Evaluating music emotion recognition

Lessons from music genre recognition?

Sturm, Bob L.

Published in:
International Conference on Multimedia and Expo

Publication date:
2013

Document Version
Early version, also known as pre-print

EVALUATING MUSIC EMOTION RECOGNITION: LESSONS FROM MUSIC GENRE RECOGNITION?

Bob L. Sturm

Audio Analysis Lab, AD:MT, Aalborg University Copenhagen, A.C. Meyers Vænge 15, DK-2450

ABSTRACT

A fundamental problem with nearly all work in music genre recognition (MGR) is that evaluation lacks validity with respect to the principal goals of MGR. This problem also occurs in the evaluation of music emotion recognition (MER). Standard approaches to evaluation, though easy to implement, do not reliably differentiate between recognizing genre or emotion from music, or by virtue of confounding factors in signals (e.g., equalization). We demonstrate such problems for evaluating an MER system, and conclude with recommendations.

***Index Terms*—** Evaluation, music genre recognition, music emotion recognition, machine learning

1. INTRODUCTION

We discuss, specifically for music genre recognition (MGR), fundamental problems with standard approaches to evaluation — experimental designs, data, and figures of merit — and relate them to evaluation in music emotion recognition (MER) [1–4]. Training machines to recognize emotions from music resembles training machines to recognize genres used by music: the significant amount of subjectivity [5]; the debates about whether genre is even *in* the music signal [6]; the assumed existence of abstract categories that are difficult if not impossible to systematize, to define directly with a set of clear unambiguous rules, or even to define indirectly by exemplars that are indisputably representative [7]; ground truths that are difficult if not impossible to generate for datasets that are deceptively simple to assemble; and the difficulty of evaluation [8]. Genre and mood have been shown to be correlated in some cases [1, 9]; and some systems proposed for MER are just adapted from MGR [1, 2]. Thus, since MGR is one of the most studied areas in music information research [2, 10] — MGR has been called its "flagship application" [11] — we argue that the challenges MGR research faces in evaluation inform those faced in MER.

In this paper, we do not argue whether MER is ill-defined or not, whether it has value or not, whether music induces, arouses, evokes, or conveys mood or not [12]. We do not aim to provide a state-of-the-art of MER, which is already presented excellently in [1–4]. We only address the fact that researchers are and have been pursuing the development of MER systems, that MER has been a part of MIREX since 2007 [13], and that the approaches to evaluation used, and the conclusions drawn from it, are similar to those in MGR. The fundamental problem is that the majority of this evaluation lacks the validity for making any useful conclusion.

In the next section, we articulate the principal goals of MGR, and discuss how standard approaches to evaluation in MGR research do not have validity to address them. In Section 3, we adapt this discussion to MER, and discuss specific conclusions drawn from such evaluation. The fourth section provides an illustration of problems with evaluation in MER. We conclude by outlining new approaches to evaluation.

2. PROBLEMS IN MGR EVALUATION

We define the principal goals of MGR, and summarize our findings about MGR evaluation [10]. Finally, we show how most evaluation in MGR lacks validity with respect to the principal goals of MGR.

2.1. What are the principal goals of MGR?

In the MGR literature [10], we find very few explicit definitions of MGR. A straightforward one, summarizing [2, 14], is to select genre labels appropriate for describing particular music. Another definition, summarizing [15, 16], is to automatically classify an audio signal into a taxonomy of musical genre. It is debatable, however, if musical genres can be usefully expressed by a taxonomy [5, 7]. An entirely different definition comes from perhaps the first work of automatic MGR [17]: "to build a phenomenological model that will imitate the human ability to distinguish between music [genres]." This is broad (it is not only about reproducing labels), contains aspects of the others above, makes no assumption of how genre can be organized, and prescribes the involvement of the ingredient necessary for the existence of genre: humans [5].

From these then, the principal goals of MGR are *to imitate the human ability to organize, recognize, distinguish between, and imitate genres of music.* We define "organize" as to find and express characteristics of genres used by some music. We define "recognize" as to identify the genres used by some music, or to relate unknown genres to known ones. We define "distinguish between" as to describe why, or the

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

extents to which, some music uses some genres but not others. Finally, we define "imitate" as to exemplify the use of particular genres, e.g., perform a piece as if it uses Metal.

A critical point is that the principal goals of MGR do not prescribe decisions are to be made in the same ways as humans, but only that a system *imitates the human ability* to make those decisions. This encompasses more than reproducing labels for a set of music. It means, at the very least, decisions should be based upon criteria relevant to music genre, not aspects correlated somehow with labels in some dataset.

2.2. Evaluation trends in MGR

When proposing an MGR system (feature, machine learning, and train data), one must evaluate it by an experimental design, test data, and figure of merit. Evaluation in MGR is a critical aspect that has been and continues to be overlooked; indeed, there is little if anything said in reviews of MGR [2,14,16], and we find in our survey [10] that most published work uses one experimental design and singly-labeled and private data. We now summarize our findings in [10].

The experimental design most used in MGR, appearing in over 91% of experimental work, is that of comparing the discrete labels selected by a system to the "ground truth" of the test data, which we name *Classify*. Other experimental designs include selecting and comparing features, and testing and comparing systems across different test data. The least used experimental design is having a system compose music, and then using a listening test to determine how representative the excerpts are of particular music genres.

Over 58% of work in MGR uses private data, which makes the reproduction of experiments near impossible. The most-used public dataset is GTZAN [15], which appears in about 23% of MGR work since its creation in 2002. We have shown GTZAN to suffer from several faults [18], such as exact replicas, mislabelings, and distortions. We have shown that these impact evaluation in non-trivial ways such that any meaningful comparison cannot be made [19].

As *Classify* is the experimental design most used in MGR, it is not unexpected that the figure of merit most used — reported in 82% of work — is classification accuracy. Related to this are precision, recall, and F-score. Confusion tables appear in 32% of work, and are accompanied by reflection on the observed behaviors only half of the time.

2.3. Validity in MGR evaluation

Results of *Classify* on benchmark datasets are typically used to compare MGR systems and track research progress, e.g., the surveys in [2, 14, 16], and the results of several MIREX challenges since 2007 [13]. Since we now see MGR system classification accuracies rivaling that reported for humans [20], one might conclude that great progress in MGR has been made. With respect to its principal goals, however, such a conclusion is not supported by the evaluation.

Consider that one wishes to evaluate (experimental design, test data, figure of merit) the extent to which an MGR system (feature, machine learning, train data) imitates the human ability to recognize genre. When the evaluation has *internal validity* [21,22], it provides a measure of how well the system is able to recognize the genres used by music in the test data. When the evaluation has *external validity* [21,22], it provides a measure of how well the same system will recognize the genres used by music in any other test data, e.g., the real world. When an evaluation lacks internal and external validity, one cannot conclude from it whether an MGR system is even recognizing genre at all.

A fundamental problem that compromises the validity of *Classify* in MGR evaluation is the lack of control for independent variables in data. This is illustrated by the following example [23]. A system was developed to detect the presence of military tanks from a photograph. The developers assembled a dataset, and carefully split it in two so that the train and test data did not overlap. System performance was so good that it aroused suspicion. The developers assembled new test data, and found the same system performed very poor. A richer evaluation of the system found it to be discriminating using factors related to the weather. In the initial test and train data, all photographs of one condition were taken on a sunny day, while the others were taken on a cloudy day. Since *Classify* with the original data lacks control of the independent variables, its results have no internal and external validity to conclude whether the system even addresses the basic problem.

For the 91% of published work in MGR using *Classify* with data having uncontrolled independent variables, e.g., fidelity, bandwidth, dynamic compression, loudness, etc., none of it has validity for concluding whether any of thee systems are recognizing genre at all, whether any is better than another, or even whether any address the principal goals of MGR. *Classification accuracy is not enough* [8, 19]. Recalls, precisions, F-scores, confusions, and formal statistical comparisons between systems are still not enough [13]. Even if one sees 100% classification accuracy in *Classify*, as long as independent variables in the data are not controlled, one cannot logically claim from this that the system is making decisions based on the musical content in a signal. As for the system built to detect tanks, testing what is has actually learned, why it behaves as it does, and how it is making decisions, requires different approaches to evaluation.

Hardly any work closely examines the behaviors, robustness, and internal models of MGR systems. The results of Marques et al. [24] argue against the claim that low-level features of musical audio signals, e.g., features computed from 23 ms, carry genre-indicative information. In [8, 19, 25], we show that even though a system might have a high classification accuracy, its actual behavior can reveal it must not be making decisions based on the genre of the music, but based on factors in the data confounded with the labels. This argues against the claim that an MGR system achieving high performance statistics is doing so by recognizing genre, and explains why MGR systems can have high classification ac-

curacies, but behave in poor ways (not human-like), e.g., persistently labeling as metal "Mamma Mia" by ABBA [19, 25].

In summary, most evaluation in the MGR literature lacks the validity to argue that a system is addressing a principal goal of MGR. In the 467 papers we survey in MGR [10], we find no discussion about validity in evaluation. Indeed, as we show in [19], the results of 96 systems that have since 2002 been proposed and tested by *Classify* in GTZAN cannot be meaningfully compared *in any sense*. (Among these, six results come from an error in the evaluation implementation, or something else. For details, see [8, 26, 27].) A richer palette of evaluation must be used to uncover the mechanisms MGR systems employ, and thus to separate those that are really working from those that are just 'Clever Hans' [28].

3. EVALUATION IN MER

The reviews in [1–4] provide an excellent survey of MER. They suggest its principal goals are to automatically organize, label, and retrieve music according to emotion. Hence, it is necessary to evaluate MER systems in ways that reliably measure the extent to which they address these goals. None of these reviews [1–4], however, discuss valid evaluation. Certainly, all of them discuss the variety of difficulties faced when creating the "ground truth" of a dataset; but none discuss experimental designs for evaluating MER systems in ways that are valid with respect to the principal goals.

Regardless of whether an MER system is devised around a categorical or dimensional model of emotion — one difference being discrete or continuous labels — it appears from [1–4] that most evaluation uses *Classify*. (Regression, used to evaluate MER systems built upon a dimensional model of emotion, is just *Classify* with continuous labels.) Since *Classify* in a dataset having uncontrolled independent variables does not provide a valid way to evaluate MGR systems with respect to its principal goals, it can not be valid for measuring the extent to which an MER system can organize, label, and retrieve music based on emotion — a similarly vague and human-centered concept as genre. No matter how good classification accuracy is, or recall, precision, F-score or confusions, one cannot logically conclude from *Classify* in test data with uncontrolled independent variables that an MER system is working by virtue of recognizing emotion. One can only conclude how well an MER system can reproduce "ground truth" labels (discrete or continuous) of the test data, whether or not by factors completely irrelevant to emotion in music.

Since 2007, the MIREX automatic mood classification challenge (MIREX AMC) [13, 29] has employed *Classify*, a dataset of 600 singly-labeled music excerpts, and several figures of merit, including classification accuracy, confusions, and formal statistical testing for significant differences between MER systems. Hu et al. [29] describe creating the dataset from CDs of a professional music distribution service. Hu et al. appear to take some care to control for genre in each cluster (which itself carries many independent variables), but not to control for similarities of production, instrumentation, artist replication, and other independent variables. Hence, MIREX AMC as designed does not provide valid evaluation with respect to the principal goals of MER. Though it provides systematic evaluation, involves formal statistical tests, and uses a human-validated "ground truth," the uncontrolled independent variables in the test data make the results of *Classify* irrelevant to conclude what system is good, or better than another, for labeling music by emotion.

Nonetheless, showing just how subtle the problem of validity really is, the four reviews of MER [1–4] all contain such conclusions about MER systems and their components. Barthet et al. [4] write a particular MER system "is promising since [it] achieved a performance increase of approximately 20% points (60.6%) in comparison with [another] proposed at MIREX 2009." Kim et al. [1] write, "the highest performing systems in [MIREX AMC] demonstrate improvement each year using solely acoustic features. ... [It] is clear that rapid progress is being made. In the past 5 years, the performance of automated systems ... have advanced significantly." Kim et al. [1] state, "the most successful systems combine multiple acoustic feature types." Yang and Howard [3] state, "Empirical evaluations show that the incorporation of [high-level] features improves the accuracy of emotion recognition." Barthet et al. [4] state, "Timbre features have shown to provide the best performance in MER systems when used as individual features." Fu et al. [2] state, "[rhythmic information plays] a much more important role in mood classification ... than genre and other classification tasks." Not one of these conclusions are supported by the evidence of *Classify* in test data having uncontrolled independent variables.

Another subtle problem here is the conflation of performance and relevance in feature selection experiments. In their work, Song et al. [30] test a single-label categorical MER system based on a variety of features and their combinations using *Classify* and test data with uncontrolled independent variables: "Spectral features outperform [features] based on rhythm, dynamics, and, to a lesser extent, harmony." While the "spectral features" may reproduce the most "ground truth" labels in a specific test data, *Classify* with data having uncontrolled independent variables provides no evidence to conclude specific features are relevant to the task of recognizing emotion. We illustrate this with the following example.

Consider one argues that the intelligence of members of a population and its chocolate consumption are related, since there is a strong correlation between the estimated chocolate consumption in each of 22 countries and their number of Nobel laureates per capita [31]. One might even find chocolate consumption a better predictor of the number of Nobel laureates in these countries than many other factors, e.g., average age, literacy rates, etc. This, however, does not mean one is relevant to the other, either in this data (no internal validity), or in the real world (no external validity). In the same way, one cannot claim by *Classify* with data having uncontrolled independent variables that one feature is more relevant than

another for predicting emotion in music just because classification accuracy is higher when using it.

We do not aim to make examples of the authors above, but only to show how easy it is to be persuaded that standard approaches to evaluation provide valid ways to measure the performance of a MER or MGR system with respect to their principal goals. For an "algorithm-handler" to be persuaded to believe an artificial system has actually learned to recognize and discriminate between complex human notions in products of human culture is known as the "Clever Hans Effect" [28]. Instead of responding to involuntary gestures given by the body, the system is responding according to irrelevant cues unknowingly embodied in data. One cannot make a valid conclusion with respect to the principal goals of MER using *Classify* and data with uncontrolled independent variables.

4. A DEMONSTRATION

We have above essentially condemned the approach to evaluation used most widely in both MGR and MER. In this section, we demonstrate an alternative: a simple way to test whether or not an MER system is making decisions based on confounding factors. We use a simple system trained to discriminate between two mutually exclusive emotions, and evaluate it using *Classify* and *Robust* [10], and a benchmark dataset. One might argue that the system we use is not state-of-the-art in MER; however, our aim here is only to demonstrate an alternative approach to evaluating any MER (or MGR) system.

4.1. System (machine learning, features, data)

Our system uses sparse representation classification of auditory temporal modulations (SRCAM), which was originally proposed for recognizing music genres [32], but then adapted for music autotagging [33]. We take a similar approach here, but do not use tensors, make several adjustments to the algorithm, and restrict it to use only two mutually exclusive emotion tags. The system is essentially described in [8].

Given a matrix of N features $\mathbf{D} := [\mathbf{d}_1|\mathbf{d}_2|\cdots|\mathbf{d}_N]$, and the set of class identities $\cup_{k=1}^{K}\mathcal{I}_k = \{1,\ldots,N\}$, where $\mathcal{I}_k$ specifies the columns of $\mathbf{D}$ related to class k, SRCAM finds a representation of an unlabeled feature $\mathbf{x}$ by solving $\min \|\mathbf{a}\|_1$ subject to $\|\mathbf{x}-\mathbf{D}\mathbf{a}\|_2 < \epsilon$ for $\epsilon \geq 0$. SRCAM then defines a set of weights $\{\mathbf{a}_k\}$ where $\mathbf{a}_k$ are the weights in a that are related to the features in $\mathbf{D}$ from class k. When there are several feature vectors and solutions for an observation $\mathcal{X} = \{(\mathbf{x}^{(i)}, \mathbf{a}^{(i)})\}$, SRCAM classifies the set by

$$\hat{k}(\mathcal{X}) := \arg\min_k \sum_{i=1}^{|\mathcal{X}|} \|\mathbf{x}^{(i)} - \mathbf{D}\mathbf{a}_k^{(i)}\|_2^2. \qquad (1)$$

To find sparse solutions, we use SGPL1 [34] with at most 200 steps, and $\epsilon = 0.1$. Our features are auditory temporal modulations [32], computed as we describe in [8].

As in [33], we use the CAL500 dataset [35]: 502 songs, annotated with a variety of tags (e.g., emotion, genre, and usage) selected by at least two people. We consider only two classes. The first, "happy", has 116 songs tagged "Emotion-Happy" *and* "NOT-Emotion-Sad." The second, "sad", has 49 songs tagged "Emotion-Sad" *and* "NOT-Emotion-Happy." We compute features for each disjoint 29.7 s of each song, creating 619 features in the first class, and 285 in the second.

4.2. Evaluation (experimental design, data, FoM)

We first use *Classify* with 2-fold non-stratified CV. We split the songs randomly such that the first partition has 311 features from the first class, and 142 from the second class. We do not split features from any song across the training and test sets. Table 1 shows the performance statistics. This system appears more adept at applying "happy" to songs in CAL500 tagged "happy" than "sad" to songs tagged "sad." SRCAM shows a normalized classification accuracy of 0.65, which is better than chance; but a system labeling all songs "happy" obtains a normalized classification accuracy of 0.5, and an F-score for "happy" of 0.83 (but an F-score for "sad" of 0).

	True		
	H	S	Prec.
H	0.79	0.49	0.79
S	0.21	0.51	0.51
F-score	0.79	0.51	

Table 1. Classification performance of SRCAM

We now train SRCAM with all but two songs in CAL500: "ABC" by Jackson 5 (tagged "happy"), and "Lightning Crashes" by Live (tagged "sad"). Figure 1(a,c) shows the segments delimited by dashed vertical lines, and the class-specific error (1) in each (which we want small). SRCAM labels "happy" all of "ABC", and "sad" all of "Lightning Crashes". Hence, SRCAM performs perfectly with respect to labeling all segments of these songs. Note, too, that one might relate class-specific error to a dimensional model of emotion, such that when the two errors are equal the emotion content of the segment is at the origin of the valence-arousal space.

We now pass these signals through filters with the magnitude responses shown in Fig. 2. (We do this for testing MGR systems in [25].) Figures 1(b,d) shows the resulting signal waveforms and class-specific errors of SRCAM. Though these filters do not radically affect the music content in the signals, SRCAM now picks the wrong classes for all segments. Furthermore, in the case of "Lightning Crashes," we see that all "happy" class-specific errors in the filtered version are lower than all "sad" class-specific errors in the original version. We can reproduce this behavior with other excerpts.

4.3. Discussion

This demonstration shows how the perfect performance of SRCAM for these songs, and the results in Table 1, *reveal nothing* about whether SRCAM addresses the principal goals of MER in this limited two-class problem. When a change to a signal that is irrelevant to humans so severely handicaps an MER system, it must be making decisions based on factors unrelated to the music to which it is supposedly listening.

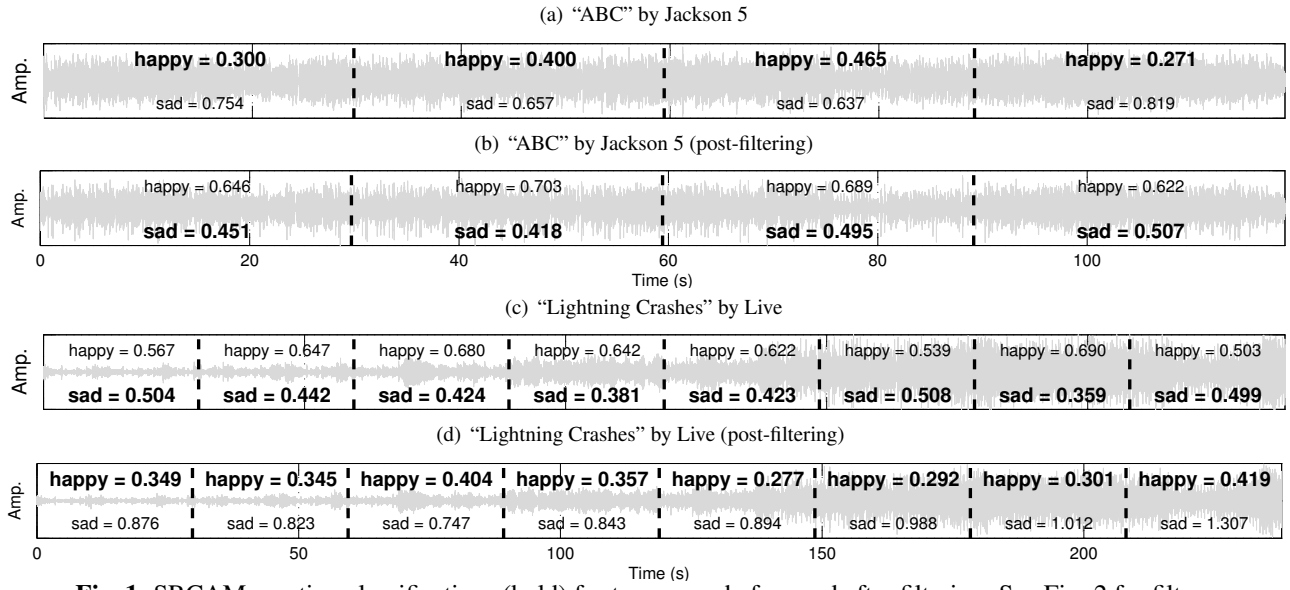

Fig. 1. SRCAM emotion classifications (bold) for two songs before and after filtering. See Fig. 2 for filters

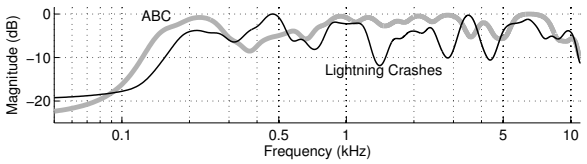

Fig. 2. Filters responses altering SRCAM emotions in Fig. 1

There are two arguments one can make. First, such sensitivity to filtering is expected when training on spectral features. However, as we quote in Section 3, Song et al. [30] and Barthet et al. [4] conclude spectral/timbral features outperform many others in MER. Our demonstration has not only shown this conclusion to be invalid with respect to the principal goals of MER, but also makes clear that when a system uses features sensitive to irrelevant changes in a signal, that is evidence against the suitability of such features for MER. Second, one might be tempted to train the system on "all filtered versions" of the data. While this might produce an MER system that is more robust to the changes in the environment, it still does not guarantee the system will then not be making decisions based on confounding factors.

5. CONCLUSION

The validity of evaluation in music information research in general has rarely been questioned before [22]. We have here reviewed problems of validity in MGR evaluation, related them to evaluation in MER, and shown the crux of the problem in both: uncontrolled independent variables in data renders any conclusions from *Classify* invalid with respect to the principal goals of MGR/MER. We show this conclusively for an MER system by using *Robust* [10].

One might argue, by a cost and benefit analysis, that *Classify* using millions of songs provides a standard and systematic approach for evaluating MGR/MER systems that is not as labor intensive as alternatives. When the benefit is always zero, the cost is irrelevant. One might argue the goal of MER is really about reproducing with high accuracy the genre/emotion labels of a particular music dataset — in which case it is valid to use *Classify* in the relevant dataset. Solving this problem, however, cannot be argued to be useful for or relevant to recognizing genre/emotion in music. One might argue MGR/MER is really about crafting the *illusion* of genre/emotion recognition, which can indeed be useful for some applications, e.g., [36]. Still, *Classify* with a dataset having uncontrolled independent variables does not reliably indicate real-world success. This requires a different approach to evaluation; and, for the principal goals of MGR/MER, there are alternatives.

Hu et al. [29] state that one evaluation option considered for MIREX AMC was *Retrieve* [10]. This experimental design is much better than *Classify* for the principal goals of MER since it emphasizes not reproducing labels, but choosing labels that are indistinguishable from those humans would choose. Whereas humans and music play little role in *Classify*, they can be put first in *Retrieve*. Related to this is performing a listening test of certain outcomes of *Classify*. In [8], we use this to determine the extent to which humans can tell which of two genre labels for a music excerpt was selected by a human. Our results show that the consistent and persistent mistakes made by two state-of-the-art MGR systems are easily detectable. We describe many other approaches in [10].

As Barthet et al. [4] observe, "[Most] approaches to [MER use] black-box models which do not take into account the interpretability of the relationships between features and emotion components; this is a disadvantage when trying to understand the underlying mechanisms." We add to this that evaluation must not be treated as a black-box either, and must be envisioned together with the proposed hypotheses [21], and argued to be relevant. A proper evaluation from which valid conclusions can be drawn about a hypothesis requires a thought and creativity that cannot be dismissed by, e.g., larger

datasets of music, i.e., posing in the same way more of the same kinds of questions to Clever Hans, horse mathematician.

6. REFERENCES

[1] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. J. Scott, J. A. Speck, and D. Turnbull, "State of the art report: Music emotion recognition: A state of the art review," in *ISMIR*, 2010, pp. 255–266.

[2] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, Apr. 2011.

[3] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 3, pp. 1–30, May 2012.

[4] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Implications for content and context-based models," in *Proc. CMMR*, 2012.

[5] J. Frow, *Genre*, Routledge, New York, NY, USA, 2005.

[6] G. A. Wiggins, "Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music," in *Proc. IEEE Int. Symp. Mulitmedia*, Dec. 2009, pp. 477–482.

[7] F. Pachet and D. Cazaly, "A taxonomy of musical genres," in *Proc. Content-based Multimedia Information Access Conference*, Paris, France, Apr. 2000.

[8] B. L. Sturm, "Classification accuracy is not enough: On the evaluation of music genre recognition systems," *J. Intell. Info. Systems (accepted)*, 2013.

[9] Y.-C. Lin, Y.-H. Yang, and H.-H. Chen, "Exploiting genre for music emotion classification," in *Proc. ICME*, 2009.

[10] B. L. Sturm, "A survey of evaluation in music genre recognition," in *Proc. Adaptive Multimedia Retrieval*, Copenhagen, Denmark, Oct. 2012.

[11] J.-J. Aucouturier and E. Pampalk, "Introduction – from genres to tags: A little epistemology of music information retrieval research," *J. New Music Research*, vol. 37, no. 2, pp. 87–92, 2008.

[12] P. N. Juslin and D. Västfjäll, "Emotional responses to music: The need to consider underlying mechanisms," *Behavioral and Brain Sciences*, vol. 31, pp. 559–621, 2008.

[13] MIREX, "`http://www.music-ir.org/mirex`," 2012.

[14] J.-J. Aucouturier and F. Pachet, "Representing music genre: A state of the art," *J. New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

[15] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, July 2002.

[16] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: A survey," *IEEE Signal Process. Mag.*, vol. 23, no. 2, pp. 133–141, Mar. 2006.

[17] B. Matityaho and M. Furst, "Neural network based model for classification of music type," in *Proc. Conv. Electrical and Elect. Eng. in Israel*, Mar. 1995, pp. 1–5.

[18] B. L. Sturm, "An analysis of the GTZAN music genre dataset," in *Proc. ACM MIRUM Workshop*, Nara, Japan, Nov. 2012.

[19] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their affect on evaluation, and its future use," 2013 (submitted).

[20] R. O. Gjerdingen and D. Perrott, "Scanning the dial: The rapid recognition of music genres," *J. New Music Research*, vol. 37, no. 2, pp. 93–100, Spring 2008.

[21] D. T. Campbell and J. C. Stanley, *Experimental and quasi-experimental designs for research*, Houghton Mifflin Company, Boston, 1963.

[22] J. Urbano, "Information retrieval meta-evaluation: Challenges and opportunities in the music domain," in *ISMIR*, 2011, pp. 609–614.

[23] K. L. Priddy and P. E. Keller, *Artificial Neural Networks: An Introduction*, SPIE Press, 2005.

[24] G. Marques, T. Langlois, F. Gouyon, M. Lopes, and M. Sordo, "Short-term feature space and music genre classification," *J. New Music Research*, vol. 40, no. 2, pp. 127–137, 2011.

[25] B. L. Sturm, "Two systems for automatic music genre recognition: What are they really recognizing?," in *Proc. ACM MIRUM Workshop*, Nara, Japan, Nov. 2012.

[26] B. L. Sturm, "On music genre classification via compressive sampling," in *Proc. ICME*, 2013.

[27] B. L. Sturm and F. Gouyon, "Comments on "automatic classification of musical genres using inter-genre similarity"," 2013 (submitted).

[28] O. Pfungst (translated by C. L. Rahn), *Clever Hans (The horse of Mr. Von Osten): A contribution to experimental animal and human psychology*, Henry Holt, New York, 1911.

[29] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, "The 2007 MIREX audio mood classification task: lessons learned," in *Proc. ISMIR*, 2008.

[30] Y. Song, S. Dixon, and M. Pearce, "Evaluation of musical features for emotion classification," in *Proc. ISMIR*, Oct. 2012.

[31] F. H. Messerli, "Chocolate consumption, cognitive function, and nobel laureates," *The New England Journal of Medicine*, vol. 367, no. 16, pp. 1562–1564, Oct. 2012.

[32] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Music genre classification via sparse representations of auditory temporal modulations," in *Proc. EUSIPCO*, Aug. 2009.

[33] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Sparse multi-label linear embedding nonnegative tensor factorization for automatic music tagging," in *Proc. EUSIPCO*, Aalborg, Denmark, Aug. 2010, pp. 492–496.

[34] E. van den Berg and M. P. Friedlander, "Probing the Pareto frontier for basis pursuit solutions," *SIAM J. on Scientific Computing*, vol. 31, no. 2, pp. 890–912, Nov. 2008.

[35] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic description using the CAL500 data set," in *ACM SIGIR*, 2007.

[36] G. Xia, J. Tay, R. Dannenberg, and M. Veloso, "Autonomous robot dancing driven by beats and emotions of music," in *Proc. Int. Conf. Autonomous Agents Multiagent Syst.*, Richland, SC, 2012, pp. 205–212.