A Simple Method to Determine if a Music Information Retrieval System is a "Horse"

Bob L. Sturm, *Member, IEEE*

***Abstract*—We propose and demonstrate a simple method to explain the figure of merit (FoM) of a music information retrieval (MIR) system evaluated in a dataset, specifically, whether the FoM comes from the system using characteristics confounded with the "ground truth" of the dataset. Akin to the controlled experiments designed to test the supposed mathematical ability of the famous horse "Clever Hans," we perform two experiments to show how three state-of-the-art MIR systems produce excellent FoM in spite of not using musical knowledge. This provides avenues for improving MIR systems, as well as their evaluation. We make available a reproducible research package so that others can apply the same method to evaluating other MIR systems.**

***Index Terms*—2-WORK system performance, 5-CONT content description and annotation, 5-SEAR multimedia search and retrieval.**

I. Introduction

MUSIC information retrieval (MIR) aims to produce systems that extract information from recorded music signals. Such information can be useful for indexing and searching music databases (archiving, identification, browsing, recommendation, organizing), for studying music (transcription, description, categorization), and even for creating music (mastering, remixing, playlist generation, composition). However, despite "human level" figures of merit (FoM) found through standard, systematic and rigorous evaluations of state of the art MIR systems [1]–[4], these same systems can appear upon closer inspection to be making decisions as if they are not even considering the *music* in recorded music signals [4]–[10].

Whether it is crucial for an MIR system to consider the music in a recorded music signal depends on a *use case* [11], [12]. For example, consider two systems purported to detect the key of a recorded piece of tonal music. Unbeknownst to all, one system considers the relationships between all pitches present and their function in the composition, and the other system selects only the first chord it finds. If a test dataset is inadvertently composed mostly of tonal music recordings that begin in the tonic, then comparing the "ground truth" labels with those produced by a system can result in excellent FoMs for both systems in spite of the fact that one of them does not address the musical problem of key detection. In the "real world," where tonal music pieces do not always begin in the tonic, and recordings of music do not always start at the beginning of a piece, the success of the second system critically depends upon the preservation of the fragile confounded characteristic it uses.

In this article, we propose a method to test the hypothesis that the FoM resulting from evaluating an MIR system in a dataset comes not from it addressing the musical problem for which it is designed, but instead from its reliance upon characteristics in the dataset confounded with the "ground truth." The standard, systematic and rigorous approach to evaluation used most in MIR research (the measurement of the amount of "ground truth" produced by a system, which we call *Classify* [3]) simply cannot distinguish between a system that operates by using characteristics confounded with the "ground truth" of a dataset, and another that actually considers the *music* in recorded music signals [9], [13]. The most fundamental reason for this is its lack of control of independent variables [4], [14]. This means that the FoMs published in many papers, articles, and annual evaluation campaigns like MIREX[1] are not valid for concluding upon the extent to which any of the work actually addresses the intended problem, not to mention how to improve any of the proposed systems [4], [14]. At worst, this can result in neglecting the development of systems that actually consider music because their FoMs are judged worse than those of systems that rely on "tricks". By explaining *why* an MIR system produces the FoM it does, our method provides a sanity test of an MIR system, suggests ways to improve it, and thus ultimately provides a way to complete the "IR research and development cycle" for which MIR has been described as falling short [14]–[16].

In the next section, we summarize the story of "Clever Hans". We then formally present our method. In the fourth section, we apply our method to evaluating three state of the art MIR systems, and show that, though they produce excellent FoMs, they are nonetheless working with confounded characteristics, i.e., they are "horses". All figures and results in this paper are reproducible using the code at: http://imi.aau.dk/~bst/software/SturmHorse.tgz.

II. The Clever Horse Named Hans

At the beginning of the 20th century, a horse named "Hans" handled by a retired mathematics teacher (Wilhelm von Osten) appeared in Germany to be capable of complex arithmetic feats, as well as many other problems requiring abstract thought [17]. When asked in front of an audience, "What are the factors of 28?" Hans would tap his right hoof once, twice, then four times, seven, and so on until 28. What is more, Hans was able to answer questions posed him by people other than von Osten. One of his

Manuscript received December 17, 2013; revised March 27, 2014; accepted May 28, 2014. Date of publication July 02, 2014; date of current version September 15, 2014. This work was supported in part by Independent Postdoc under Grant 11-105218 from Det Frie Forskningsråd. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Chong-Wah Ngo.

The author is with the Audio Analysis Lab, AD:MT, Aalborg University Copenhagen, Copenhagen DK-2450, Denmark (e-mail: bst@create.aau.dk).

Digital Object Identifier 10.1109/TMM.2014.2330697

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

most emphatic "converts" was an initially skeptical zoologist who found Hans correctly answered his questions even when von Osten was absent.

While such a display amuses—and, in fact, we see similar stories reported by the media today—it either challenges the uniqueness of humans in their ability for abstract thought, or suggests something else is at play. It was not until properly controlled experiments were designed, implemented and analyzed that Hans was definitively proven clever, but not actually solving any problem posed to him [17].

The experimentalist, Dr. Oskar Pfungst, performed several experiments in controlled conditions to determine, first, if Hans is capable of abstract thought; and, if not, then by what mechanism Hans is able to correctly answer problems. In one experiment, Pfungst alternated trials in which the answer to a problem was known by everyone, or the answer could only be known by Hans. In trials of the former, Hans was verbally asked by von Osten to add two numbers; in the latter, Hans was told to add the numbers whispered in his ears by separate persons, each not knowing the other's number. Of 31 trials of the former, Hans answered 29 correctly; but of 31 trials of the latter, he answered only 3 correctly.

After Pfungst demonstrated Hans has no ability in arithmetic, he sought to answer how Hans appears able to respond correctly. Pfungst started by applying blinders to Hans so that only what is in front of him is visible. Problems were posed to Hans in alternating trials with the questioner either in Hans' field of view or not. Of the trials in which Hans could see the questioner, Hans was correct in 86%; when Hans could not see the questioner, his score dropped to 6%. The same occurred when Hans and the questioner were separated by canvas.

After Pfungst found that Hans relied on vision to answer correctly, he sought the specific visual mechanism. He found Hans would not begin to respond to a problem until the questioner slightly tilted his head and torso; and Hans would continue to tap until they were erect again. Using such cues, Pfungst reliably elicited from Hans any response to any question. The same result occurred when Pfungst instructed other people to consciously control those visual cues. Thus, the remarkable ability of Hans was explained, and also that he really did depend on humans to solve his problems [18]. That people believed they were not giving cues when in fact they were is now known as the "Clever Hans Effect."

III. Testing for a Reliance on Confounds

The current state of evaluation in MIR is analogous to the historical episode of the horse "Clever Hans" [4]: many MIR systems have been and continue to be asked questions in uncontrolled conditions, their answers tallied and compared to a "ground truth," and conclusions made about the proximity of their FoM to that expected of humans, or their superiority to other MIR systems asked the same questions in similarly uncontrolled conditions. With properly controlled experiments, Hans was found to not be capable of human feats of intelligence. These experiments directly motivate the method we propose to determine whether an MIR system is actually a "horse:" *a system appearing capable of a remarkable human feat, e.g., music genre recognition, but actually working by using irrelevant characteristics (confounds).*

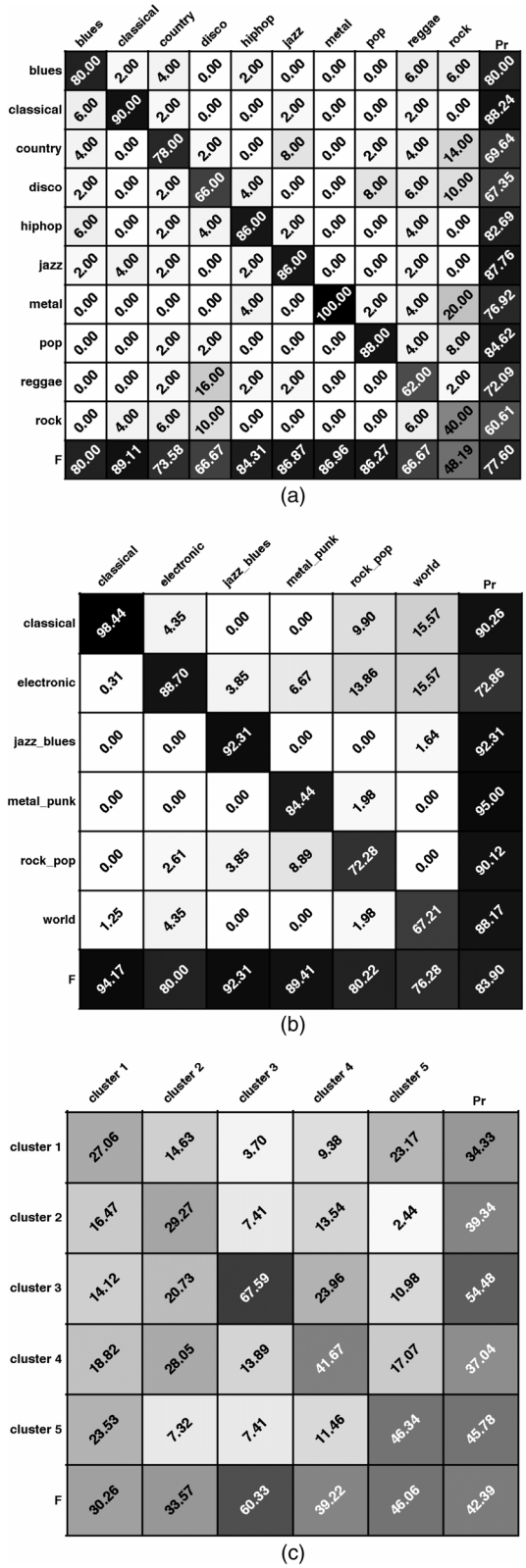


Fig. 1. Figures of merit (FoM, $\times 100$) for three MDS. Column is "true" label, and row is selected class. Off diagonals are confusions. Precision is the rightmost column, F-score is the bottom row, recall is the diagonal, and normalized accuracy (mean recall) is at bottom-right corner. (a) S_G tested on GTZAN fold 2. (b) S_I tested on ISMIR2004 validation. (c) S_P tested on PandaMood.

As Pfungst did for Hans [17], we do for an MIR system: *we seek to explain its FoM*. For instance, Fig. 1(a) shows the FoM

of an MIR system: its classification accuracy is over 0.77, which is significantly larger than the 0.1 expected of a random system. It is also greater than about half of the highest classification accuracies reported using the same dataset [19]. The "null hypothesis"—the default position when it comes to an artificial system appearing capable of a remarkable human feat [4]—is that this system achieves its FoM by relying on confounds. We now define and formalize this problem, present the "method of irrelevant transformations", and finally contrast it with robustness testing.

A. Definitions and Formalization

We define a *system* as "a connected set of interacting and interdependent components that together address a goal" [13]. There are four types of components in which we are interested when analyzing a system [13]: *algorithm* (a finite set of ordered instructions that transduce input into output), *operator* (the one who runs the system), *instructions* (directions on how one runs and uses the system), and *environment* (connections between components, external training databases, etc.). We thus conceptualize a system as a "black box" run by an operator according to its instructions. It is critical to note that we define a system as a "finished product," ready to run, and not, for example, just a pairing of a machine learning algorithm with a feature extraction algorithm. The FoM in Fig. 1(a) comes from one test of a specific system, and it is this FoM of this specific system that we wish to explain.

Now, denote a *music universe* Ω, and a *vocabulary* $\mathcal{V}$. One member $\omega \in \Omega$ could be all or a portion of Beethoven's Fifth Symphony; and some members of $\mathcal{V}$ could be "Classical", "deaf composer", "knock knock", a symbolic transcription, a musical analysis, and the set of reals $\mathbb{R}$. In the physical world, any member of Ω is not directly accessible, but can manifest as content in a *physical recording*, e.g., a digitized mastered recording of Bernstein conducting the New York Philharmonic playing Beethoven's Fifth Symphony. Denote the *recording universe* as $\mathcal{X}_\Omega$. A member of $\mathcal{X}_\Omega$ is a *recording* of an $\omega \in \Omega$, which we define here as a sequence of digital samples $\mathbf{x}_\omega := (x_n \in \mathbb{R})_n$ (implicit are ω, and parameters like sampling rate). We say the music ω is *embedded* in the recording $\mathbf{x}_\omega$. There can be many recordings of an ω in $\mathcal{X}_\Omega$.

We define the *semantic universe* as

$$\mathcal{S}_{\mathcal{V},A} := \{s = (v_1, \ldots, v_n) | n \in \mathbb{N}, \forall 1 \leq i \leq n [v_i \in \mathcal{V}] \wedge A(s)\} \quad (1)$$

where $A(\cdot)$ encompasses a semantic rule, for instance, restricting $\mathcal{S}_{\mathcal{V},A}$ to consist of sequences of cardinality 1. We define *music description* as pairing an element of Ω or $\mathcal{R}_\Omega$ with an element of $\mathcal{S}_{\mathcal{V},A}$. The *problem of music description* is to make the pairing "acceptable" with respect to a use case. A *use case* is a specification of Ω (e.g., "Classical" music) and $\mathcal{X}_\Omega$ (e.g., 30 second digital recordings of "Classical" music), $\mathcal{V}$ (e.g., instruments) and $A(\cdot)$, and success criteria (e.g., determine instrumentation of recorded "Classical" music). We define a *music description system* (MDS) as a map from the recording universe to the semantic universe: $S : \mathcal{X}_\Omega \rightarrow \mathcal{S}_{\mathcal{V},A}$. *Building* an MDS means making a map according to well-specified criteria, e.g., find the S maximizing recall in a training dataset. Finally, we build a *dataset* by assembling tuples of recordings and elements of a semantic universe, e.g., $\mathcal{D} := \{\{\mathbf{x}_\omega \in \mathcal{X}_\Omega, s \in \mathcal{S}_{\mathcal{V},A}\}_i\}_i$. We call the sequence $(s_i)_i$ the "ground truth" of $\mathcal{D}$.

Of interest to a user of an MDS is how well it addresses a specific use case [11], [12]. One thus evaluates the system by performing experiments to estimate a variety of statistics, or FoM, that characterize the performance expected of the system when employed in the use case. One standard way this is attempted is by *Classify* [3]: have the MDS "treat" the recordings of a standard dataset, compare its mappings with the "ground truth", and summarize them using, e.g., mean classification accuracy, recalls, and confusions. The *real world performance* of an MDS are the FoM that result from an experiment using a dataset of every member, rather than a random sampling, of $\mathcal{X}_\Omega$. Depending on the use case, this might not be possible; and so statistical tests are performed to determine significant differences in performance between two systems, or that of picking an element of $\mathcal{S}_{\mathcal{V},A}$ independent of $\mathcal{X}_\Omega$. These statistical tests are all subject to implicit and strict assumptions on the measurement model and its appropriateness to describe the measurements made in the experiment [13], [20], [21].

A danger to the validity of experiments—the quality of the estimates of real world performance—are uncontrolled independent variables, or characteristics, that might be confounded with the "ground truth" [14], [20]. A characteristic is *confounded with the "ground truth" of* $\mathcal{D}$ when the recordings in $\mathcal{D}$ are a finite sampling of $\mathcal{X}_\Omega$, and the characteristic and the "ground truth" of $\mathcal{D}$ are correlated while the characteristic and the "ground truth" of $\mathcal{X}_\Omega$ are independent. A well-known example of such a characteristic in MIR is "artist" [5], [6].

To make the formalism above more clear, consider a concrete example. One of the most prevalent kinds of MIR systems is an MDS using a vocabulary that is indicative of characteristics such as genre, mood, and instrumentation. Since people find it useful to talk about and search for music in such terms (they constitute 77% of the top 500 "tags" of 7 million manually applied to artists on the last.fm music service [22]), a large amount of research has been devoted to creating systems to automatically recognize such high-level characteristics from music recordings [2], [4], [9], [23]–[25]. More specifically, consider the 100 MDS built using the most-used public dataset in music genre recognition research: GTZAN[2] [4], [26]. Since a use case is rarely, if ever, specified in the creation of all these systems [11], [12], implicit in the use of GTZAN are: a music universe Ω consisting of, at least, music excerpts of, e.g., songs by Willie Nelson, compositions by Mozart, and performances of Coleman Hawkins [4]; a vocabulary, $\mathcal{V} = \{\text{Blues}, \text{Classical}, \text{Country}, \text{Disco}, \text{Hiphop}, \text{Jazz}, \text{Metal}, \text{Pop}, \text{Reggae}, \text{Rock}\}$, and a semantic rule limiting $\mathcal{S}_{\mathcal{V},A}$ to be single elements of $\mathcal{V}$; a recording universe $\mathcal{X}_\Omega$ of digital audio of durations of about 30 seconds; and a success criterion of building a map to reproduce the most "ground truth" of GTZAN. Most MDS created with GTZAN are built using only one portion of it, and then evaluated on the rest, without

[2]http://marsyas.info/download/data_sets

taking into consideration the content of GTZAN [4], [19]. The design of this evaluation is most often *Classify* [3]. The implicit assumption is that this approach to evaluation, and the resulting FoM, provide a reasonable reflection of the extent to which an MDS addresses the problem of music description in $\mathcal{S}_{\mathcal{V},A}$.

B. The Method of Irrelevant Transformations

We now present an experimental methodology to test the hypothesis: S achieves an FoM in $\mathcal{D}$ by relying on characteristics confounded with the "ground truth". In the case of Clever Hans, Pfungst showed that though Hans correctly answered most math questions in front of an audience, Hans has no ability in mathematics. Pfungst also showed Hans answers most mathematical questions correctly by relying on vision, and, in particular, specific visual cues of the questioner. Likewise, our method tests the sanity of a FoM (and the validity of the evaluation underlying it) for concluding that an MIR system is addressing the problem for which it is designed, and, further, illuminates the *nature* of the characteristics a system uses to make its decisions.

Consider Ω, $\mathcal{X}_\Omega$ and $\mathcal{S}_{\mathcal{V},A}$ are specified by a use case. Define the *transformation*, $T : \mathcal{X}_\Omega \to \mathcal{X}_\Omega$. We call T an *irrelevant transformation with respect to* $\mathcal{S}_{\mathcal{V},A}$ if the "ground truth" of each element of $\mathcal{X}_\Omega$ is the "ground truth" of its transformation. Now consider treating the recordings of a dataset $\mathcal{D}$ by MDS S, and comparing its outputs to the "ground truth" to compute a FoM. Assume this FoM is higher than that expected of randomly picking from $\mathcal{S}_{\mathcal{V},A}$. To test whether this FoM results from S using confounded characteristics in $\mathcal{D}$, we search for a set of irrelevant transformations that, when applied to recordings in $\mathcal{D}$, result in S producing an FoM consistent with that expected of randomly picking an element of $\mathcal{S}_{\mathcal{V},A}$, or result in S performing perfectly. Finding such transformations supports the hypothesis that S uses confounded characteristics in $\mathcal{D}$. What is more, the kinds of irrelevant transformations we find capable of doing this reveal the nature of the confounds in $\mathcal{D}$ upon which S relies.

To make this more clear, we return to Clever Hans. Pfungst controls characteristics in an experiment with Hans by using irrelevant transformations of mathematical questions. Examples of these irrelevant transformations are: Pfungst asking a question instead of von Osten; having a question asked out of Hans' view instead of in his view; and having a question asked by someone who does not know the answer instead of someone who does. The irrelevant transformations for which Hans becomes unable to correctly respond reveals the nature of the confounds responsible for his mathematical "abilities": questioner knows correct answer *and* questioner is in Hans' eyesight. Once Pfungst discovered the specific visual stimulus to which Hans responds, he was able to elicit any response at all from Hans regardless of the question. Thereby, the validity of an experiment in which Hans appears capable of mathematics, not to mention any claim that Hans is capable of mathematics, are clearly rendered dubious.

Returning to evaluating MIR systems, we can create transformations by filtering, amplifying, cropping, equalization, time stretching, companding, remixing, effects like reverberation, phase vocoding, autotune, combinations of these, and so on. Whether or not a transformation is irrelevant depends on $\mathcal{S}_{\mathcal{V},A}$. Examples of irrelevant transformations for a semantic universe denoting the key of tonal music include: equalization; time stretching by up to 2% (where pitch is preserved); normalization; playing the piece on a tuned organ instead of tuned piano; and cropping the recording to remove the first chord. Considering the two MDS we discuss in the introduction, the FoM of the one that chooses the first chord in the recording as the key could be significantly affected by the irrelevant transformation of cropping the first chord.

C. Contrasting With Robustness Evaluation

One might see the method of irrelevant transformations as *robustness evaluation* [27], which aims to answer questions of how the FoM of a system changes with environmental conditions. For instance, if a use case requires a system to have some minimum performance on a cellular telephone, then robustness evaluation addresses questions related to how the FoM of the system is impacted by noise in the environment, the fidelities of microphones of cellular telephones, coding specifications, and so on. Put in terms of Clever Hans, a robustness evaluation measures how Hans responds to questions posed in different weather conditions, at different times of the day, at different locations, when he is hungry or when he is not, and so on. The method of irrelevant transformations, however, tests the hypothesis that Hans is not actually capable of remarkable human feats. By applying the method of irrelevant transformations to an MIR system, we seek to determine whether the system is actually considering the music at all. In this respect, our aims in using the method of irrelevant transformations are more aligned with the work of Porter and Neuringer with pigeons [28], Watanabe and Nemoto with sparrows [29], Chase with koi carp [30], and Marques *et al.* [7], [8] with MDS.

IV. Experiments

We now perform two experiments applying the method of irrelevant transformations to MDS. We first describe the MDS, then the methodology, and finally our results.

A. Systems and Preliminary Evaluation

We create three different MDS, each of which maps a recording to a single element of a vocabulary. One system, S_G, uses the $\mathcal{V}$ of the GTZAN dataset [4], [19], [26], and the other, S_I, uses that of the ISMIR2004 dataset.[3] The third system, S_P, uses the $\mathcal{V}$ of the audio portion of the music emotion dataset in [31] (PandaMood). The semantic universes of all three systems are single elements of their vocabularies. We build S_G and S_I using sparse representation classification (SRC) of auditory temporal modulations [9]. The environment of S_G includes as training data one half of GTZAN, which we partition randomly. The environment of S_I uses the "training" dataset of ISMIR2004. Since the recording universe of S_G encompasses audio signals of 30 second duration, it extracts one auditory temporal modulation from each input signal, and classifies that single feature using SRC. The recording universe of S_I, however, encompasses audio signals of at least 30 second duration; and when it extracts several auditory temporal modulations from an input (from contiguous segments of 30 second duration) its makes a classification using majority voting from SRC performed on each individual feature. This approach and the settings we use for these two systems are fully described

[3]http://ismir2004.ismir.net/genre_contest/index.html

as "SRCAM" elsewhere [9], and are freely available in our accompanying code. Finally, we create S_P according to the specifications in [32]: a support vector machine with radial basis function, and $C = 1.0$, trained on "spectral features" extracted from amplitude-normalized music signals using the *MIRtoolbox* [33]. These features include mean and standard deviations of 16 features, such as spectrum centroid, kurtosis, flux, and MFCCs. We standardize the dimensions of the features. The environment of S_P has as training data one half of PandaMood, which we create randomly.

Fig. 1 shows FoMs of our three MDS produced by a preliminary evaluation. We input to S_G and S_P all elements from the half of their respective datasets on which they are not trained; and we input to S_I all elements of the "validation" dataset of ISMIR 2004. We compare the output of each system with the "ground truth" of the dataset, and compute confusions, recalls, precisions, F-scores, and mean normalized classification accuracy, all shown in Fig. 1. The labels of the columns except the last show the vocabulary for each system. Along the diagonal are the recalls of each label, except the bottom right corner, which shows the normalized accuracy (mean recall). The bottom-most row is the F-score for each label, and the right-most column are their precisions. The off-diagonal elements show percent confusions. For example, Fig. 1(a) shows that S_G produces a normalized accuracy of 0.776; and we see 80% of the excerpts with the "ground trurth" of "Blues" are classified "Blues," but 6% are classified "Classical." The FoM of each system is significantly better than that expected when picking randomly from each $\mathcal{S}_{\mathcal{V},A}$; and all actually appear competitive with other systems tested in the same datasets [4], [31], [34]–[36].

B. Experiment 1: Uncovering a Reliance on Confounds

In this experiment, we investigate how the FoM of each MDS changes with irrelevant transformations of the test dataset. We first attempt to "inflate" the FoM for a given system S and test dataset $\mathcal{D}$:

1) Find the recordings in $\mathcal{D}$ that S maps "incorrectly"
2) Create irrelevant transformation T
3) Apply T to all recordings found in (1)
4) Have S map transformed recordings
5) Find the recordings that S maps "correctly"
6) For each recording in (1) that S now maps "correctly" in (5), replace it in $\mathcal{D}$ with its irrelevant transformation
7) Return to (1), repeat $20\times$, or until FoM of S is perfect.

We then attempt to "deflate" the performance of S to an FoM that is consistent with that expected of picking randomly from $\mathcal{S}_{\mathcal{V},A}$. We do this using the same approach above, but replacing the recordings in $\mathcal{D}$ that S maps "correctly" with irrelevant transformations for which S maps "incorrectly."

For the systems we test here, we generate each irrelevant transformation in the following supervised way. We take a 96-band near perfect reconstruction filterbank,[4] randomly choose several bands, and reduce their gains from 1 to 0.1 We then apply this time-invariant filter to one test recording, listen to the result, and either accept it as irrelevant, or generate and listen to the effects of new filters created in the same way, until we find an acceptable one. Fig. 2 shows the magnitude responses of some of these irrelevant transformations. The qualitative effects from them can be heard here: http://imi.aau.dk/~bst/research/TM_expt2/.

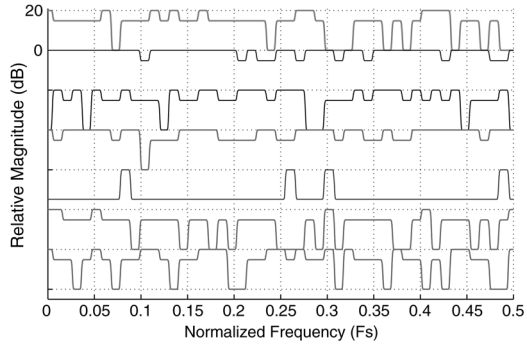


Fig. 2. Relative magnitude responses of several irrelevant transformations we use in Experiment 1 for S_G.

Fig. 3 shows the FoM for the three MDS resulting from the above "inflation" and "deflation" process. We can "deflate" each FoM to be consistent with that expected from randomly selecting from $\mathcal{S}_{\mathcal{V},A}$ even though the *music* embedded in any of the recordings of the test dataset $\mathcal{D}$ does not change. Though we do not reach perfect FoMs in 20 iterations, we are able to "inflate" them to be much better than those in Fig. 1.

C. Experiment 2: Using the Confounds to Produce Any Class

In this experiment, we investigate whether each MDS can be made to classify the same music embedded in a recording to any element of $\mathcal{S}_{\mathcal{V},A}$. We perform the following procedure with a system S and a recording $\mathbf{x}$. Starting with $\mathcal{R} = \{\}$

1) Create transformation T
2) If $S[T(\mathbf{x})]$ has not been selected, update $\mathcal{R} \leftarrow \mathcal{R} \cup \{T(\mathbf{x}), S[T(\mathbf{x})]\}$
3) Return to (1) and repeat $300\times$, or until every element of $\mathcal{S}_{\mathcal{V},A}$ has been selected.

We make transformations as in the first experiment, but not in a supervised way. Once this procedure has finished, we find which classes have been applied to the music embedded in $\mathbf{x}$, and listen to the transformations to confirm that they are irrelevant. We perform this procedure for ten recordings of different music, spanning a variety of genres and emotions.

Table I shows which classes we are able to elicit from each system for the ten different recordings. The results can be heard here: http://imi.aau.dk/~bst/research/TM_expt2. Clearly, we are able to make each MDS map most music recordings to any element of $\mathcal{S}_{\mathcal{V},A}$ by irrelevant transformations, even though one might take the FoM in Fig. 1 as suggesting each system has learned and is using something relevant to describing music with respect to $\mathcal{S}_{\mathcal{V},A}$.

D. Discussion

From our results, we see that the FoM of each MDS in Fig. 1 do not result from their use of criteria relevant to the task of music description. Though they are giving "right" answers most

[4]W. Lubberhuizen, "Near Perfect Reconstruction Polyphase Filterbank," MATLAB Central, http://www.mathworks.com/matlabcentral/fileexchange/15813-near-perfect-reconstruction-polyphase-filterbank (last accessed Mar. 23, 2014).

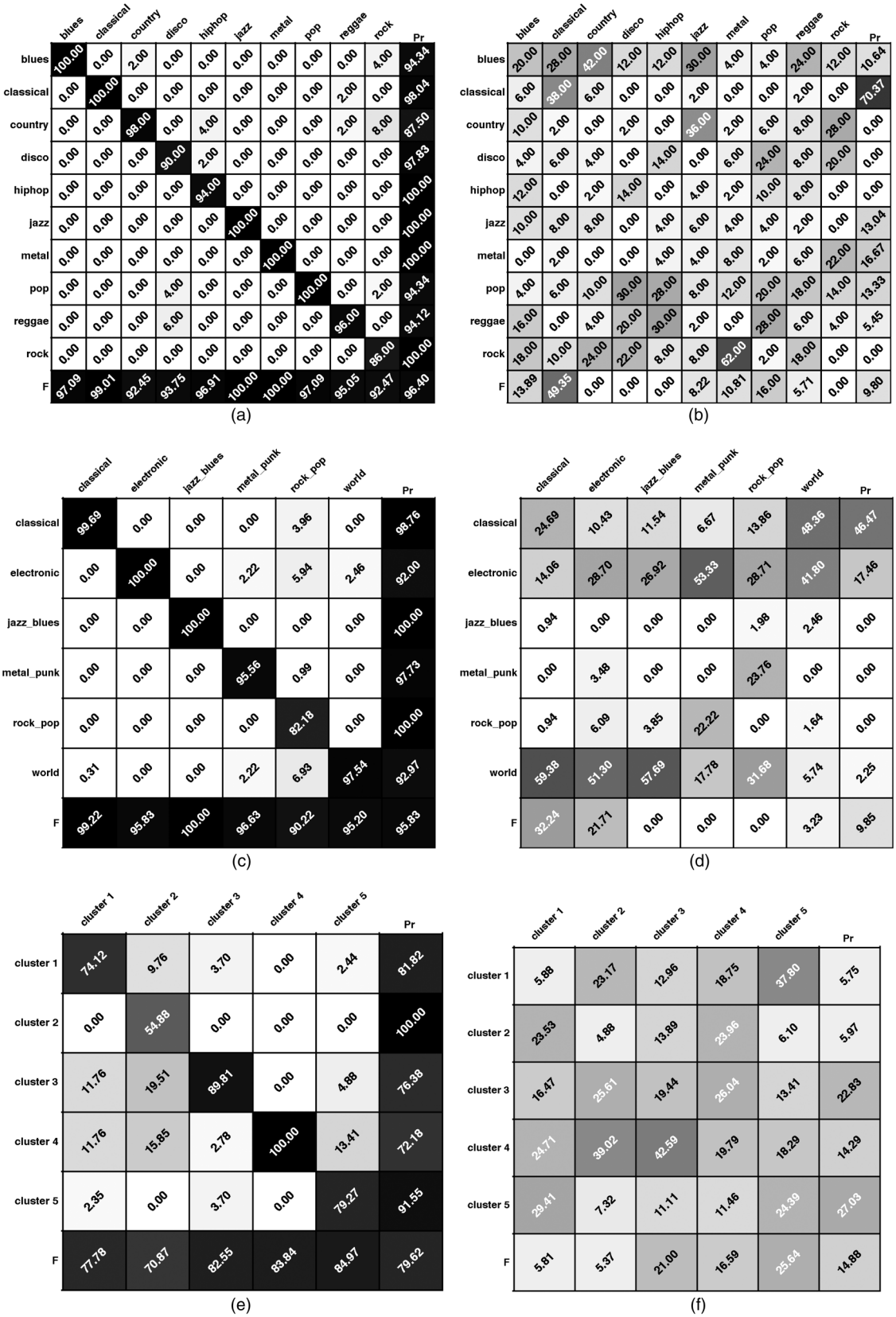


Fig. 3. "Inflated" and "deflated" FoMs ($\times 100$) for three MDS tested on irrelevant transformations of recordings in three datasets. Compare with initial results in Fig. 1 . (a) S_G: Inflation with GTZAN. (b) S_G: Deflation with GTZAN. (c) S_I: Inflation with ISMIR2004. (d) S_I: Deflation with ISMIR2004. (e) S_P: Inflation with PandaMood. (f) S_P: Deflation with PandaMood.

of the time in the "original" dataset, they are not for the "right" reasons. Just as Pfungst applied blinders to Clever Hans to disrupt his reliance on confounds, we disrupt by irrelevant transformations the fragile confounds learned by each MDS in these

TABLE I
RESULTS OF EXPERIMENT 2. EACH BLACK BOX IN A ROW SHOWS THE CLASS SELECTED BY EACH SYSTEM FOR AN IRRELEVANT TRANSFORMATION OF THE MUSIC EXCERPT SHOWN IN THE FIRST COLUMN. THESE RESULTS CAN BE HEARD HERE: HTTP://IMI.AAU.DK/~BST/RESEARCH/TM_EXPT2

	A_G										A_I					A_P				
Music Excerpt: *Title*, artist	*Blues*	*Classical*	*Country*	*Disco*	*Hip hop*	*Jazz*	*Metal*	*Pop*	*Reggae*	*Rock*	*Classical*	*Electronic*	*Jazz/Blues*	*Metal/Punk*	*Rock/Pop*	*World*	*Cluster 1*	*Cluster 2*	*Cluster 3*	*Cluster 4*	*Cluster 5*
"Last Year's Race Horse (Can't Run in this Year's Race)", Little Richard	■	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■		■	■	
"William Tell Overture", Rossini	■	■	■	■	■	■	■	■	■	■	■	■		■	■	■			■	■	■
"A Horse Called Music", Willie Nelson	■	■	■	■	■	■		■	■	■	■	■	■	■	■	■			■	■	
"10000 Horses Can't be Wrong", Simian Mobile Disco	■	■		■	■		■	■	■	■	■	■		■	■	■	■	■	■	■	■
"Horse Outside", The Rubberbandits	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■		■		■	
"(Ghost) Riders in the Sky", Leonard Gaskin	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
"Heavy Horses", Jethro Tull	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
"Bring on the Dancing Horses", Echo & The Bunnymen	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
"Mule Train", Count Prince Miller	■	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	
"Wild Horses", The Rolling Stones	■	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

datasets. In this way, we "deflate" the performance of each MDS to be consistent with that expected of choosing randomly from $\mathcal{S}_{\mathcal{V},A}$. We also see that these irrelevant transformations are not always destructive to the FoM of an MDS system, since we can use them to "inflate" performance. For both S_G and S_I, we can make them perform near perfect; and, while the performance measured for S_P is not close to perfect, it is much better than in Fig. 1.

Regardless of "right" or "wrong" labels, we clearly see from the results of Experiment 2 that each MDS maps to most of $\mathcal{S}_{\mathcal{V},A}$ *the same music*. Auditioning the resulting transformations makes clear just how irrelevant they are to the task of genre or emotion recognition—amounting in effect to particular equalization settings on a home stereo. We have applied the same approach to explain the FoM of the system reproducing most of the "ground truth" in the 2013 MIREX Audio Latin Music Genre task [37]. We find that very subtle time stretching can increase to perfect, or reduce to random, the FoM of this system. Just as Pfungst discovered and showed Clever Hans could be made to give any answer by performing the specific cues to which he was responding, we have found there exists irrelevant cues for these MDS. Hence, these results are incompatible with the claim that any of these MDS is using characteristics of music relevant to the task of *describing music*. In other words, each is a "horse."

V. OF "HORSES" IN MIR RESEARCH

That we have shown the three MIR systems above to be "horses" does not necessarily mean they are useless; and it does not necessarily mean that the approaches taken to build these systems (feature extraction and classification algorithms, datasets, and so on) are useless for extracting information from recordings. That we call an MIR system a "horse" is not meant to be an aspersion. As an intentional nod to Clever Hans, a "horse" is just a system that is not actually addressing the problem it appears to be solving. The judgment of whether a "horse" is useful or not completely revolves around a use case of a system [11], [12]: can the requirements demanded of a use case be satisfied by a system that relies on characteristics confounded with the "ground truth"? For instance, consider creating a dancing robot built to demonstrate a successful merging of controls with machine music listening [38]. The use case might specify that the robot *appears* to be dancing in a synchronized way with the music it hears, both "in time" with the music, and in a way that reflects the emotional content of the music. As emotion in music can be vague for humans [24], the success criteria for such synchrony could be less strict than that for synchrony in time. Hence, the emotion recognition success criteria of the robot might be met by a "horse."

The standard, systematic and rigorous approaches to evaluation used in MIR research, however, cannot differentiate between "horses" and systems that actually use relevant musical knowledge. Our recent survey [3] of nearly all published work addressing the problem of music genre recognition (467 publications up to the end of 2012) identifies ten different evaluation designs, but finds a poverty of evaluation in that work since 91% of it (i.e., 397 papers of 435 with experimental components) employs one design in particular: *Classify*, which compares the "ground truth" to the labels selected by a system to compute classification accuracy, and other FoM. In [9], we show how *Classify* is not valid for making a conclusion with regards to the capacity of a system to recognize genre. In [10], we show why *Classify* is not valid for making a conclusion with regards to whether a system is recognizing genre, or emotion in music. Since for music autotagging the majority of tags appear indicative of genre and mood [25], the problem of validity occurs in music autotagging evaluation as well.

Resulting from its lack of control over independent variables, two other subtle problems arise in the standard, systematic and rigorous approach to evaluation used in MIR research: 1) its measurement model relies on very strict assumptions; and, 2) error bounds on FoM have rarely, if ever, been calculated or specified [13], [20], [21]. One might argue that since error bounds can become smaller with a growing number of observations, what MIR evaluation really needs is much more data so that the resulting FoM better reflects real world performance. Along these same lines, one might argue that evaluating an MIR system with several datasets is better for predicting real-world performance than just using one dataset. These arguments, however, require several essential assumptions. First, the amount of music data needed to reach some specified statistical confidence cannot be known until the measurement model of the experiment is completely specified [20], [21]. Furthermore, when the number of independent variables in an experiment is unquantified, one is left to guess how much data is needed to validly answer a scientific question. Second, an increased amount of testing data does not necessarily lead to better error bounds. These bounds depend on the statistical deconstruction of the scientific question being addressed [39], the validity of the experimental approach designed, implemented and analyzed, how the data sampling is performed, the applicability

of the measurement model, and so on. Third, more data is not necessarily good data. *Classify* requires "ground truth," which itself requires a significant amount of labor, and its use assumes that the existence of a "ground truth" makes sense for the problem being addressed. "Ground truth" in music genre recognition, music emotion recognition and music autotagging, for instance, has notorious ambiguity [4], [40]–[44]. Fourth, while testing on independently constructed datasets seems like it can better predict the performance of a system in the real world, our experiments above shows this is not necessarily the case. On both *GTZAN* and *ISMIR2004*, we see SRCAM builds systems that appear to perform quite well, but are "horses" nonetheless.

Essentially, increasing the amount of testing data used with *Classify* amounts to asking a system more questions without addressing the real problem: a lack of control. As Pfungst showed for Clever Hans, more questions *in uncontrolled conditions* are not useful for detecting a "horse"; instead, well-designed experiments with carefully controlled conditions are necessary. For use cases in which a "horse" is not satisfactory, a method is needed to test when a system learns and uses confounded characteristics in a dataset undetected by the experimentalist. We cannot just ask an artificial system whether it gave the "right" answer for the "right" reasons. As Pfungst did for Clever Hans, creative and valid experiments must be designed and implemented to detect "horses."

That we have shown to be "horses" the three MIR systems above, and the one in [37], does not mean that all MIR systems are "horses." However, when one evaluates a highly specified artificial system with some dataset and finds an FoM appearing to support the claim that it is addressing a complex and human-centered problem requiring abstract thinking and a large amount of extrinsic data, the default position (null hypothesis) should be that the system is a "horse" [4]. To be persuaded by a high classification accuracy, or "sensible" confusions, that an artificial system displays a capacity for a complex human task—in some cases dependent upon human culture and all of its oddities—is but a manifestation of the *Clever Hans Effect*. One need not look deep into the MIR literature to find such examples [4], [9], [10], [13].

The outcomes of our experiments above make it clear that any MIR system employing low-level spectral features, like MFCCs, are in a precarious position. Given that it is difficult to find any audio-based MIR system that does not use low-level spectral features, that "MFCC is inarguably the single most important feature type that was widely used for genre classification" [2] (at least half of the systems competing in the 2012 MIREX MGR task use MFCCs), and that "spectral features" outperform other kinds of features for MER [32], most MIR systems should then be sensitive to the irrelevant transformations we use above. Such a prediction, in fact, is patently obvious by the very design of such an MIR system.

Finally, our findings above motivate the following hypothesis. Many researchers have observed a "performance limit" to MIR systems, which has been called a "glass ceiling," and attributed to "signal only approaches" and the "semantic gap" [45]–[47], the lack considering the involvement of humans [44], or the inappropriateness of standard features and machine learning methods [48]. Of this situation, Wiggins [44] writes, "There can be two responses … One is to keep trying. The other is to ask, 'Was it the right question?"' We posit instead that the question to answer first is if the evaluation of MIR systems is even relevant to make conclusions about their real world performance. In the words of Richard Hamming [49]: "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." We thus posit that the "performance limit" many have observed for MIR systems can be explained by something more simple than the claim that information extraction from audio signals has been maximized: the extents to which characteristics are confounded with the "ground truth" in standard datasets.

VI. Conclusion

We have proposed the simple method of irrelevant transformations, which seeks to *explain* the FoM resulting from a standard, systematic and rigorous evaluation of an MIR system. We demonstrate this method by performing two experiments with three state of the art MIR systems. We see that each system, like Clever Hans, relies upon characteristics confounded with the "ground truth." An immediate avenue for improving the systems above is to address their sensitivity to such irrelevant transformations, and then perhaps determine whether the resulting systems can better address a particular use case. One might augment the training data with transformed versions, or preprocess by whitening. Our current work explores these possibilities. In another direction, the method of irrelevant transformations leads to the analysis of MIR datasets. From our use of it in [37], we have discovered that tempo in the BALLROOM dataset [50] is confounded with the labels. As a result, a system considering *only* tempo can appear highly capable at identifying and discriminating between several different dance styles [51]. Our simple method of irrelevant transformations thus provides a means to address the incomplete "MIR research and development cycle" [14]–[16]: *evaluation approaches that not only reliably and meaningfully quantify the performance of systems, but that can also aid in their improvement.*

Acknowledgment

The author would like to thank F. Gouyon, N. Collins, A. Flexer, J.-J. Aucouturier, J. Urbano, V. Emiya, M. Davies, R. Bardeli, H. Purwins, C. Kereliuk, A. Pikrakis, and C. Sturm for many helpful discussions.

References

[1] 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.

[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] B. L. Sturm, "A survey of evaluation in music genre recognition," in *Post-Proc. 2012 Int. Workshop Adapt. Multimedia Retrieval, LNCS, 2014*, to be published.

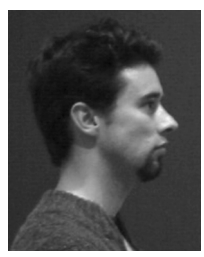
[4] B. L. Sturm, "The state of the art ten years after a state of the art: Future research in music information retrieval," *J. New Music Res.*, vol. 43, no. 2, pp. 147–172, 2014.

[5] E. Pampalk, A. Flexer, and G. Widmer, "Improvements of audio-based music similarity and genre classification," in *Proc. ISMIR*, London, U.K., Sep. 2005, pp. 628–233.

[6] A. Flexer, "A closer look on artist filters for musical genre classification," in *Proc. ISMIR*, Sep. 2007, pp. 341–344.

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

[8] G. Marques, M. Domingues, T. Langlois, and F. Gouyon, "Three current issues in music autotagging," in *Proc. ISMIR*, 2011, pp. 795–800.
[9] B. L. Sturm, "Classification accuracy is not enough: On the evaluation of music genre recognition systems," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 371–406, 2013.
[10] B. L. Sturm, "Evaluating music emotion recognition: Lessons from music genre recognition?," in *Proc. ICME*, 2013, pp. 1–6.
[11] C. C. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic, "The need for music information retrieval with user-centered and multimodal strategies," in *Proc. Int. ACM Workshop Music Inf. Retrieval User-Centered Multimodal Strategies*, 2011, pp. 1–6.
[12] M. Schedl, A. Flexer, and J. Urbano, "The neglected user in music information retrieval research," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 523–539, Dec. 2013.
[13] B. L. Sturm, "Making explicit the formalism underlying evaluation in music information retrieval research: A look at the MIREX automatic mood classification task," in *Post-Proc. 2013 Comput. Music Model. Res.*, to be published.
[14] J. Urbano, M. Schedl, and X. Serra, "Evaluation in music information retrieval," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 345–369, Dec. 2013.
[15] J. Urbano, "Evaluation in audio music similarity," Ph.D. dissertation, Dept. Informat., Univ. Carlos III of Madrid, Madrid, Spain, 2013.
[16] X. Serra *et al., Roadmap for Music Information ReSearch*, G. Peeters, Ed. Paris, France: Creative Commons, 2013.
[17] O. Pfungst*, Clever Hans (The Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology*, C. L. Rahn, Ed. New York, NY, USA: Henry Holt, 1911.
[18] C. Lesimple, C. Sankey, M.-A. Richard, and M. Hausberger, "Do horses expect humans to solve their problems?," *Frontiers Psychol.*, vol. 3, pp. 1–4, Aug. 2012.
[19] B. L. Sturm, "The GTZAN dataset: Its contents, its faults, their effects on evaluation, and its future use," *Clinical Orthopaedics Related Res.*, vol. 1306.1461, Jun. 2013 [Online]. Available: http://dblp.uni-trier.de/rec/bibtex/journals/corr/Sturm13
[20] R. A. Bailey*, Design of Comparative Experiments*. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[21] E. R. Dougherty and L. A. Dalton, "Scientific knowledge is possible with small-sample classification," *EURASIP J. Bioinf. Syst. Biol.*, vol. 2013:10, 2013.
[22] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere, "Autotagger: A model for predicting social tags from acoustic features on large music databases," *J. New Music Res.*, vol. 37, no. 2, pp. 115–135, 2008.
[23] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *IEEE Trans. Audio, Speech, Signal Process.*, vol. 17, no. 1, pp. 174–184, Jan. 2009.
[24] 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 *Proc. ISMIR*, 2010, pp. 255–266.
[25] T. Bertin-Mahieux, D. Eck, and M. Mandel, "Automatic tagging of audio: The state-of-the-art," in *Machine Audition: Principles, Algorithms and Systems*, W. Wang, Ed. Hershey, PA, USA: IGI Global, 2010.
[26] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.
[27] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *Proc. ISMIR*, 2013, pp. 83–88.
[28] D. Porter and A. Neuringer, "Music discriminations by pigeons," *Exp. Psychol.: Animal Behavior Processes*, vol. 10, no. 2, pp. 138–148, 1984.
[29] S. Watanabe and M. Nemoto, "Reinforcing property of music in Java sparrows (Padda oryzivora)," *Behavioural Processes*, vol. 43, no. 2, pp. 211–218, 1998.
[30] A. Chase, "Music discriminations by carp "Cyprinus carpio," *Animal Learn. Behavior*, vol. 29, no. 4, pp. 336–353, 2001.
[31] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "Multimodal music emotion recognition: A new dataset, methodology and comparative analysis," in *Proc. CMMR*, 2013, pp. 570–582.
[32] Y. Song, S. Dixon, and M. Pearce, "Evaluation of musical features for emotion classification," in *Proc. ISMIR*, Oct. 2012, pp. 523–528.
[33] O. Lartillot and P. Toiviainen, "A MATLAB toolbox for musical feature extraction from audio," in *Proc. DAFx*, 2007, pp. 237–244.
[34] Y. Costa, L. Oliveira, A. Koerich, F. Gouyon, and J. Martins, "Music genre classification using LBP textural features," *Signal Process.*, vol. 92, no. 11, pp. 2723–2737, Nov. 2012.
[35] J.-M. Ren and J.-S. R. Jang, "Discovering time-constrained sequential patterns for music genre classification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1134–1144, May 2012.
[36] K. Markov and T. Matsui, "Music genre classification using self-taught learning via sparse coding," in *Proc. ICASSP*, Mar. 2012, pp. 1929–1932.
[37] B. L. Sturm, C. Kereliuk, and A. Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in *Proc. Int. Workshop Cognitive Inf. Process.*, to be published.
[38] 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*, 2012, pp. 205–212.
[39] D. J. Hand, "Deconstructing statistical questions," *J. Royal Statist. Soc. A (Statist. Soc.)*, vol. 157, no. 3, pp. 317–356, 1994.
[40] J.-J. Aucouturier and F. Pachet, "Representing music genre: A state of the art," *J. New Music Res.*, vol. 32, no. 1, pp. 83–93, 2003.
[41] A. Craft, G. A. Wiggins, and T. Crawford, "How many beans make five? The consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems," in *Proc. ISMIR*, 2007, pp. 73–76.
[42] A. Craft, "The role of culture in music information retrieval: A model of negotiated musical meaning, and its implications in methodology and evaluation of the music genre classification task," Ph.D. dissertation, Dept. Comput., Goldsmiths College, Univ. London, London, U.K., 2008.
[43] M. Levy and M. Sandler, "Learning latent semantic models for music from social tags," *J. New Music Res.*, vol. 37, no. 2, pp. 137–150, 2008.
[44] 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.
[45] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?," *J. Neg. Results Speech Audio Sci.*, vol. 1, no. 1, 2004.
[46] J.-J. Aucouturier, F. Pachet, P. Roy, and A. Beurivé, "Signal + context = better classification," in *Proc. ISMIR*, 2007, pp. 425–430.
[47] J.-J. Aucouturier, "Sounds like teen spirit: Computational insights into the grounding of everyday musical terms," in *Language, Evolution and the Brain: Frontiers in Linguistic Series*, J. Minett and W. Wang, Eds. Taipei, Taiwan: Academia Sinica, 2009.
[48] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Feature learning and deep architectures: New directions for music informatics," *J. Intell. Inf. Syst.*, vol. 41, no. 3, pp. 461–481, 2013.
[49] R. Hamming,, "You get what you measure," lecture at Naval Postgraduate School. Monterey, CA, USA, Jun. 1995 [Online]. Available: http://www.youtube.com/watch?v=LNhcaVi3zPA
[50] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *Proc. Int. Audio Eng. Soc. Conf.*, 2004, pp. 196–204.
[51] F. Gouyon and S. Dixon, "Dance music classification: A tempo-based approach," in *Proc. ISMIR*, 2004, pp. 501–504.



Bob L. Sturm (S'06–M'09) received the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara, California, USA, in 2009. In 2009, he was a Chateaubriand Post-Doctoral Fellow at the Institut Jean Le Rond d'Alembert, Equipe Lutheries, Acoustique, Musique (LAM), Université Pierre et Marie Curie (UPMC), Paris, France. For two years, starting in 2013, he was supported by an Independent Postdoc Grant from the Danish Agency for Science, Technology and Innovation. He is currently an Associate Professor with the Department of Architecture, Design and Media Technology, Aalborg University Copenhagen, Copenhagen, Denmark. His research interests include: signal processing, machine learning, and their applications to audio; and music information retrieval and evaluation.