTZAN* category that are represented by at least 5% of the excerpts identified in that category. With the numbers of excerpts we have identified in each category, this means a category top tag is one appearing as a top tag in at least 4 excerpts of a *GTZAN* category.

Figure 5 shows the resulting category top tags for *GTZAN*. We can see large differences in the numbers of category top tags, where *GTZAN* Rock has the most, and *GTZAN* Classical and *GTZAN* Hiphop have the least. As seen in Table 1, most of the tags for excerpts in *GTZAN* Classical come from those that taggers have entered for artists, which explains why the tag "composer" appears. Most of the category top tags appear indicative of genre, and those appearing most in each *GTZAN* category is the category label, except for Metal. We

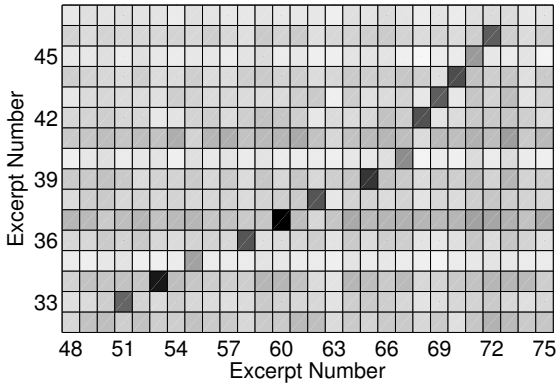

Figure 6: Taken from *GTZAN* Jazz, exact repetitions appear clearly with pair-wise comparisons of their fingerprint hashes. The darker a square, the higher the number of matching hashes.

also see signs of a specific tagger in three category top tags of *GTZAN* Jazz. With our considerations above, and by listening to entire dataset, it does not seem unreasonable to make some claims about the content of each *GTZAN* category. For instance, the category top tags of *GTZAN* Blues reveal its contents to include a large amount of music tagged "cajun" and "zydeco" by a majority of taggers. *GTZAN* Disco appears more broad than dance music from the late seventies (Shapiro, 2005), but also includes a significant amount of music that has been tagged "80s", "pop" and "funk" by a majority of taggers. Listening to the excerpts in these categories confirms these observations.

3.3 What faults does *GTZAN* have?

We now delimit three kinds of faults in *GTZAN*: repetitions, mislabelings, and distortions. Table 2 summarizes these, which we reproduce online with sound examples.[15]

3.3.1 Repetitions

We consider four kinds of repetition, from high to low specificity: exact, recording, artist, and version. We define an *exact repetition* as when two excerpts are the same to such a degree that their time-frequency fingerprints are the same. This means the excerpts are not only extracted from the same recording, they are essentially the same excerpt, either sample for sample (up to a multiplicative factor), or displaced in time by only a small amount. To find exact repetitions, we implement a simplified version of the Shazam fingerprint (Wang, 2003). This means we compute sets of anchors in the time-frequency plane, compare hashes of the anchors of every pair of excerpts, and then listen to those excerpts sharing many of the same anchors to confirm them to be exact replicas. Figure 6 shows the clear appearance of exact repetitions in *GTZAN* Jazz. The second column of Table 2 lists these. Our comparison of hashes across categories reveal one exact repetition in two categories: the same excerpt of "Tie Your Mother Down" by Queen appears as Rock 58 and Metal 16. In total, we find 50 exact repetitions in *GTZAN*.

[15] `http://imi.aau.dk/~bst/research/GTZANtable2`

We define a *recording repetition* as when two excerpts come from the same recording, but are not detected with the fingerprint comparison detailed above. We find these by artist name and song repetitions in the index we create above, and by listening. For instance, Country 8 and 51 are both from "Never Knew Lonely" by Vince Gill, but excerpt 8 comes from later in the recording than excerpt 51. The second and third columns of Table 2 shows the excerpts we suspect as coming from the same recording. We find *GTZAN* Pop has the most exact and suspected recording repetitions (16): "Lady Marmalade" sung by Christina Aguilera et al., as well as "Bootylicious" by Destiny's Child, each appear four times. In total, we find 21 suspected recording repetitions in *GTZAN*.

The last two kinds of repetitions are not necessarily "faults", but we show in Section 3.4 why they must be taken into consideration when using *GTZAN*. We define *artist repetition* as excerpts performed by the same artist. We find these easily using the index we create above. Figure 4 and Table 2 show how every *GTZAN* category has artist repetition. Finally, we define a *version repetition* as when two excerpts are of the same song but performed differently. This could be a studio version, a live version, performed by the same or different artists (covers), or possibly a remix. We identify these with the index we create above, and then confirm by listening. For instance, Classical 44 and 48 are from "Rhapsody in Blue" by George Gershwin, but presumably performed by different orchestras. Metal 33 is "Enter Sandman" by Metallica, and Metal 74 is a parody of it. Pop 16 and 17 are both "I can't get no satisfaction" by Britney Spears, but the latter is performed live. In total, we find 13 version repetitions in *GTZAN*.

3.3.2 Potential mislabelings

The collection of concepts and criteria Tzanetakis used to assemble *GTZAN* is of course unobservable; and in some sense, the use of *GTZAN* for training and testing an MGR system aims to reproduce or uncover it by reverse engineering. Regardless of whether *GTZAN* was constructed in a well-definable manner, or whether it makes sense to restrict the membership of an excerpt of music to one *GTZAN* category, we are interested here in a different question: are any excerpts mislabeled? In other words, we wish to determine whether the excerpts might be arranged in a way such that their membership to one *GTZAN* category, as well as their exclusion from every other *GTZAN* category, is in some sense "less debatable" to a system, human or artificial?

Toward this end, we previously (2012a) considered two kinds of mislabelings in *GTZAN*: "contentious" and "conspicuous." We based these upon non-concrete criteria formed loosely around musicological principles associated with the names of the categories in *GTZAN*. Now that we have shown in Section 3.2 that those names do not necessarily reflect the content of the categories, we instead consider here the content of each *GTZAN* category. In summary, we identify excerpts that might be "better" placed in another category, placed across several categories, or excluded from the dataset altogether, by comparing their tags to the category top tags of each *GTZAN* category.

We consider an excerpt *potentially mislabeled* if not one of its tags match the category top tags of its category, or if the "number of votes" for its own category is equal to or less than the "number of votes" for another category. We define the *number of votes* in a category as the sum of the counts of tags for an excerpt matching category top tags. (We consider a match

as when the category top tag appears in a tag, e.g., "blues" appears in "blues guitar".) As an example, consider the category top tags of *GTZAN* Country in Fig. 5, and the pairs of tags and counts we find for Country 39 ("Johnnie Can't Dance" by Wayne Toups & Zydecajun): {("zydeco", 100), ("cajun", 50), ("folk", 33), ("louisiana", 33), ("bayou", 16), ("cruise", 16), ("Accordeon", 16), ("New Orleans", 16), ("accordion", 16), ("everything", 16), ("boogie", 16), ("dance", 16), ("swamp", 16), ("country", 16), ("french", 16), ("rock", 16)}. We find its two first tags among the category top tags of *GTZAN* Blues, and so the number of votes for *GTZAN* Blues is, $100 + 50 = 150$. We only find one of its tags ("country") among the category top tags of *GTZAN* Country, and so the number of votes there is 16. Hence, we argue that Country 39 is potentially mislabeled. Listening to the excerpt in relation to all the others in *GTZAN* Country, as well as the music tagged "cajun" and "zydeco" in *GTZAN* Blues, also supports this conclusion. It is important to emphasize that we are not claiming Country 39 *should* be labeled *GTZAN* Blues, but only that if a MGR system learning from *GTZAN* Blues labels Country 39 as "blues", then that might not be considered a mistake.

Of the excerpts we have identified in *GTZAN*, we find 13 with no tags that match the category top tags of their category, 43 that have more votes in the category top tags of a different category, and 18 that have the same number of votes in its own category and at least one other. Of these, 20 appear to be due to a lack of tags, and two appears to be a result of bad tags. For instance, among the tags with count of 100 for "Can't Do Nuttin' For Ya, Man!" by Public Enemy (Hip hop 93) are, "glam rock", "Symphonic Rock", "instrumental rock", "New York Punk" and "Progressive rock." We list the remaining 52 potential misclassifications in Table 2.

As we have pointed out (2012a; 2012c), *GTZAN* has other potential problems with excerpts in its categories. Disco 47 is of Barbra Streisand and Donna Summer singing the introduction to "No More Tears", but an excerpt from a later point this song might better exemplify *GTZAN* Disco. Hip hop 44 is of "Guantanamera" by Wyclef Jean, but the majority of the excerpt is a sample of musicians playing the traditional Cuban song "Guantanamera." Hence, *GTZAN* Hip hop might or might not be appropriate. Hip hop 5 and 30 are Drum and Bass dance music that are quite unique among the rest of the excerpts; and Reggae 51 and 55 are electronic dance music that are again quite contrasting to the other excerpts. Finally, Reggae 73 and 74 are exact replicas of "Hip-Hopera" by Bounty Killer, which might be better more appropriate in *GTZAN* Hip hop. We do not include these as potential mislabelings in Table 2.

3.3.3 Identifying faults: Distortions

The last column of Table 2 lists some distortions we find by listening to every excerpt in *GTZAN*. This dataset was purposely created to have a variety of fidelities in the excerpts (Tzanetakis and Cook, 2002); however, one of the excerpts (Reggae 86) is so severely distorted that the value of its last 25 seconds is debatable.

GTZAN **Category**	**Repetitions** *Exact*	*Recording*	*Artist*	*Version*	**Potential Mislabelings**	**Distortions**
Blues			John Lee Hooker (0-11); Robert Johnson (12-28); Kelly Joe Phelps (29-39); Stevie Ray Vaughn (40-49); Magic Slim (50-60); Clifton Chenier (61-72); Buckwheat Zydeco (73-84); Hot Toddy (85-97); Albert Collins (98, 99)			
Classical		(42,53) (51,80)	J. S. Bach (00-10); Mozart (11-29); Debussy (30-33); Ravel (34-37); Dutilleux (38-41); Schubert (63-67); Haydn (68-76); Grainger (82-88); Vivaldi (89-99); and others	(44,48)		static (49)
Country		(08,51) (52,60)	Willie Nelson (19,26,65-80); Vince Gill (50-64); Brad Paisley (81-93); George Strait (94-99); and others	(46,47)	Ray Peterson "Tell Laura I Love Her" (20); Burt Bacharach "Raindrops Keep Falling on my Head" (21); Karl Denver "Love Me With All Your Heart" (22); Wayne Toups & Zydecajun "Johnnie Can't Dance" (39); Johnny Preston "Running Bear" (48)	static distortion (2)
Disco	(50,51,70) (55,60,89) (71,74) (98,99)	(38,78)	Gloria Gaynor (1,21, 38,78); Ottawan (17,24,44,45); The Gibson Brothers (22,28,30,35,37); KC and The Sunshine Band (49-51,70,71,73,74); ABBA (67,68,72); and others	(66,69)	Boz Scaggs "Lowdown" (00); Cheryl Lynn "Encore" (11); The Sugarhill Gang "Rapper's Delight" (27); Evelyn Thomas "Heartless" (29); Barbra Streisand and Donna Summer "No More Tears (Enough Is Enough)" (47); Tom Tom Club "Wordy Rappinghood" (85); Blondie "Heart Of Glass" (92); Bronski Beat "WHY?" (94)	clipping distortion (63)
Hip hop	(39,45) (76,78)	(01,42) (46,65) (47,67) (48,68) (49,69) (50,72)	Wu-Tang Clan (1,7,41,42); Beastie Boys (8-25); A Tribe Called Quest (46-51,62-75); Cypress Hill (55-61); Public Enemy (81-98); and others	(02,32)	3LW "No More (Baby I'ma Do Right)" (26); Aaliyah "Try Again" (29); Pink "Can't Take Me Home" (31); Lauryn Hill "Ex-Factor" (40)	clipping distortion (3,5); skip at start (38)
Jazz	(33,51) (34,53) (35,55) (36,58) (37,60) (38,62) (39,65) (40,67) (42,68) (43,69) (44,70) (45,71) (46,72)		James Carter (2-10); Joe Lovano (11-24); Branford Marsalis Trio (25-32); Coleman Hawkins (33-46,51,53,55,57, 58,60,62,65,67-72); Dexter Gordon (73-77); Miles Davis (87-92); Joe Henderson (94-99); and others		Leonard Bernstein "On the Town: Three Dance Episodes, Mvt. 1" (00) and "Symphonic dances from West Side Story, Prologue" (01)	clipping distortion (52,54,66)
Metal	(04,13) (34,94) (40,61) (41,62) (42,63) (43,64) (44,65) (45,66) (58) is Rock (16)		Dark Tranquillity (12-15); Dio (40-45,61-66); The New Bomb Turks (46-57); Queen (58-60); Metallica (33,38,73, 75,78,83,87); Iron Maiden (2,5,34,92-94); Rage Against the Machine (95-99); and others	(33,74) (85) is Ozzy Osbourne covering Disco (14)	Creed "I'm Eighteen" (21); Living Colour "Glamour Boys" (29); The New Bomb Turks "Hammerless Nail" (46), "Jukebox Lean" (50), "Jeers of a Clown" (51); Queen "Tie Your Mother Down" (58), "Tear it up" (59), "We Will Rock You" (60)	clipping distortion (33,73,84)
Pop	(15,22) (30,31) (45,46) (47,80) (52,57) (54,60) (56,59) (67,71) (87,90)	(68,73) (15,21,22) (47,48,51) (52,54) (57,60)	Mandy Moore (00,87-96); Mariah Carey (2,97-99); Alanis Morissette (3-9); Celine Dion (11,39,40); Britney Spears (15-38); Christina Aguilera (44-51,80); Destiny's Child (52-62); Janet Jackson (67-73); Jennifer Lopez (74-78,82); Madonna (84-86); and others	(10,14) (16,17) (74,77) (75,82) (88,89) (93,94)	Diana Ross "Ain't No Mountain High Enough" (63); Prince "The Beautiful Ones" (65); Kate Bush "Couldbusting" (79); Ladysmith Black Mambazo "Leaning On The Everlasting Arm" (81); Madonna "Cherish" (86)	(37) is from same recording as (15,21,22) but with sound effects
Reggae	(03,54) (05,56) (08,57) (10,60) (13,58) (41,69) (73,74) (80,81,82) (75,91,92)	(07,59) (33,44) (85,96)	Bob Marley (00-27,54-60); Dennis Brown (46-48,64-68,71); Prince Buster (85,94-99); Burning Spear (33,42,44,50, 63); Gregory Isaacs (70,76-78); and others	(23,55)	Pras "Ghetto Supastar (That Is What You Are)" (52); Marcia Griffiths "Electric Boogie" (88)	last 25 s of (86) are useless
Rock	(16) is Metal (58)		Morphine (0-9); Ani DiFranco (10-15); Queen (16-26); The Rolling Stones (28-31,33,35,37); Led Zeppelin (32,39-48); Simple Minds (49-56); Sting (57-63); Jethro Tull (64-70); Simply Red (71-78); Survivor (79-85); The Stone Roses (91-99)		Morphine "Hanging On A Curtain" (9); Queen "(You're So Square) Baby I Don't Care" (20); Billy Joel "Movin' Out" (36); Guns N' Roses "Knockin' On Heaven's Door" (38); Led Zeppelin "The Song Remains The Same" (39), "The Crunge" (40), "Dancing Days" (41), "The Ocean" (43), "Ten Years Gone" (45), "Night Flight" (46), "The Wanton Song" (47), "Boogie With Stu" (48); Simply Red "She's Got It Bad" (75), "Wonderland" (78); Survivor "Is This Love" (80), "Desperate Dreams" (83), "How Much Love" (84); The Tokens "The Lion Sleeps Tonight" (90)	jitter (27)

Table 2: The repetitions, potential mislabelings and distortions we find in *GTZAN*. Excerpt numbers are in parentheses. Exact repetitions are those excerpts that are the same with respect to a comparison of their time-frequency content. Recording repetitions are those excerpts we suspect coming from the same recording. Artist repetitions are those excerpts featuring the same artists. Version repetitions are covers of the same song. Potential mislabelings are excerpts we argue are misplaced with regards to the category top tags of its *GTZAN* category. Distortions are those excerpts that we regard as having a significant amount of distortion. This table can be auditioned online at http://imi.aau.dk/∼bst/research/GTZANtable2.

3.4 How do the faults of *GTZAN* affect evaluation?

We now study how the faults of *GTZAN* affect the evaluation of MGR systems, e.g., the classification accuracies in the nearly 100 published works seen in Fig. 2. Through our experiments below, we see the following two claims are false: 1) "all MGR systems and evaluations are affected in the same ways by the faults of *GTZAN*"; and 2) "the performances of all MGR systems in *GTZAN*, working with the same data and faults, are still meaningfully comparable." Thus, *regardless of how the systems are performing the task*, the results in Fig. 2 cannot be meaningfully interpreted.

It is not difficult to predict how some faults can affect the evaluations of MGR systems built using different approaches. For instance, when exact replicas are distributed across train and test sets, the evaluation of some systems can be more biased than others: a nearest neighbor classifier will find features in the training set with zero distance to the test feature, while a Bayesian classifier with a parametric model may not so strongly benefit when its model parameters are estimated from all training features. If there are replicas in the test set only, then they will bias an estimate of a figure of merit because they are not independent tests — if one is classified (in)correctly then its replicas are also classified (in)correctly. In addition to exact repetitions, we show above that the number of artists in *GTZAN* is at most 329. Thus, as Seyerlehner (2010); Seyerlehner et al. (2010) predict for *GTZAN*, its use in evaluating systems will be biased due to the *artist effect* (Pampalk et al., 2005; Flexer, 2007; Flexer and Schnitzer, 2009, 2010), i.e., the observation that a music similarity system can perform significantly worse when artists are disjoint in training and test datasets, than when they are not. Since all results in Fig. 2 come from evaluations without using an artist filter, they are quite likely to be optimistic.

To investigate the faults of *GTZAN*, we create several MGR systems using a variety of feature extraction and machine learning approaches paired with different training data in *GTZAN*. We define a *system* not as an abstract proposal of a machine learning method, a feature description, and so on, but as a real and working implementation of the components necessary to produce an output from an input (Sturm, 2013c). A system, in other words, has already been trained, and might be likened to a "black box" operated by a customer in an environment, according to some instructions. In our case, each system specifies the customer input 30 s of monophonic audio data uniformly sampled at 22050 Hz, for which it outputs one of the ten *GTZAN* category names.

Some systems we create by combining the same features with three classifiers (Duin et al., 2007): nearest neighbor (NN), minimum distance (MD), and minimum Mahalanobis distance (MMD) (Theodoridis and Koutroumbas, 2009). These systems create feature vectors from a 30 s excerpt in the following way. For each 46.4 ms frame, and a hop half that, it computes: 13 MFCCs using the approach by Slaney (1998), zero crossings, and spectral centroid and rolloff. For each 130 consecutive frames, it computes the mean and variance of each dimension, thus producing nine 32-dimensional feature vectors. From the feature vectors of the training data, the system finds a normalization transformation that maps each dimension to $[0, 1]$, i.e., the minimum of a dimension is subtracted from all values in that dimension, and then those are divided by the maximum magnitude difference between any two values. Each system applies the same transformation to the feature vectors extracted from an input. Each classifier chooses a label for an excerpt as follows: NN selects the class by majority vote, and breaks

GTZAN Category	**Fold 1**	**Fold 2**
Blues	John Lee Hooker, Kelly Joe Phelps, Buckwheat Zydeco, Magic Slim & The Teardrops	Robert Johnson, Stevie Ray Vaughan, Clifton Chenier, Hot Toddy, Albert Collins
Classical	J. S. Bach, Percy Grainger, Maurice Ravel, Henri Dutilleux, Tchaikovsky, Franz Schubert, *Leonard Bernstein*, misc.	Beethoven, Franz Joseph Haydn, Mozart, Vivaldi, Claude Debussy, misc.
Country	Shania Twain, Johnny Cash, Willie Nelson, misc.	Brad Paisley, George Strait, Vince Gill, misc.
Disco	Donna Summer, KC and The Sunshine Band, Ottawan, The Gibson Brothers, Heatwave, Evelyn Thomas, misc.	Carl Douglas, Village People, The Trammps, Earth Wind and Fire, Boney M., ABBA, Gloria Gaynor, misc.
Hip hop	De La Soul, Ice Cube, Wu-Tang Clan, Cypress Hill, Beastie Boys, 50 Cent, Eminem, misc.	A Tribe Called Quest, Public Enemy, *Lauryn Hill*, Wyclef Jean
Jazz	*Leonard Bernstein*, Coleman Hawkins, Branford Marsalis Trio, misc.	James Carter, Joe Lovano, Dexter Gordon, Tony Williams, Miles Davis, Joe Henderson, misc.
Metal	Judas Priest, Black Sabbath, *Queen*, Dio, Def Leppard, Rage Against the Machine, *Guns N' Roses*, New Bomb Turks, misc.	AC/DC, Dark Tranquillity, Iron Maiden, Ozzy Osbourne, Metallica, misc.
Pop	Mariah Carey, Celine Dion, Britney Spears, Alanis Morissette, Christina Aguilera, misc.	Destiny's Child, Mandy Moore, Jennifer Lopez, Janet Jackson, Madonna, misc.
Reggae	Burning Spear, Desmond Dekker, Jimmy Cliff, Bounty Killer, Dennis Brown, Gregory Isaacs, Ini Kamoze, misc.	Peter Tosh, Prince Buster, Bob Marley, *Lauryn Hill*, misc.
Rock	Sting, Simply Red, *Queen*, Survivor, *Guns N' Roses*, The Stone Roses, misc.	The Rolling Stones, Ani DiFranco, Led Zeppelin, Simple Minds, Morphine, misc.

Table 3: Composition of each fold of the artist filter partitioning (500 excerpts in each). Italicized artists appear in two *GTZAN* categories.

ties by selecting randomly among those classes that are ties; MD and MMD both select the label with the maximum log posterior sum over the nine feature vectors. Both MD and MMD model the feature vectors as independent and identically distributed multivariate Gaussian.

While the above approaches provide "baseline" systems, we also use two state-of-the-art approaches that produce MGR systems measured to have high classification accuracies in *GTZAN*. The first is SRCAM — proposed by Panagakis et al. (2009b) but modified by us (2013a) — which uses psychoacoustically-motivated features of 768 dimensions. SRCAM classifies an excerpt by using sparse representation classification (Wright et al., 2009). We implement this using the SPGL1 solver (van den Berg and Friedlander, 2008) with at most 200 iterations, and define $\epsilon^2 = 0.01$. The second approach is MAPsCAT, which uses feature vectors of "scattering transform" coefficients (Andén and Mallat, 2011). This produces 40 feature vectors of 469 dimensions. MAPsCAT models the features in the training set as independent and identically distributed multivariate Gaussian, and computes for a test feature the log posterior in each class. We define all classes equally likely for MAPsCAT, as well as for MD and MMD. As for the baseline systems above, systems built using SRCAM and MAPsCAT normalize input feature vectors according to the training set. (See Sturm (2013a) for further details of SRCAM and MAPsCAT.)

We use four different partition strategies to create train and test datasets from *GTZAN*: ten realizations of standard non-stratified 2fCV (ST); ST without the 67 exact and recording repetitions and 2 distortions (ST'); a non-stratified 2fCV with artist filtering (AF); AF without the 67 exact and recording repetitions and 2 distortions (AF'). Table 3 shows the composition of each fold in AF in terms of artists. We created AF manually to ensure that: 1) each *GTZAN* category is approximately balanced in terms of the number of training and testing excerpts; and 2) each fold of a *GTZAN* category has music tagged with category top tags (Fig. 5). For instance, the two Blues folds have 46 and 54 excerpts, and both have

music tagged "blues" and "zydeco." Unless otherwise noted, we do not take into account any potential mislabelings. In total, we train and test 220 MGR systems.

We look at several figures of merit computed from a comparison of the the outputs of the systems to the "ground truth" of testing datasets: confusion, precision, recall, F-score, and classification accuracy. Define the set of *GTZAN* categories $\mathcal{G}$. Consider that we input to a system $N^{(g)}$ number of excerpts from *GTZAN* category $g \in \mathcal{G}$, and that of these the system categorizes as r $\in \mathcal{G}$ the number $M^{(g \text{ as } r)} \leq N^{(\mathrm{g})}$. We define the *confusion* of *GTZAN* category g as r for a system

$$C^{(g \text{ as } r)} := M^{(g \text{ as } r)} \big/ N^{(g)}. \tag{1}$$

The *recall* for *GTZAN* category g of a system is then $C^{(g \text{ as } g)}$. We define the *normalized accuracy* of a system by

$$A := \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} C^{(g \text{ as } g)}. \tag{2}$$

We use normalized accuracy because the number of inputs in each *GTZAN* category may not be equal in a test set. We define the *precision* of a system for *GTZAN* category g as

$$P^{(\mathrm{g})} := M^{(\mathrm{g \text{ as } g})} \Big/ \sum_{r \in \mathcal{G}} M^{(\mathrm{r \text{ as } g})}. \tag{3}$$

Finally, we define the *F-score* of a system for *GTZAN* category g as

$$F^{(g)} := 2P^{(g)} C^{(g \text{ as } g)} \big/ \left[P^{(g)} + C^{(g \text{ as } g)} \right]. \tag{4}$$

To test for significant differences in the performance between two systems in the same test dataset, we build a contingency table (Salzberg, 1997). Define the random variable N to be the number of times the two systems choose different categories, but one is correct. Let t_{12} be the number for which system 1 is correct but system 2 is wrong. Thus, $N - t_{12}$ is the number of observations for which system 2 is correct but system 1 is wrong. Define the random variable T_{12} from which t_{12} is a sample. The null hypothesis is that the systems perform equally well given $N = n$, i.e., $E[T_{12}|N = n] = n/2$, in which case T_{12} is distributed binomialy, i.e.,

$$p_{T_{12}|N=n}(t) = \binom{n}{t} (0.5)^n, 0 \leq t \leq n. \tag{5}$$

The probability we observe a particular performance given the systems actually perform equally well is

$$\begin{aligned} p &:= P[T_{12} \leq \min(t_{12}, n - t_{12})] + P[T_{12} \geq \max(t_{12}, n - t_{12})] \\ &= \sum_{t=0}^{\min(t_{12}, n-t_{12})} p_{T_{12}|N=n}(t) + \sum_{t=\max(t_{12}, n-t_{12})}^{n} p_{T_{12}|N=n}(t). \end{aligned} \tag{6}$$

We define statistical significance as $\alpha = 0.05$, and reject the null hypothesis if $p < \alpha$.

Figure 7 shows a summary of the normalized accuracies of all our MGR systems, with respect to the five approaches and four partitions we use. As each partition breaks *GTZAN* into two folds, the left and right end points of a segment correspond to using the first or second fold for training, and the other for testing. For the ten random partitions of ST and

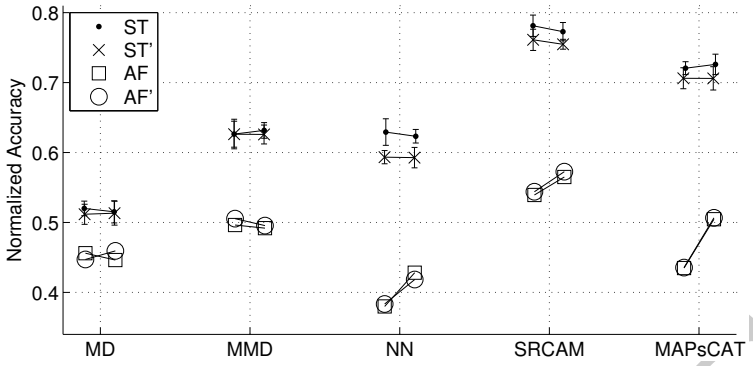

Figure 7: Normalized accuracy (2) of each approach (x-axis) for each fold (left and right) of different partition strategies (legend). One standard deviation above and below the mean are shown for ST and ST'.

ST', we show the mean normalized accuracy, and a vertical line segment of two standard deviations centered on the mean. It is immediately clear that the faults of *GTZAN* affect an estimate of the classification accuracy for an MGR system. For systems created with all approaches except NN, the differences between conditions ST and ST' are small. As we predict above, the performance of systems created using NN appears to benefit more than the others from the exact and recording repetition faults of *GTZAN*, boosting their mean normalized accuracy from below 0.6 to 0.63. Between conditions AF and AF', we see that removing the repetitions produces very little change, presumably because the artist filter keeps exact and recording repetitions from being split across train and test datasets. Most clearly, however, we see for all systems a large decrease in performance between conditions ST and AF, and that each system is affected to different degrees. The difference in normalized accuracy between conditions ST and AF appears the smallest for MD (7 points), while it appears the most for MAPsCAT (25 points). Since we have thus found systems with performance evaluations affected to different degrees by the faults in *GTZAN* — systems using NN are hurt by removing the repetitions while those using MMD are not — we have disproven the claim that the faults of *GTZAN* affect all MGR systems in the same ways.

When we test for significant differences in performance between all pairs of system in the conditions ST and ST', only for those created with MMD and NN do we fail to reject the null hypothesis. In terms of classification accuracy in the condition ST, and the binomial test described above, we can say with statistical significance: those systems created using MD perform worse than those systems created using MMD and NN, which perform worse than those systems created using MAPsCAT, which perform worse than those systems created using SRCAM. The systems created using SRCAM and MAPsCAT are performing significantly better than the baseline systems. In the conditions AF and AF', however, we fail to reject the null hypothesis for systems created using MAPsCAT and MD, and systems created using MAPsCAT and MMD. MAPsCAT, measured as performing significantly better than the baseline in the conditions ST and ST' — and hence motivating conclusions that the features it uses are superior for MGR (Andén and Mallat, 2011; Andén and Mallat, 2013) — now performs no better than the baseline in the conditions AF and AF'. Therefore, this disproves that the performances of all MGR systems in *GTZAN* — all results shown in Fig. 2 — are still meaningfully comparable for MGR. Some of them benefit significantly more

	GTZAN category										
	Blues	Classical	Country	Disco	Hip hop	Jazz	Metal	Pop	Reggae	Rock	Precision
Blues	100	0	1	0	0	0	0	0	0	0	99.0
Classical	0	100	0	0	0	2	0	0	0	0	98.0
Country	0	0	95	0	0	0	0	0	0	1	99.0
Disco	0	0	0	92	1	0	0	2	0	0	97.0
Hip hop	0	0	0	1	96	0	0	0	1	0	98.0
Jazz	0	0	0	0	0	98	0	0	0	0	100.0
Metal	0	0	0	0	0	0	92	0	0	13	87.6
Pop	0	0	0	2.5	3	0	3	95	1	3	88.3
Reggae	0	0	0	0	0	0	0	0	98	1	99.0
Rock	0	0	4	4.5	0	0	5	3	0	82	83.3
F-score	99.5	99.0	96.9	94.4	97.2	99.0	89.8	91.6	98.5	82.7	Acc: 94.8

Table 4: The worst figures of merit ($\times 10^{-2}$) of a "perfect" classifier evaluated with *GTZAN*, which takes into account the 52 potential mislabelings in Table 2, and assuming that the 81 excerpts we have yet to identify have "correct" labels.

than others due to the faults in *GTZAN*, but which ones they are cannot be known.

We now focus our analysis upon systems created using SRCAM. Figure 8 shows averaged figures of merit for systems built using SRCAM and the 10 realizations of ST and ST', as well as for the single partition AF'. Between systems built in conditions ST and ST', we see very little change in the recalls for *GTZAN* categories with the fewest exact and recording repetitions: Blues, Classical and Rock. However, for the *GTZAN* categories having the most exact and recording repetitions, we find large changes in recall. In fact, Fig. 9 shows that the number of exact and recording repetitions in a *GTZAN* category is correlated with a decrease in the recall of SRCAM. This makes sense because SRCAM is like adaptive nearest neighbors (Noorzad and Sturm, 2012). When evaluating the systems we created using SRCAM in the condition AF', Fig. 8 shows the mean classification accuracy is 22 points lower than that in condition ST. With respect to F-score, excerpts in *GTZAN* Classical and Metal suffer the least; but we see decreases for all other categories by at least 10 points, e.g., 62 points for *GTZAN* Blues, 29 points for *GTZAN* Reggae, and 25 points for *GTZAN* Jazz. We see little change in classification accuracy between conditions AF' and AF (not shown).

The question remains, what are the worst figures of merit that a "perfect" MGR system can obtain in *GTZAN*? We first assume that the 81 excerpts yet to be identified are "correct" in their categorization, and that the last 25 seconds of Reggae 86 are ignored because of severe distortion. Then, we assume a "perfect" system categorizes all *GTZAN* excerpts in their own categories, except for the 52 potential misclassifications in Table 2. For each of those excerpts, we consider that the system chooses the other *GTZAN* categories in which its number of votes are higher than, or as high as, the number of votes in its own category. For instance, Country 39 has its largest vote in *GTZAN* Blues, so we add one to the Blues row in the *GTZAN* Country column. When a vote is split between K categories, we add $1/K$ in the relevant positions of the confusion table. For instance, Disco 11 has the same number of votes in *GTZAN* Disco, Pop and Rock, so we add 0.5 to the Pop and to the Rock rows of the *GTZAN* Disco column.

Table 4 shows the worst figures of merit we expect for this "perfect" system, which is also imposed as the thick gray line in Fig. 2. If the figures of merit of an MGR system tested in *GTZAN* are better that in this, it might actually be performing worse than the "perfect" system. Indeed, we have found that the classification accuracies for six MGR

	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	87.50	0.00	1.80	0.80	1.60	1.70	0.00	0.30	2.60	6.30	85.56
cl	2.40	92.90	3.20	1.60	0.10	4.00	0.20	0.10	1.10	1.50	87.08
co	1.10	1.30	72.60	2.10	1.70	3.00	0.80	1.30	2.40	8.60	76.75
di	1.70	0.30	2.30	61.80	4.10	0.60	0.50	5.30	5.90	5.10	70.81
hi	2.10	0.20	0.20	4.70	78.50	0.80	0.90	2.10	7.30	0.90	80.67
ja	0.80	2.00	1.80	0.20	1.00	86.10	0.40	0.30	0.90	1.00	91.21
me	0.90	0.90	1.10	1.80	3.20	1.30	93.00	1.70	1.30	18.40	75.45
po	0.00	0.00	5.40	7.70	3.60	0.70	0.30	84.60	7.70	2.00	75.71
re	1.80	0.10	1.90	8.60	5.80	0.30	0.00	2.80	66.90	3.10	73.43
ro	1.70	2.30	9.70	10.70	0.40	1.50	3.90	1.50	3.90	53.10	60.25
F	86.39	89.74	74.42	65.67	79.37	88.51	83.23	79.82	69.89	56.21	77.70

(a) SRCAM, ST

	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	87.60	0.41	4.08	0.53	1.51	1.93	0.55	0.47	3.51	5.60	83.50
cl	2.00	93.04	2.65	2.35	0.33	4.25	0.00	0.47	1.17	0.90	87.65
co	1.70	1.81	69.69	2.70	1.73	4.36	0.56	2.61	3.73	8.90	72.69
di	1.80	0.30	1.63	57.34	4.25	0.57	0.33	6.33	6.44	6.40	67.82
hi	1.50	0.30	0.61	5.48	77.57	0.92	0.88	2.49	9.81	0.30	78.45
ja	1.70	0.70	1.62	0.22	0.66	83.52	0.11	0.48	1.02	1.20	91.05
me	0.70	0.81	1.13	2.56	3.16	1.15	93.62	2.12	1.61	16.00	75.61
po	0.00	0.00	5.44	7.43	4.56	1.03	0.00	79.98	6.95	3.40	72.15
re	1.10	0.10	1.95	10.76	5.91	0.67	0.22	2.58	60.87	2.60	69.59
ro	1.90	2.52	11.21	10.63	0.33	1.60	3.74	2.48	4.89	54.70	60.38
F	85.37	90.21	70.98	61.85	77.81	86.89	83.50	75.61	64.38	57.17	75.79

(b) SRCAM, ST'

	bl	cl	co	di	hi	ja	me	po	re	ro	Pr
bl	18.12	1.92	6.17	2.13	3.24	6.05	1.04	1.16	5.90	10.00	33.97
cl	1.09	89.32	2.08	1.06	0.00	6.05	0.00	1.22	1.14	1.00	88.00
co	6.80	0.96	59.08	3.24	3.29	4.35	0.00	1.16	2.33	17.00	60.31
di	9.30	1.92	3.04	45.12	7.49	1.00	2.08	14.12	18.34	13.00	40.57
hi	10.39	0.00	0.00	6.43	55.44	0.00	0.00	2.33	15.10	0.00	62.20
ja	19.00	1.92	1.04	0.00	1.11	60.43	1.16	0.00	1.14	2.00	66.15
me	7.29	0.96	1.04	0.00	5.41	1.00	87.26	3.49	0.00	15.00	70.51
po	1.09	0.00	8.25	15.17	7.59	10.76	0.00	70.53	10.44	6.00	52.78
re	11.15	0.00	5.04	8.51	13.14	7.00	0.00	3.55	40.91	4.00	44.72
ro	15.78	2.99	14.25	18.34	3.29	3.35	8.45	2.44	4.71	32.00	31.37
F	23.61	88.60	59.63	42.49	58.63	63.15	77.89	59.60	40.79	31.65	55.82

(c) SRCAM, AF'

Figure 8: 100× confusion, precision (Pr), F-score (F), and normalized accuracy (bottom right corner) for systems built using SRCAM, trained and tested in ST and ST' (averaged over 10 realizations), as well as in AF' (with artist filtering and without replicas). Columns are "true" labels; rows are predictions. Darkness of square corresponds to value. Labels: Blues (bl), Classical (cl), Country (co), Disco (di), Hip hop (hi), Jazz (ja), Metal (me), Pop (po), Reggae (re), Rock (ro).

approaches appearing close to this limit (each marked by an "x" in Fig. 2) are due instead to inappropriate evaluation designs (discussed further in Section 4.5).

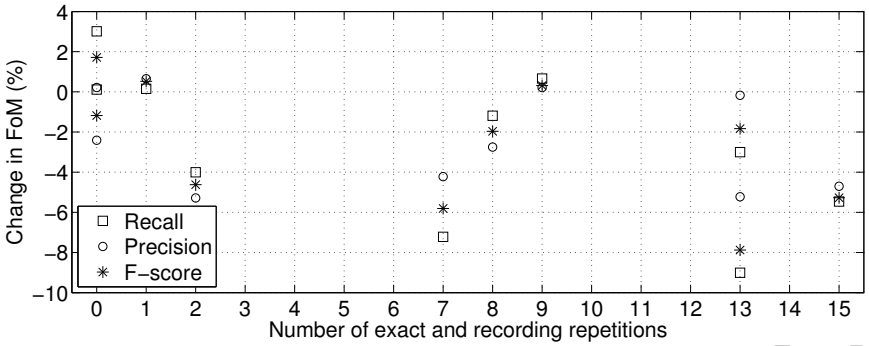

Figure 9: Scatter plot of percent change in mean figures of merit (FoM) for SRCAM between the conditions ST and ST' (Fig. 8) as a function of the number of exact and recording repetitions in a *GTZAN* category.

3.5 Discussion

Though it may not have been originally designed to be *the* benchmark dataset for MGR, *GTZAN* is currently honored by such a position by virtue of its use in some of the early influential works (Tzanetakis and Cook, 2002), its continued wide-spread use, and its public availability. It was, after all, one of the first publicly-available datasets in MIR. Altogether, *GTZAN* appears more than any other dataset in the evaluations of nearly 100 MGR publications (Sturm, 2012b). It is thus fair to claim that researchers who have developed MGR systems have used, or been advised to use, *GTZAN* for evaluating success. For instance, high classification accuracy in *GTZAN* has been argued as indicating the superiority of approaches proposed for MGR (Li and Ogihara, 2006; Bergstra et al., 2006; Panagakis et al., 2009b). Furthermore, performance in *GTZAN* has been used to argue for the relevance of features to MGR, e.g., "the integration of timbre features with temporal features are important for genre classification" (Fu et al., 2011). Despite its ten-year history spanning most of the research in MGR since the review by Aucouturier and Pachet (2003), only five works indicate someone has taken a look at what is in *GTZAN*. Of these, only one (Li and Chan, 2011) implicitly reports listening to all of *GTZAN*; but another completely misses the mark in its assessment of the integrity of *GTZAN* (Bergstra et al., 2006). When *GTZAN* has been used, it has been used for the most part without question.

GTZAN has never had metadata identifying its contents, but our work finally fills this gap, and shows the extent to which *GTZAN* has faults. We find 50 exact repetitions, 21 suspected recording repetitions, and a significant amount of artist repetition. Our analysis of the content of each *GTZAN* category shows the excerpts in each are more diverse than what the category names suggest, e.g., the *GTZAN* Blues excerpts include music using the blues genre, as well as the zydeco and cajun and genres. We find potential mislabelings, e.g., two excerpts of orchestral music by Leonard Bernstein appear in *GTZAN* Jazz, but might be better placed with the other orchestral excerpts of Bernstein in *GTZAN* Classical; an excerpt by Wayne Toups & Zydecajun appears in *GTZAN* Country, but might be better placed with the other cajun and zydeco music in *GTZAN* Blues. Finally, we find some excerpts suffer from distortions, such as jitter, clipping, and extreme degradation (Reggae 86).

Having proven that *GTZAN* is flawed, we then sought to measure their real effects on the evaluation of MGR systems. Using three baseline and two state of the art approaches,

we disproved the claims that all MGR systems are affected in the same ways by the faults of *GTZAN*, and that the performances of MGR systems in *GTZAN* are still meaningfully comparable no matter how the systems are performing the task. In other words, evaluating systems in *GTZAN* with *Classify* without taking into consideration its contents provides numbers that may be precise and lend themselves to a variety of formal statistical tests, but that are nonetheless irrelevant for judging which system can satisfy the success criteria of some use case in the real world that requires musical intelligence, not to mention whether the system demonstrates a capacity for recognizing genre *at all*. This lack of validity in standard MGR evaluation means one cannot simply say, "94.8 in *GTZAN* is the new 100," and proceed with "business as usual."

Since all results in Fig. 2 come from experimental conditions where independent variables are not controlled, e.g., by artist filtering, the "progress" in MGR seen over the years is very suspect. Systems built from MAPsCAT and SRCAM, previously evaluated in *GTZAN* to have classification accuracies of 83% without considering the faults of *GTZAN* (Sturm, 2012c, 2013a), now lay at the bottom in Fig. 2 below 56%. Where the performances of all the other systems lie, we do not yet know. The incredible influence of the other uncontrolled independent variables confounded with the *GTZAN* category names becomes clear with irrelevant transformations (Sturm, 2013g).

Some may argue that our litany of faults in *GTZAN* ignores what they say is the most serious problem with a dataset like *GTZAN*: that it is too small to produce meaningful results. In some respects, this is justified. While personal music collections may number thousands of pieces of music, commercial datasets and library archives number in the millions. The 1000 excerpts of *GTZAN* is most definitely an insufficient random sample of the population of excerpts "exemplary" of the kinds of music between which one may wish a MGR system to discriminate if it is to be useful for some use case. Hence, one might argue, it is unreasonably optimistic to assume an MGR system can learn from a fold of *GTZAN* those rules and characteristics people use (or at least Tzanetakis used) when describing music as belonging to or demonstrating aspects of the particular genres used by music in *GTZAN*. Such warranted skepticism, compounded with the pairing of well-defined algorithms and an ill-defined problem, highlights the absurdity of interpreting the results in Fig. 2 as indicating real progress is being made in suffusing computers with musical intelligence.

One might hold little hope that *GTZAN* could ever be useful for tasks such as evaluating systems for MGR, audio similarity, autotagging, and the like. Some might call for *GTZAN* to be "banished" — although we have yet to find any paper that says so. There are, however, many ways to evaluate an MGR system using *GTZAN*. Indeed, its faults are representative of data in the real-world, and they can be used in the service of evaluation (Sturm, 2013a,g). For instance, by using the *GTZAN* dataset, we (2012c; 2013a; 2013g) perform several different experiments to illuminate the (in)sanity of a system's internal model of music genre. In one experiment (2012c; 2013a), we look at the kinds of pathological errors of an MGR system rather than what it labels "correctly." We design a reduced Turing test to determine how well the "wrong" labels it selects *imitate* human choices. In another experiment (2012c; 2013g), we attempt to fool a system into selecting any genre label for the same piece of music by changing factors that are irrelevant to genre, e.g., by subtle time-invariant filtering. In another experiment (2012c), we have an MGR system compose music excerpts it hears as highly representative of the "genres of *GTZAN*", and then we perform a formal listening

test to determine if those genres are recognizable. The lesson is not to banish *GTZAN*, but to use it with full consideration of its musical content. It is currently not clear what makes a dataset "better" than another for MGR, and whether any are free of the kinds of faults in *GTZAN*; but at least now with *GTZAN*, one has a manageable, public, and finally well-studied dataset.

4 Implications for future work in MGR and MIR

The most immediate point to come from this work is perhaps that one should not take for granted the integrity of any given dataset, even when it has been used in a large amount of research. Any researcher evaluating an MIR system using a dataset must know the data, know real data has faults, and know such faults have real impacts. However, there are five other points important for the next ten years of MIR. First, problems in MIR should be unambiguously defined through specifying use cases and with formalism. Second, experiments must be designed, implemented, and analyzed such that they have validity and relevance with respect to the intended scientific questions. Third, system analysis should be deeper than just evaluation. Fourth, the limitations of all experiments should be acknowledged, and appropriate degrees of skepticism should be applied. Finally, as a large part of MIR research is essentially algorithm- and data-based, it is imperative to make all such work reproducible.

4.1 Define problems with use cases and formalism.

As mentioned in Section 2, few works of our survey (2012b) define "the problem of music genre recognition," but nearly all of them treat genre as categories to which objects belong, separated by boundaries, whether crisp or fuzzy, objective or subjective, but recoverable by machine learning algorithms. The monopoly of this Aristotelean viewpoint of genre is evinced by the kind of evaluation performed in MGR work (Sturm, 2012b): 397 of 435 published works pose as relevant the comparison of labels generated by an MGR system with a "ground truth" (*Classify*). All but ten of these works consider MGR as a single label classification problem. We have posed (2013b) "the principal goal of MGR," but the discussion of what its principal goal is might be premature without the consideration of well-defined use cases.

The parameters and requirements of a music recommendation service are certainly different from those of a person wanting to organize their home media collection, which are different from those of a music library looking to serve the information needs of musicologists. The light provided by a use case can help define a problem, and show how the performance of a solution should be measured; however, only a few works in the MGR literature consider use cases. In the same direction, the MIReS Roadmap (Serra et al., 2013) recommends the development of "meaningful evaluation tasks." A use case helps specify the solutions, and what is to be tested by evaluations. One good example is provided by one of the first works in MGR: Dannenberg et al. (1997) seeks to address the specific scenario of interactive music performance between a musician and machine listener. It poses several requirements: the musician must play styles consistently; the classification must take no longer than five seconds; and the number of false positives should be minimized. Dannenberg et al. evaluate solutions not only in the laboratory, but also in the intended situation. This work shows how a use case can help define a problem, and the success criteria for the solution.

To fully define a problem with no uncertain terms, it must be formalized. Formalization is a tool that can disambiguate all parts of a problem in order to clarify assumptions, highlight the existence of contradictions and other problems, suggest and analyze solutions, and provide a path for designing, implementing and analyzing evaluations. By and large, such formalization remains implicit in much evaluation of MIR, which has been uncritically and/or unknowingly borrowed from other disciplines, like machine learning and text information retrieval (Sturm, 2013c; Urbano et al., 2013). We have attempted to finally make this formalism explicit (2013c): we define what a system is, and what it means to analyze one; we use the formalism of the design and analysis of experiments, e.g., Bailey (2008), to dissect an evaluation into its aims, parts, design, execution, and analysis; and we show how an evaluation in MIR *is* an experiment, and thus makes several assumptions. When this formalism is made explicit, it becomes clear why the systematic and rigorous evaluations performed using standardized datasets in many MIREX tasks can still not be *scientific* evaluation (Dougherty and Dalton, 2013; Sturm, 2013c). Consequently, what must be done in order to address this becomes clearer.

4.2 Design valid and relevant experiments.

No result in Fig. 2 provides a reasonable direction for addressing a use case that requires an artificial system having musical intelligence. That a MGR system evaluated using *Classify* achieves even a perfect classification accuracy in a dataset having uncontrolled independent variables provides no logical support for the claim that the system is using criteria relevant to the meaning behind those labels (Sturm, 2012c, 2013a,b). The same is true of the results from the past several years of the MIREX MGR task, or the MIREX audio mood classification task, or the various MIREX autotagging tasks (Sturm, 2013c). An unwary company looking to build and sell a solution for a particular use case *requiring a system capable of music intelligence* will be misled by all these results when following the intuition that a system with a classification accuracy of over 0.8 must be "better" than one below 0.7. This is not to say that none of the systems in Fig. 2 are using relevant criteria to decide on the genre labels (e.g., instrumentation, rhythm, form), but that such a conclusion does not follow from the experiment, *no matter its outcome*. In short, none of this evaluation is valid.

In discussing the *music intelligence* of a system, and the lack of validity in evaluation for measuring it, we might be mistaken as misrepresenting the purpose of machine learning, or overstating its capabilities. Of course, machine learning algorithms are agnostic to the intentions of an engineer employing them, and are unperturbable by any accusation of being a "horse" (Sturm, 2013g). Machine learning does not create systems that think or behave as humans, or mimic the decision process of humans. However, there are different use cases where "by any means" is or is not acceptable. Even when a problem is underspecified, e.g., "The mob boss asked me to make the guy disappear," then any solution might not be acceptable, e.g., "so, I hired a magician." If one seeks only to give the illusion that a robot can recognize genre and emotion in music (Xia et al., 2012), or that a musical water fountain is optimal (Park et al., 2011), then "whatever works" may be fine. However, if one seeks to make a tool that is useful to a musicologist for exploring stylistic influence in a music archive, then a system that uses musically meaningful criteria to make judgements is likely preferable over one that works by confounded but irrelevant factors.

Validity in experiments is elegantly demonstrated with the case of "Clever Hans" (Pfungst, 1911; Sturm, 2013g). Briefly, Hans was a horse that appeared able to solve complex arithmetic feats, as well as many other problems that require humans to think abstractly. Many people were convinced by Hans' abilities; and a group of intellectuals was unable to answer the question, "Is Hans capable of abstract thought?" It was not until valid experiments were designed and implemented to test well-posed and scientific questions that Hans was definitively proven to be incapable of arithmetic — as well as his many other intellectual feats — and only appeared so because he was responding to unintended but confounded cues of whoever asked him a question (Pfungst, 1911). The question of Hans' ability in arithmetic is not validly addressed by asking Hans more arithmetic problems and comparing his answers with the "ground truth" (*Classify*), but by designing valid and relevant experiments in which careful control is exercised over independent variables.

Confounds in machine learning experiments create serious barriers to drawing conclusions from evaluations — which can also seen as the problem of "overfitting." Consider that we wish to create a classifier to predict the sexes of people. We begin by randomly sampling 48,842 people from around the United States of America (USA) in 1996. Consider that we are not aware that the sex of a person is determined (for the most part) by the presence or absence of the "Y" chromosome, and so we ask about age, education, occupation, marital status, relationship to spouse, and many other things that might be relevant, and create a multivariate dataset from the entries.[16] As is common in machine learning, we split this into a training dataset (2/3) and testing dataset (1/3), and then perform feature selection (Theodoridis and Koutroumbas, 2009) using the training dataset to find the best features of sex with respect to, e.g., maximizing the ratio of intra- to inter-class separation, or by considering the Fisher discriminant. Finally, we create several classifiers using these feature combinations and training dataset, evaluate each using *Classify* on the testing dataset, and compute its normalized classification accuracy (accuracy).

Figure 10 plots the accuracies of the classifiers as a function of the number of best features. We find that with the "relationship" feature[17] we can build a classifier with an accuracy of almost 0.7 in the testing dataset. If we add to this the "occupation" feature[18] then accuracy increases to about 0.8 in the test dataset. If we also consider the "marital-status" feature[19] then we can build a classifier having an accuracy that exceeds 0.85 in the test dataset. Since we believe random classifiers would have an accuracy of 0.5 in each case, we might claim from these results, "Our systems predict the sex of a person better than random." We might go further and conclude, "The two most relevant features for predicting the sex of a person is 'relationship' and 'occupation'."

Now consider these results while knowing how sex is truly determined. While it is clear that the "relationship" feature is a direct signifier of sex given a person is married, we know that the "occupation" feature is *irrelevant* to the sex of *any* person. It is only a factor that happens to be confounded in our dataset with the attribute of interest through the economic

[16]We use here the multivariate dataset, `http://archive.ics.uci.edu/ml/datasets/Adult`

[17]This feature takes a value in { "Wife", "Husband", "Unmarried", "Child"}.

[18]This feature takes a value in { "Tech-support", "Craft-repair", "Other-service", "Sales", "Exec-managerial", "Prof-specialty", "Handlers-cleaners", "Machine-op-inspct", "Adm-clerical", "Farming-fishing", "Transport-moving", "Priv-house-serv", "Protective-serv", "Armed-Forces"}.

[19]This feature takes a value in {"Married", "Divorced", "Never-married", "Widowed", "Separated"}.

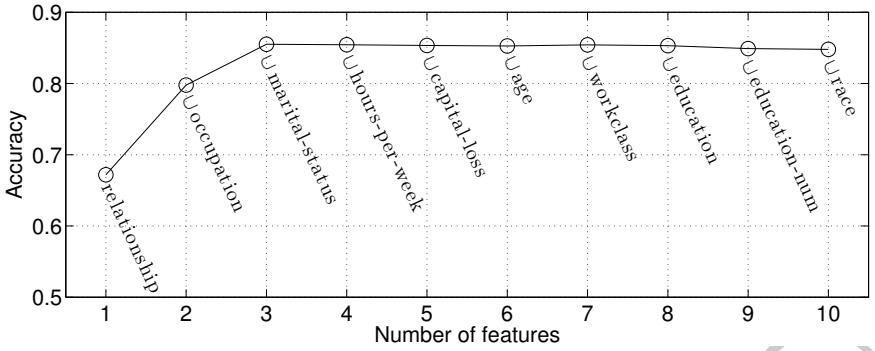

Figure 10: Normalized classification accuracy of male/female classifier (in test dataset) as a function of the number of features selected by maximizing inter- and intra-class separation. The three features producing the highest classification accuracy in the training set are "relationship", "occupation" and "marital-status."

and societal norms of the population from which we collected samples, i.e., 1996 USA. If we test our trained systems in populations with economics and societal norms that are different from 1996 USA, the confounds could break and performance thus degrade. Furthermore, the "occupation" feature only appears to be relevant to our task because: 1) it is a confound factor of sex more correlated than the others; and 2) it is among the attributes we chose for this task from the outset, few of which have any relevance to sex. Hence, due to confounds — independent variables that are not controlled in the experimental design — the evaluation lacks validity, and neither one of the conclusions can follow from the experiment.

Our work joins the growing chorus of concerns about validity in MIR evaluation. A number of interesting and practical works have been produced by Urbano et al. (Urbano et al., 2011; Urbano, 2011; Urbano et al., 2012, 2013). The work in Urbano et al. (2011) is concerned with evaluating the performance of audio similarity algorithms, and is extended in Urbano et al. (2012) to explore how well an MIR system must perform in a particular evaluation in order to be potentially useful in the real world. Urbano (2011) deals specifically with the question of experimental validity of evaluation in MIR; and Urbano et al. (2013) extend this to show that evaluation *of* evaluation remains neglected by the MIR community. Aucouturier and Bigand (2013) look at reasons why MIR has failed to acquire attention from outside computer science, and pose it is due in no small part to its lack of a scientific approach. While evaluation is also cited in the MIReS roadmap (Serra et al., 2013) as an current area of concern, such concerns are by no means only recent. The initiatives behind MIREX were motivated in part to address the fact that evaluation in early MIR research did not facilitate comparisons with new methods (Downie, 2003, 2004). MIREX has since run numerous systematic, rigorous and standardized evaluation campaigns for many different tasks since 2005 (Downie, 2008; Downie et al., 2010), which has undoubtedly increased the visibility of MIR (Cunningham et al., 2012). Furthermore, past MIR evaluation has stumbled upon the observations that evaluations that split across training and testing datasets music from the same albums (Mandel and Ellis, 2005) or artists (Pampalk et al., 2005) results in inflated performances than when the training and testing dataset do not share songs from the same album or by the same artists. Just as we have shown above for *GTZAN*, the effects of this can be quite significant (Flexer, 2007), and can now finally be explained through our work as being a result of confounds.

Evaluation in MIR is not just a question of what figure of merit to use, but of how to draw a valid conclusion by designing, implementing, and analyzing an experiment relevant to the intended scientific question (Sturm, 2013c). For any experiment, thorough consideration must be given to its "materials," "treatments," "design," "responses," "measurements," and "models" of those measurements. Beginning ideally with a well-specified and scientific question, the experimentalist is faced with innumerable ways of attempting to answer that question. However, a constant threat to the validity of an experiment is the selection of the components of an experiment because of convenience, and not because they actually address the scientific question. Another threat is the persuasion of precision. Richard Hamming has argued,[20] "There is a confusion between what is reliably measured, and what is relevant. ... What can be measured precisely or reliably does not mean it is relevant." Another threat is the blind application of statistics, resulting in "errors of the third kind" (Kimball, 1957): *correct answers to the wrong questions.* Hand (1994) shows why one must compare the intended scientific questions to the statistical questions that are actually answered. While the absence of formal statistics in MIR research has been highlighted by Flexer (2006), statistics is agnostic to the quality of data, the intention of the experiment, and the ability of the experimentalist. Statistics provides powerful tools, but none of them can rescue an invalid evaluation.

One might be inclined to say that since one problem with Fig. 2 is that *GTZAN* has faults that are not considered in any of the results, better datasets could help. However, though a dataset may be larger and more modern than *GTZAN* does not mean that it is free of the same kinds of faults found in *GTZAN*. Furthermore, that a dataset is large does not free the designer of an MIR system of the necessarily difficult task of designing, implementing, and analyzing an evaluation having the validity to conclude — if such a thing is indeed relevant to a use case — whether the decisions and behaviors of a system are related to the musical content *that is supposedly behind those decisions* (Sturm, 2013b,g). We have shown instead that the significant root of this problem, of which data plays a role, is in the validity and relevance of experiments (Sturm, 2013c,b). All independent variables must be accounted for an in evaluation, by controlling the experimental conditions, and by guaranteeing that enough material is sampled in a random enough way that the effects of the independent variables that are out of reach of the experimentalist will satisfy the assumptions of the measurement model (Sturm, 2013c). What amount of data is necessary depends on the scientific question of interest; but even a million songs may not be enough (Bertin-Mahieux et al., 2011).

An excellent example providing a model for the design, implementation, analysis and description of valid and relevant scientific experiments is that of Chase (2001) — which complements the earlier and pioneering results of Porter and Neuringer (1984). This work shows the effort necessary to planning and executing experiments to take place over a duration of about two years such that waste is avoided and results are maximized. Chase designs four experiments to answer two specific questions: 1) Can fish learn to discriminate between complex auditory stimuli that humans can also discriminate? 2) If so, does this behavior generalize to novel auditory stimuli with the same discriminability? Chase uses three koi fish

[20] R. Hamming, "You get what you measure", lecture at Naval Postgraduate School, June 1995. `http://www.youtube.com/watch?v=LNhcaVi3zPA`

as the material of the experiment, and specifies as the stimuli (treatments) recorded musical audio, both real and synthesized, in the categories "blues" and "classical." Chase designs the experiments to have hundreds of trials, some of which investigate potential confounds, probe the generalizability of the discrimination of each fish, confirm results when the reinforcement stimuli are switched, and so on. From these experiments, one can conclude whether fish can indeed learn to discriminate between complex auditory stimuli like music styles.

Unfortunately, much evaluation in MGR, music emotion recognition, and autotagging has been just the kind of show performed by Clever Hans (Pfungst, 1911): certainly, an evaluation observed by an audience and done not to deceive, but entirely irrelevant for proving any capacity of these systems for musical intelligence. This has thus contributed to the training of horses of many exotic breeds, and competitions to show whose horse is better-trained; but, aside from this pageantry, one does not know whether any meaningful problem has ever been addressed in it all. The design and implementation of a valid experiment requires creativity and effort that cannot be substituted with collecting more data; and a flawed design or implementation cannot be rescued by statistics. If one can elicit any response from a system by changing irrelevant factors, then one has a horse (Sturm, 2013g).

4.3 Perform system analysis deeper than just evaluation.

In our work (2013c), we define a system ("a connected set of interacting and interdependent components that together address a goal"), system analysis (addressing questions and hypotheses related to the past, present and future of a system), and highlight the key role played by evaluation ("a 'fact-finding campaign' intended to address a number of relevant questions and/or hypotheses related to the goal of a system"). A system analysis that only addresses a question of how well a system reproduces the labels of a testing dataset is quite shallow and of limited use. Such an analysis does not produce knowledge about how the system is operating or whether its performance is satisfactory with respect to some use case. More importantly, an opportunity is lost for determining how a system can be improved, or adapted to different use cases. One of the shortcomings of evaluation currently practiced in MIR is that system analysis remains shallow, and does not help one to understand or improve systems.

The MIReS Roadmap (Serra et al., 2013) also points to the need for deeper system analysis. One challenge is identifies is the design of evaluation that provides "qualitative insights on how to improve [systems]." Another challenge is that entire systems should be evaluated, and not just their unconnected components. In a similar light, Urbano et al. (2013) discuss how the development cycle that is common in Information Retrieval research is unsatisfactorily practiced in MIR: the definition of a task in MIR often remains contrived or artificial, the development of solutions often remains artificial, the evaluation of solutions is often unsatisfactory, the interpretation of evaluation results is often biased and unilluminating, the lack of valid results lead to improving solutions "blindly," and returning to the beginning of the cycle rarely leads to a qualitative improvement of research.

4.4 Acknowledge limitations and proceed with skepticism.

The third point to come from this work is that the limitations of experiments should be clearly acknowledged. In her experiments with music discriminating fish, Chase (2001) is

not terse when mentioning their limitations:

> Even a convincing demonstration of categorization can fail to identify the stimulus features that exert control at any given time ... In particular, there can be uncertainty as to whether classification behavior had been under the stimulus control of the features *in terms of which the experimenter had defined the categories* or whether the subjects had discovered an effective discriminant of which the experimenter was unaware. ... [A] constant concern is the possible existence of a simple attribute that would have allowed the subjects merely to discriminate instead of categorizing. (emphasis ours)

Chase thereby recognizes that, with the exception of timbre (experiment 3), her experiments do not validly answer questions about *what* the fish are using in order to make their decisions, or even whether those criteria are related in any way to how the music was categorized in the first place. In fact, her experiments were not designed to answer such a question.

This same limitation exists for all MIR experiments that use *Classify* (Sturm, 2013a,c,b,g). That a MIR system is able to perfectly reproduce the "ground truth" labels of a testing dataset gives no reason to believe the system is using criteria relevant to the meaning of those labels. The existence of many independent variables in the complex signals of datasets like *GTZAN*, and the lack of their control in any evaluation using them, severely limits what one can say from the results. Even if one is to be conservative and restrict the conclusion to only that music in *GTZAN*, or *ISMIR2004*, we have shown (2013g) that even such a conclusion is not valid.[21]

Other limitations come from the measurement model specified by the experimental design. Any estimate of the responses from the measurements in an experiment (Sturm, 2013c) must also accompanied by its quality (confidence interval), and the conditions under which it holds, e.g., "simple textbook model" (Bailey, 2008). (This is not the variance of the measurements!) With random-, fixed- and mixed-effects models, these estimates generally have wider confidence intervals than with the standard textbook model, i.e., the estimates have more uncertainty (Bailey, 2008). In the field of bioinformatics, an evaluation that does not give the confidence interval and conditions under which it holds has been called "scientifically vacuous" (Dougherty and Dalton, 2013).

It is of course beyond the scope of this paper whether or not an artificial system can learn from the few examples in *GTZAN* the meaning behind the various tags shown in Fig. 5; but, that SRCAM is able to achieve a classification accuracy of over 0.75 from only 50 training dataset feature vectors of 768 dimensions in each of 10 classes, is *quite* impressive — almost miraculous. To then claim, "SRCAM is recognizing music genre", is as remarkable a claim as, "Hans the horse can add." Any artificial system purported to be capable of, e.g., identifying thrilling Disco music with male lead vocals that is useful for driving, or romantic Pop music with positive feelings and backing vocals that is useful for getting ready to go out, should be approached with as much skepticism as Pfungst approached Hans (Pfungst, 1911). "An extraordinary claim requires extraordinary proof" (Truzzi, 1978).

[21] What conclusion is valid in this case has yet to be determined.

4.5 Make reproducible work reproducible.

The final implication of our work here is in the reproducibility of experiments. One of the hallmarks of modern science is the reproduction of the results of an experiment by an independent party (with no vested interests). The six "×" in Fig. 2 come from our attempts at reproducing the work in several publications. In (Sturm and Noorzad, 2012), we detail our troubles in reproducing the work of Panagakis et al. (2009b). Through our collaboration with those authors, they found their results in that paper were due to a mistake in the experimental procedure. This also affected other published results of theirs (Panagakis et al., 2009a; Panagakis and Kotropoulos, 2010). Our trouble in reproducing the results of Marques et al. (2011a) led to the discovery that the authors had accidentally used the training dataset as the testing dataset.[22] Our significant trouble in reproducing the results of Bağci and Erzin (2007) led us to analyze the algorithm (Sturm and Gouyon, 2013), which reveals that their results could not have been due to the proper operation of their system. Finally, we have found the approach and results of Chang et al. (2010) to be irreproducible and determined that they are contradicted by several well-established principles (Sturm, 2013f). In only one of these cases (Marques et al., 2011a) did we find the code to be available and accessible enough for easily recreating the published experiments. For the others, we had to make many assumptions and decisions from the beginning, sometimes in consultation with the authors, to ultimately find that the description in the paper does not match what was produced, or that the published results are in error.

Among other things, "reproducible research" (Vandewalle et al., 2009) is motivated by the requirements of modern science, and by the algorithm-centric nature of disciplines like signal processing. Vandewalle et al. (2009) define a piece of research "reproducible" if: "all information relevant to the work, including, but not limited to, text, data and code, is made available, such that an independent researcher can reproduce the results." This obviously places an extraordinary burden on a researcher, and is impossible when data is privately owned; but when work can be reproduced, then it should be reproducible. Making work reproducible can not only increase citations, but also greatly aid in peer review. For instance, the negative results in our work (Sturm and Gouyon, 2013) were greeted by skepticism in peer review; but, because we made entirely reproducible the results of our experiments and the figures in our paper, the reviewers were able to easily verify our analysis, and explore the problem themselves. The results of all our work, and the software to produce them, are available online for anyone else to run and modify.[23] The MIReS Roadmap (Serra et al., 2013) also highlights the need for reproducibility in MIR. Great progress is being achieved by projects such as SoundSoftwatre.[24] Vandewalle et al. (2009) provides an excellent guide of how to make research reproducible.

5 Conclusions

Our "state of the art" here takes a much different approach than the previous four reviews of MGR (Aucouturier and Pachet, 2003; Scaringella et al., 2006; Dannenberg, 2010; Fu et al.,

[22] Personal communication with J. P. Papa.

[23] `http://imi.aau.dk/~bst/software/index.html`

[24] `http://soundsoftware.ac.uk`

2011). It aims not to summarize the variety of features and machine learning approaches used in MGR systems in the ten years since Aucouturier and Pachet (2003), or the new datasets available and how "ground truths" are generated. It is not concerned with the validity, well-posedness, value, usefulness, or applicability of MGR; or whether MGR is "replaced by," or used in the service of, e.g., music similarity, autotagging, or the like. These are comprehensively addressed in other works, e.g., (McKay and Fujinaga, 2006; Craft et al., 2007; Craft, 2007; Wiggins, 2009; Sturm, 2012c,b, 2013a,b,g). It is not concerned with how, or even whether it is possible, to create "faultless" datasets for MGR, music similarity, autotagging, and the like. Instead, this article is concerned with evaluations performed in the past ten years of work in MGR, and its implications for the next ten years of work in MIR. It aims to look closely at whether the results of Fig. 2 imply any progress has been made in *solving* the problem.

If one is to look at Fig. 2 as a "map" showing which pairing of audio features with machine learning algorithms is successful, or which is better than another, or whether progress has been made in charting an unknown land, then one must defend such conclusions as being valid with these evaluations — which are certainly rigorous, systematic, and standardized evaluations, but that may not address the scientific question of interest. One simply cannot take for granted that a systematic, standardized and rigorous evaluation design using a benchmark dataset produces scientific results, let alone results that reflect anything to do with *music intelligence*. Questions of music intelligence cannot be addressed with validity by an experimental design that counts the number of matches between labels output by a system and the "ground truth" of a testing dataset (*Classify*), but that does not account for the numerous independent variables of a testing dataset with flawed integrity. Only a few works have sought to determine whether the decisions made by an MGR system come from anything at all relevant to genre (Sturm, 2012c, 2013a), e.g., musicological dimensions such as instrumentation, rhythm (Dixon et al., 2004), harmony (Anglade et al., 2009), compositional form, and lyrics (Mayer et al., 2008). Metaphorically, whereas only a few have been using a tape measure, a level and a theodolite to survey the land, the points of Fig. 2 have been found using microscopes and scales — certainly scientific instruments, but irrelevant to answering the questions of interest. It may be a culmination of a decade of work, but Fig. 2 is no map of reality.

These problems do not just exist for research in MGR, but are also seen in research on music emotion recognition (Sturm, 2013c,b), and autotagging (Marques et al., 2011b; Gouyon et al., 2013). These kinds of tasks constitute more than half of the evaluation tasks in MIREX since 2005, and are, in essence, representative of the IR in MIR. How, then, has this happened, and why does it continue? First, as we have shown for *GTZAN* (Sturm, 2012a, 2013d), many researchers assume a dataset is a good dataset because many others use it. Second, as we have shown for *Classify* (Sturm, 2013c,b), many researchers assume evaluation that is standard in machine learning or information retrieval, or that is implemented to be systematic and rigorous, is thus relevant for MIR and is scientific. Third, researchers fail to define proper use cases, so problems and success criteria remain by and large ill-defined — or, as in the case of *GTZAN*, defined by and large by the artificial construction of a dataset. Fourth, the foundation of evaluation in MIR remains obscure because it lacks an explicit formalism in its design and analysis of systems and experiments (Urbano et al., 2013; Sturm, 2013c). Hence, evaluations in all MIR are accompanied by

many assumptions that remain implicit, but which directly affect the conclusions, relevance, and validity of experiments. Because these assumptions remain implicit, researchers fail to acknowledge limitations of experiments, and are persuaded that their solutions are actually addressing the problem, or that their evaluation is measuring success — the *Clever Hans Effect* (Pfungst, 1911; Sturm, 2013g). While we are not claiming that all evaluation in MIR lacks validity, we do claim there is a "crisis of evaluation" in MIR more severe than what is reflected by the MIReS Roadmap (Serra et al., 2013). The single most important

The situation in MIR is reminiscent of computer science in the late 1970s. In an editorial from 1979 (McCracken et al., 1979), the President, Vice-President and Secretary of the Association for Computing Machinery acknowledge that "experimental computer science" (to distinguish it from theoretical computer science) is in a crisis in part because there is "a severe shortage of computer scientists engaged in, or qualified for, experimental research in computer science." This led practitioners of the field to define more formally the aims of experimental computer science, how they relate to the scientific pursuit of knowledge (Denning, 1980), and find model examples of scientific practice in their field (Denning, 1981). Six years later, Basili et al. (1986) contributed a review of fundamentals in the design and analysis of experiments, how it has been applied to and benefited software engineering, and suggestions for continued improvement. Almost a decade after this, however, Fenton et al. (1994) argue "there are far too few examples of moderately effective research in software engineering ... much of what we believe about which approaches are best is based on anecdotes, gut feelings, expert opinions, and flawed research, not on careful, rigorous software-engineering experimentation." This is not a surprise to Fenton, however, since "topics like experimental design, statistical analysis, and measurement principles" remain neglected in contemporary computer science educations (Fenton et al., 1994). The conversation is still continuing (Feitelson, 2006).

In summary, even statistics is not immune to these problems (Hand, 1994). Donald Preece provides a stratospherically high-level of summary of Hand (1994): "[it speaks on] the questions that the researcher wishes to consider ...: (a) how do I obtain a statistically significant result?; (b) how do I get my paper published?; (c) when will I be promoted?" There are no shortcuts to the design, implementation, and analysis of a valid experiment. It requires hard work, creativity, and intellectual postures that are uncomfortable for many. Good examples abound, however. Chase (2001) shows the time and effort necessary to scientifically and efficiently answer real questions, and exemplifies the kinds of considerations that are taken for granted in disciplines where thousands of experiments can be run in minutes on machines that do not currently need positive reinforcement. Just as Pfungst does with Hans (Pfungst, 1911), Chase scientifically tests whether Oro, Beauty and Pepi are actually discriminating styles of music. Neither Pfungst nor Chase give their subjects free passes of accountability. Why, then, should it be any different for artificial systems?

Acknowledgments

Thanks to: Fabien Gouyon, Nick Collins, Arthur Flexer, Mark Plumbley, Geraint Wiggins, Mark Levy, Roger Dean, Julián Urbano, Alan Marsden, Lars Kai Hansen, Jan Larsen, Mads G. Christensen, Sergios Theodoridis, Aggelos Pikrakis, Dan Stowell, Rémi Gribonval, Geoffrey Peeters, Diemo Schwarz, Roger Dannenberg, Bernard Mont-Reynaud, Gaël Richard,

Rolf Bardeli, Jort Gemmeke, Curtis Roads, Stephen Pope, George Tzanetakis, Constantine Kotropoulos, Yannis Panagakis, Ulaş Bağci, Engin Erzin, and João Paulo Papa for illuminating discussions about these topics (which does not mean any endorse the ideas herein). Mads G. Christensen, Nick Collins, Cynthia Liem, and Clemens Hage helped identify several excerpts in *GTZAN*, and my wife Carla Sturm endured my repeated listening to all of its excerpts. Thanks to the many, many associate editors and anonymous reviewers for the comments that helped move this work closer and closer to being publishable.

References

Ammer, C. (2004). *Dictionary of Music*. The Facts on File, Inc., New York, NY, USA, 4 edition.

Andén, J. and Mallat, S. (2011). Multiscale scattering for audio classification. In *Proc. ISMIR*, pages 657–662.

Andén, J. and Mallat, S. (2013). Deep scattering spectrum. *http://arxiv.org/abs/1304.6763*.

Anglade, A., Ramirez, R., and Dixon, S. (2009). Genre classification using harmony rules induced from automatic chord transcriptions. In *Proc. ISMIR*.

Aucouturier, J.-J. (2009). Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. In Minett, J. and Wang, W., editors, *Language, Evolution and the Brain: Frontiers in Linguistic Series*. Academia Sinica Press.

Aucouturier, J.-J. and Bigand, E. (2013). Seven problems that keep MIR from attracting the interest of cognition and neuroscience. *J. Intell. Info. Systems*.

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

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

Bailey, R. A. (2008). *Design of comparative experiments*. Cambridge University Press.

Barbedo, J. G. A. and Lopes, A. (2007). Automatic genre classification of musical signals. *EURASIP J. Adv. Sig. Process.*

Barreira, L., Cavaco, S., and da Silva, J. (2011). Unsupervised music genre classification with a model-based approach. In *Proc. Portugese Conf. Progress Artificial Intell.*, pages 268–281.

Barrington, L., Yazdani, M., Turnbull, D., and Lanckriet, G. R. G. (2008). Combining feature kernels for semantic music retrieval. In *ISMIR*, pages 614–619.

Basili, V., Selby, R., and Hutchens, D. (1986). Experimentation in software engineering. *IEEE Trans. Software Eng.*, 12(7):733–743.

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

Bergstra, J., Casagrande, N., Erhan, D., Eck, D., and Kégl, B. (2006). Aggregate features and AdaBoost for music classification. *Machine Learning*, 65(2-3):473–484.

Bertin-Mahieux, T., Eck, D., Maillet, F., and Lamere, P. (2008). Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135.

Bertin-Mahieux, T., Eck, D., and Mandel, M. (2010). Automatic tagging of audio: The state-of-the-art. In Wang, W., editor, *Machine Audition: Principles, Algorithms and Systems*. IGI Publishing.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., and Lamere, P. (2011). The million song dataset. In *Proc. ISMIR*.

Bowker, G. C. and Star, S. L. (1999). *Sorting things out: Classification and its consequences*. The MIT Press.

Chang, K., Jang, J.-S. R., and Iliopoulos, C. S. (2010). Music genre classification via compressive sampling. In *Proc. ISMIR*, pages 387–392, Amsterdam, The Netherlands.

Chase, A. (2001). Music discriminations by carp "Cyprinus carpio". *Animal Learning & Behavior*, 29(4):336–353.

Craft, A. (2007). The role of culture in the music genre classification task: human behaviour and its effect on methodology and evaluation. Technical report, Queen Mary University of London.

Craft, A., Wiggins, G. A., and 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. ISMIR*.

Cunningham, S. J., Bainbridge, D., and Downie, J. S. (2012). The impact of MIREX on scholarly research. In *Proc. ISMIR*.

Dannenberg, R. B. (2010). Style in music. In Argamon, S., Burns, K., and Dubnov, S., editors, *The Structure of Style*, pages 45–57. Springer Berlin Heidelberg.

Dannenberg, R. B., Thom, B., and Watson, D. (1997). A machine learning approach to musical style recognition. In *Proc. ICMC*, pages 344–347.

Denning, P. J. (1980). What is experimental computer science? *Communications of the ACM*, 23(10):543–544.

Denning, P. J. (1981). Performance analysis: experimental computer science as its best. *Communications of the ACM*, 24(11):725–727.

Dixon, S., Gouyon, F., and Widmer, G. (2004). Towards characterisation of music via rhythmic patterns. In *Proc. ISMIR*, pages 509–517, Barcelona, Spain.

Dougherty, E. R. and Dalton, L. A. (2013). Scientific knowledge is possible with small-sample classification. *EURASIP J. Bioinformatics and Systems Biology*, 10.

Downie, J., Ehmann, A., Bay, M., and Jones, M. (2010). The music information retrieval evaluation exchange: Some observations and insights. In Ras, Z. and Wieczorkowska, A., editors, *Advances in Music Information Retrieval*, pages 93–115. Springer Berlin / Heidelberg.

Downie, J. S. (2003). Toward the scientific evaluation of music information retrieval systems. In *ISMIR*, Baltimore, USA.

Downie, J. S. (2004). The scientific evaluation of music information retrieval systems: Foundations and future. *Computer Music Journal*, 28(2):12–23. Cited By (since 1996): 12.

Downie, J. S. (2008). The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. *Acoustical Science and Tech.*, 29(4):247–255.

Duin, R. P. W., Juszczak, P., de Ridder, D., Paclik, P., Pekalska, E., and Tax, D. M. J. (2007). PR-Tools4.1, a matlab toolbox for pattern recognition. Delft University of Technology. `http://prtools.org`.

Fabbri, F. (1980). A theory of musical genres: Two applications. In *Proc. Int. Conf. Popular Music Studies*, Amsterdam, The Netherlands.

Fabbri, F. (1999). Browsing musical spaces. In *Proc. IASPM*, Amsterdam, The Netherlands.

Feitelson, D. G. (2006). Experimental computer science: The need for a cultural change. Technical report, The Hebrew University of Jerusalem.

Fenton, N., Pfleeger, S. L., and Glass, R. L. (1994). Science and substance: a challenge to software engineers. *IEEE Software*, 11(4):86–95.

Flexer, A. (2006). Statistical evaluation of music information retrieval experiments. *J. New Music Research*, 35(2):113–120.

Flexer, A. (2007). A closer look on artist filters for musical genre classification. In *Proc. ISMIR*, Vienna, Austria.

Flexer, A. and Schnitzer, D. (2009). Album and artist effects for audio similarity at the scale of the web. In *Proc. SMC*, pages 59–64, Porto, Portugal.

Flexer, A. and Schnitzer, D. (2010). Effects of album and artist filters in audio similarity computed for very large music databases. *Computer Music J.*, 34(3):20–28.

Frow, J. (2005). *Genre*. Routledge, New York, NY, USA.

Fu, Z., Lu, G., Ting, K. M., and Zhang, D. (2011). A survey of audio-based music classification and annotation. *IEEE Trans. Multimedia*, 13(2):303–319.

Gjerdingen, R. O. and Perrott, D. (2008). Scanning the dial: The rapid recognition of music genres. *J. New Music Research*, 37(2):93–100.

Gouyon, F., Sturm, B. L., Oliveira, J. L., Hespanhol, N., and Langlois, T. (2013). On evaluation in music autotagging research. *(submitted)*.

Guaus, E. (2009). *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers*. PhD thesis, Universitat Pompeu Fabra, Barcelona, Spain.

Hand, D. J. (1994). Deconstructing statistical questions. *J. Royal Statist. Soc. A (Statistics in Society)*, 157(3):317–356.

Hartmann, M. A. (2011). Testing a spectral-based feature set for audio genre classification. Master's thesis, University of Jyväskylä.

Humphrey, E. J., Bello, J. P., and LeCun, Y. (2013). Feature learning and deep architectures: New directions for music informatics. *J. Intell. Info. Systems*, 41(3):461–481.

Kemp, C. (2004). Towards a holistic interpretation of musical genre classification. Master's thesis, University of Jyväskylä.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *J. American Statistical Assoc.*, 52(278):133–142.

Law, E. (2011). Human computation for music classification. In Li, T., Ogihara, M., and Tzanetakis, G., editors, *Music Data Mining*, pages 281–301. CRC Press.

Levy, M. and Sandler, M. (2009). Music information retrieval using social tags and audio. *Multimedia, IEEE Transactions on*, 11(3):383 –395.

Li, M. and Sleep, R. (2005). Genre classification via an LZ78-based string kernel. In *Proc. ISMIR*.

Li, T. and Chan, A. (2011). Genre classification and the invariance of MFCC features to key and tempo. In *Proc. Int. Conf. MultiMedia Modeling*, Taipei, China.

Li, T. and Ogihara, M. (2006). Toward intelligent music information retrieval. *IEEE Trans. Multimedia*, 8(3):564–574.

Lidy, T. (2006). Evaluation of new audio features and their utilization in novel music retrieval applications. Master's thesis, Vienna University of Tech.

Lidy, T. and Rauber, A. (2005). Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In *Proc. ISMIR*.

Lippens, S., Martens, J., and De Mulder, T. (2004). A comparison of human and automatic musical genre classification. In *Proc. ICASSP*, pages 233–236.

Mandel, M. and Ellis, D. P. W. (2005). Song-level features and support vector machines for music classification. In *Proc. Int. Symp. Music Info. Retrieval*, London, U.K.

Markov, K. and Matsui, T. (2012a). Music genre classification using self-taught learning via sparse coding. In *Proc. ICASSP*, pages 1929 –1932.

Markov, K. and Matsui, T. (2012b). Nonnegative matrix factorization based self-taught learning with application to music genre classification. In *Proc. IEEE Int. Workshop Machine Learn. Signal Process.*, pages 1–5.

Marques, C., Guiherme, I. R., Nakamura, R. Y. M., and Papa, J. P. (2011a). New trends in musical genre classification using optimum-path forest. In *Proc. ISMIR*.

Marques, G., Domingues, M., Langlois, T., and Gouyon, F. (2011b). Three current issues in music autotagging. In *Proc. ISMIR*.

Matityaho, B. and Furst, M. (1995). Neural network based model for classification of music type. In *Proc. Conv. Electrical and Elect. Eng. in Israel*, pages 1–5.

Mayer, R., Neumayer, R., and Rauber, A. (2008). Rhyme and style features for musical genre classification by song lyrics. In *Proc. ISMIR*.

McCracken, D. D., Denning, P. J., and Brandin, D. H. (1979). An ACM executive committee position on the crisis in experimental computer science. *Communications of the ACM*, 22(9):503–504.

McKay, C. and Fujinaga, I. (2006). Music genre classification: Is it worth pursuing and how can it be improved? In *Proc. ISMIR*, Victoria, Canada.

McLuhan, M. (1964). *Understanding Media: The Extensions of Man*. The MIT Press, 94 edition.

Moerchen, F., Mierswa, I., and Ultsch, A. (2006). Understandable models of music collections based on exhaustive feature generation with temporal statistics. In *Int. Conf. Knowledge Discover and Data Mining*.

Noorzad, P. and Sturm, B. L. (2012). Regression with sparse approximations of data. In *Proc. EUSIPCO*, Bucharest, Romania.

Pachet, F. and Cazaly, D. (2000). A taxonomy of musical genres. In *Proc. Content-based Multimedia Information Access Conference*, Paris, France.

Pampalk, E. (2006). *Computational Models of Music Similarity and their Application in Music Information Retrieval*. PhD thesis, Vienna University of Tech., Vienna, Austria.

Pampalk, E., Flexer, A., and Widmer, G. (2005). Improvements of audio-based music similarity and genre classification. In *Proc. ISMIR*, pages 628–233, London, U.K.

Panagakis, Y. and Kotropoulos, C. (2010). Music genre classification via topology preserving non-negative tensor factorization and sparse representations. In *Proc. ICASSP*, pages 249–252.

Panagakis, Y., Kotropoulos, C., and Arce, G. R. (2009a). Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In *Proc. ISMIR*, pages 249–254, Kobe, Japan.

Panagakis, Y., Kotropoulos, C., and Arce, G. R. (2009b). Music genre classification via sparse representations of auditory temporal modulations. In *Proc. EUSIPCO*.

Park, S., Park, J., and Sim, K. (2011). Optimization system of musical expression for the music genre classification. In *Proc. Int. Conf. Control, Auto. Syst.*, pages 1644–1648.

Pfungst, O. (1911). *Clever Hans (The horse of Mr. Von Osten): A contribution to experimental animal and human psychology*. Henry Holt, New York.

Porter, D. and Neuringer, A. (1984). Music discriminations by pigeons. *Experimental Psychology: Animal Behavior Processes*, 10(2):138–148.

Salzberg, S. L. (1997). *Data mining and knowledge discovery*, chapter On comparing classifiers: Pitfalls to avoid and a recommended approach, pages 317–328. Kluwer Academic Publishers.

Scaringella, N., Zoia, G., and Mlynek, D. (2006). Automatic genre classification of music content: A survey. *IEEE Signal Process. Mag.*, 23(2):133–141.

Serra, X., Magas, M., Benetos, E., Chudy, M., Dixon, S., Flexer, A., Gómez, E., Gouyon, F., Herrera, P., Jordà, S., Paytuvi, O., Peeters, G., Schlüter, J., Vinet, H., and Widmer, G. (2013). *Roadmap for Music Information ReSearch*. Creative Commons.

Seyerlehner, K. (2010). *Content-based Music Recommender Systems: Beyond Simple Frame-level Audio Similarity*. PhD thesis, Johannes Kepler University, Linz, Austria.

Seyerlehner, K., Widmer, G., and Pohle, T. (2010). Fusing block-level features for music similarity estimation. In *DAFx*.

Shapiro, P. (2005). *Turn the Beat Around: The Secret History of Disco*. Faber & Faber, London, U.K.

Slaney, M. (1998). Auditory toolbox. Technical report, Interval Research Corporation.

Sordo, M., Celma, O., Blech, M., and Guaus, E. (2008). The quest for musical genres: Do the experts and the wisdom of crowds agree? In *Proc. ISMIR*.

Sturm, B. L. (2012a). An analysis of the GTZAN music genre dataset. In *Proc. ACM MIRUM Workshop*, Nara, Japan.

Sturm, B. L. (2012b). A survey of evaluation in music genre recognition. In *Proc. Adaptive Multimedia Retrieval*, Copenhagen, Denmark.

Sturm, B. L. (2012c). Two systems for automatic music genre recognition: What are they really recognizing? In *Proc. ACM MIRUM Workshop*, Nara, Japan.

Sturm, B. L. (2013a). Classification accuracy is not enough: On the evaluation of music genre recognition systems. *J. Intell. Info. Systems*, 41(3):371–406.

Sturm, B. L. (2013b). Evaluating music emotion recognition: Lessons from music genre recognition? In *Proc. ICME*.

Sturm, B. L. (2013c). Formalizing evaluation in music information retrieval: A look at the mirex automatic mood classification task. In *Proc. CMMR*.

Sturm, B. L. (2013d). The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use. *http://arxiv.org/abs/1306.1461*.

Sturm, B. L. (2013e). Music genre recognition with risk and rejection. In *Proc. ICME*.

Sturm, B. L. (2013f). On music genre classification via compressive sampling. In *Proc. ICME*.

Sturm, B. L. (2013g). A simple method to determine if a music information retrieval system is a horse. *submitted*.

Sturm, B. L. and Gouyon, F. (2013). Revisiting inter-genre similarity. *IEEE Sig. Proc. Letts.*, 20(11):1050 – 1053.

Sturm, B. L. and Noorzad, P. (2012). On automatic music genre recognition by sparse representation classification using auditory temporal modulations. In *Proc. CMMR*, London, UK.

Theodoridis, S. and Koutroumbas, K. (2009). *Pattern Recognition*. Academic Press, Elsevier, Amsterdam, The Netherlands, 4 edition.

Truzzi, M. (1978). On the extraordinary: An attempt at clarification. *Zetetic Scholar*, 1:11–22.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.*, 10(5):293–302.

Urbano, J. (2011). Information retrieval meta-evaluation: Challenges and opportunities in the music domain. In *ISMIR*, pages 609–614.

Urbano, J., McFee, B., Downie, J. S., and Schedl, M. (2012). How significant is statistically significant? the case of audio music similarity and retrieval. In *Proc. ISMIR*.

Urbano, J., Mónica, M., and Morato, J. (2011). Audio music similarity and retrieval: Evaluation power and stability. In *Proc. ISMIR*, pages 597–602.

Urbano, J., Schedl, M., and Serra, X. (2013). Evaluation in music information retrieval. *J. Intell. Info. Systems (in press)*.

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

Vandewalle, P., Kovacevic, J., and Vetterli, M. (2009). Reproducible research in signal processing — what, why, and how. *IEEE Signal Process. Mag.*, 26(3):37–47.

Wang, A. (2003). An industrial strength audio search algorithm. In *Proc. Int. Soc. Music Info. Retrieval*, Baltimore, Maryland, USA.

Wiggins, G. A. (2009). Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music. In *Proc. IEEE Int. Symp. Mulitmedia*, pages 477–482.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 31(2):210–227.

Xia, G., Tay, J., Dannenberg, R., and Veloso, M. (2012). Autonomous robot dancing driven by beats and emotions of music. In *Proc. Int. Conf. Autonomous Agents Multiagent Syst.*, pages 205–212, Richland, SC.