Seven problems that keep MIR from attracting the interest of cognition and neuroscience

Jean-Julien Aucouturier, Emmanuel Bigand

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

Seven problems that keep MIR from attracting the interest of cognition and neuroscience

Jean-Julien Aucouturier · Emmanuel Bigand

Received: date / Accepted: date

Abstract Despite one and a half decade of research and an impressive body of knowledge on how to represent and process musical audio signals, the discipline of *Music Information Retrieval* still does not enjoy broad recognition outside of computer science. In music cognition and neuroscience in particular, where MIR's contribution could be most needed, MIR technologies are scarcely ever utilized - when they're not simply brushed aside as irrelevant. This, we contend here, is the result of a series of misunderstandings between the two fields, about deeply different methodologies and assumptions that are not often made explicit. A collaboration between a MIR researcher and a music psychologist, this article attempts to clarify some of these assumptions, and offers some suggestions on how to adapt some of MIR's most emblematic signal processing paradigms, evaluation procedures and application scenarios to the new challenges brought forth by the natural sciences of music.

PACS 43.75.Xz · 43.66.Ba

1 Introduction

The *raison-d'être* of audio Music Information Retrieval has never been the empirical pursuit of music cognition. A discipline founded on the premises of audio signal processing, pattern recognition and information retrieval, MIR[1] has found

JJ. Aucouturier
IRCAM/UPMC/CNRS STMS UMR 9912, Paris, France
E-mail: aucouturier@gmail.com

E. Bigand
LEAD/CNRS UMR 5022, Dijon, France
E-mail: bigand@u-bourgogne.fr

[1] This work primarily addresses the subset of MIR research concerned with the automatic ranking and classification of audio signals, and not the equally-important work based on symbolic musical formats. In the following, we will take the shortcut of referring to audio MIR as, merely, MIR. This does not presume that symbolic MIR should take a secondary role in this debate, of course - see e.g. [61].

its academic base in Engineering, Computer Science and Information Science departments. Yet, borrowings from such experimental and natural sciences as psychophysics, cognitive psychology and cognitive neuroscience[2] have been numerous. For instance, findings from psychoacoustical studies of musical timbre, such as the importance of spectral centroid and spectral roll-off [24], have made their way into software toolboxes [28] and technology standards [42]; the criteria used to evaluate melody transcription and retrieval systems are taken to correspond closely to the perceptual and performance thresholds measured by experimental psychology [8]; many machine learning models, e.g. most recently Deep Learning [25], take inspiration from recent findings about neural architectures in the visual and auditory cortices. More generally, there seems to be increasing awareness within the MIR community that the scientific knowledge developed by a decade of research in the field should find direct application in the "science of music". In commonsense terms, if MIR is able to classify genres and moods on tens of thousands of mp3 files with 90-something percent accuracy, "surely music cognition researchers would have a thing or two to learn from us".

If they would only listen.

In practice, recognition of MIR research in the natural sciences of music has been scarce. In many cases, musical stimuli used in cognitive psychology experiments are still characterized "by ear", either by the participants or the experimenters. For instance, a study like Balkwill & Thompson's [4] asks participants to rate music in terms of its pitch, intensity, complexity and emotions, and correlate the first three with the emotions. When computerized audio analysis is used at all, it is mostly assimilated to some sophisticated psychoacoustics (adding, say, the 6th-band spectral flux to the usual spectral centroid for timbre perception [1]), or as a means to convince readers that the experimenters' specific set of stimuli was varied and unbiased (exhibiting whatever MIR feature distribution being a convenient shortcut for the task of listening to the whole dataset [6]). The reader is left with the impression that MIR technology is an incidental element in such studies, paying lip service to an increasingly consensual requirement of "computational modeling". When "hard knowledge" about music cognition is discussed, MIR is not much invoked at all. Of the nearly 1,000 pages of the Handbook of Music and Emotion [26], for instance, not a single one is devoted to computerized signal analysis.

This is all the more so frustrating because music has become a central element in the natural sciences, and has attracted ever increasing media attention, in the same 10 years that saw the emergence and consolidation of the MIR discipline [34]. For the researcher, the mysterious "power of music" is said to hold the key to understand the evolution and development of some of our most advanced cognitive capacities: language, empathy and social cohesion [37]. For the clinician, music seems to hold promise for therapeutic applications ranging from linguistic rehabilitation, regulation of affective disorders and even stimulation of neural plasticity [48]. If ever there were an opportunity for MIR to prove its scientific and societal worth, it is now.

[2] In the following, we will collectively refer to these disciplines as the "natural sciences of music", by which we mean the study by experimental methods of the principles of perception and cognition of music; we do not address in this article other areas that either study music as a cultural artefact (e.g. musicology) or as social capital (e.g. anthropology, sociology, economy)

Scholarly interactions between MIR and the natural sciences of music are laden with mutual misunderstandings[3]. This can be illustrated by a personal story, that of the first author's failed attempt to submit what he deemed was a cognitive science argument, based on MIR experiments, to a well-known Psychology journal (an argument we later recycled, in typical academic cynicism, as a book chapter [2]). The argument was based on a large-scale study of more than 800 musical categories, each documented for a dataset of 10,000 songs, which we subjected to MIR classification with hundreds of signal processing features and machine learning optimizations (a problem that later came to be called "auto-tagging" [5]). We analysed the distribution of classification accuracy over all categories, to find that more than three-quarters of them did not reach better-than-random scores: this was true for cognitive evaluations as varied as "old-fashioned", "danceable", "makes me want to drive a fast car", but also surprisingly for categories such as genres, moods, and even instruments ("this piece includes piano"). We argued that, given that we had used a sophisticated and exhaustive optimization resulting from the collective effort of more than 10 years of MIR research, this result had probably cognitive significance: it was establishing that there was a lot less than previously thought "in the sound" of music that explained our everyday cognitive judgements. How much of it was auditory perception, and how much was extrinsically constructed without a strong auditory base? Our write-up went into review, but was rejected unequivocally: *how can you prove,* comments said, *that the failure to categorize music is not simply a failure of the algorithm? Because you do not base your argument on human judgements, but an algorithmic simulation of human judgement, your results are telling something about the properties of music recognition algorithms, but not about the properties of human cognition. You should submit this work to an Engineering journal.* These last words struck us most: it wasn't that our work was evaluated as poor cognitive psychology (which, to be fair, it probably was); it wasn't being evaluated at all. What we had naively thought was a large conceptual leap outside of our discipline into the realm of psychology was still deemed far outside of the field by its own practioners. There was no bug with our submission. It simply did not "compile" in the language of natural sciences.

This article is an attempt to identify and correct some of the "syntax errors" that MIR research makes when it attempts to address the natural sciences. These errors are not necessarily incompetencies or flaws; for the most part, they are the consequence of deeply different methodologies and assumptions that are often implicit in both fields. Questioning these assumptions is both useful extrinsically, if MIR research is to reach broader recognition outside its field, and intrinsically: it will lead to to better features (Section 2), more fruitful tasks (Section 3), and a more reflexive sense-of-purpose (Section 4), which will benefit the field directly. The reflexions in this article are based on a public conversation between the first author (a MIR researcher) and the second author (a psychologist) which occurred at the 2012 International Conference on Music Information Retrieval [3].

[3] We do need to acknowledge a few, recent positive examples [33, 53] which all the more so encouraged us to write this piece

2 Features

Problem 1: features that don't have a clear neural/cognitive modularity[4]

On first look, it is tempting to say that most, if not all, of the signal processing features used by MIR are inspired by properties of the human auditory system. For instance, Mel-Frequency Cepstrum Coefficients (MFCCs), a mathematical construct derived from the signal's Fourier transform, are often said to be designed to reproduce the non-linear tonotopic scale of the cochlea (using a Mel-scale frequency warping) and the dynamical response of the hair cells of the basilar membrane (using a logarithm as a simple compression algorithm) [47]. However, this is only partly correct. From this model-based point of departure, parts of the algorithm were added to improve the feature's computational value for machine learning, and not at all to improve their cognitive relevance. For instance, the MFCC's final discrete cosine transform (DCT) is used to reduce correlations between coefficients, which would make their statistical modeling more complex [32]. It could be argued, to some extent, that the brain uses a similar computational trick (authors like Lewicki [29] have demonstrated similar properties in physiological neural networks), but it remains to be tested whether a DCT is an appropriate model for such mechanisms. If anything, such efficient coding of information is likely to be implemented in early auditory nuclei, i.e. after the signal is encoded as spikes on the auditory nerve; but why then are the temporal integration properties of the auditory nerve not implemented before the DCT? Clearly the logics of designing features like MFCCs have more to do with maths than empirical data. The same could be said of many of the classical MIR features, e.g. the feature of spectral skewness, the 3rd spectral moment, in relation to the spectral centroid, the first moment: the latter comes from psychoacoustics, the former is only justified because it is an obvious mathematical derivation thereof.

In these conditions, constructs like the MFCCs are not easily amenable to psychological investigation, even when efforts are made to validate them with psychophysics: for instance, a study by Terasawa and colleagues [58] resynthesized sounds from MFCCs and showed that human timbre dissimilarity ratings between sounds correlated exactly with the MFCCs. This proves that an algorithmic construction, the MFCC, closely predicts a cognitive judgement. But should one conclude that the brain implements a discrete cosine transform? Probably not. It would be like concluding that jet planes demonstrate how birds fly just because they both move in air.

The traditional MIR position regarding such remarks is to argue that more is known in the psychophysiology of e.g. the visual cortex than of the auditory cortex, and it is therefore only logical that it should try to compensate the lack of physiological data by mathematic intuition. While this position may have held when the field of MIR emerged in the early 2000s, it seems less and less tenable. We now have a good understanding of the response patterns of neurons throughout the auditory pathway (the so-called spectro-temporal receptive fields [20]), and computational models even exist to model them [13,41]. When such physiologically and psychologically validated alternatives exist, it is therefore increasingly difficult

[4] we use the term modularity in its "modularity of mind" definition, following Fodor [19]

Table 1 Acoustic characteristics that best correlate with the valence and arousal of a set of musical stimuli, as computed by the MIRToolbox. Despite near-perfect predictive power, these characteristics are too mathematical to yield any cognitive or physiological interpretation. Data reproduced from Aucouturier & Bigand, 2012 [3].

Regression for valence	
Feature	β
`tonal_chromagram_peak_PeakMagPeriodEntropy`	-0.75
`tonal_mode_Mean`	0.13
`spectral_mfcc_PeriodAmp_8` (600 Hz +/- 66)	0.12
`spectral_ddmfcc_PeriodEntropy_1` (133.3)	-0.11
Regression for arousal	
Feature	β
`spectral_mfcc_Std_6` (466.6)	-0.34
`spectral_mfcc_Mean_3` (266.6)	0.28
`tonal_keyclarity_Std`	-0.28
`spectral_mfcc_Std_7` (533.3)	0.24

to justify the use of features like MFCCs in MIR studies pretending to have any relevance for cognition.

Problem 2: algorithms that look like cognitive mechanisms, but are not

Progress in both feature extraction as well as feature selection algorithms [17] means that techniques now exist to automatically combine and test very many features (predictors) to match any type of target. A software library such as the MIRToolbox [28] now offers more than 300 of such algorithms; the EDS system claims to be able to generate and explore millions of features [40]. This trend (more features, more complex, more combined) accounts for a general and constant improvement in the accuracy of MIR's classification or regression problems [60]. However, it remains very difficult to derive useful psychological intuitions from what MIR claims to be good "predictors". This is best illustrated with an example, taken from our own collaboration with the team of Bigand et al. [7]. Valence and arousal ratings were collected for a dataset of short musical extracts, and a multiple regression was conducted using the full predictive power of the MIRToolbox in an attempt to find what acoustical characteristics of sound create the emotions reported by the subjects. As shown in Table 1, we "found" that the valence of the music is very well "explained" by the *entropy of the period of the magnitude of the maximum peak detected every 50ms in the signal's chromagram*[5]. Arousal, on the other hand, was found to result from the *variance of the 6th Mel-Frequency Cepstrum Coefficient* and the *average of the 3rd* - but, apparently, not reciprocally.

Such explanations are obviously superficial. But what if cognitive psychologists were to take them literally? They would be confronted with a formidable mix-bag of a proposal: we have here an evaluation mechanism of emotional valence, of which neuroscience tells us it is at least partly pre-attentive and subcortical [10], and which MIR research seemingly suggests to explain with a series of steps involving constructs as diverse as statistical learning ("entropy"), rhythmic

[5] the chromagram, yet-another MIR construct, gives, at every time step, the energy found in the signal's Fourier spectrum in the frequency bands corresponding to each note of the octave - *c*, *c#*, *d*, etc.

entrainment ("period"), temporal integration ("maximum peak"), harmonic analysis ("chromagram") and even requiring the agent's musical training in a western culture (because the chromagram relies on the 12-tone western pitch system).

Problem 3: lack of low-level psychological validation

MIR research practices are notoriously goal-oriented. As algorithm designers, we select (manually, or automatically) a given feature because putting it in the mix helps the global task: if adding, say, spectral flux to MFCCs improves the precision of an emotion classication algorithm on a test dataset, compared to not adding it, then we add it. This design process is deeply at odds with scientific practice in the natural sciences, where dependent variables or regressors are not added to a model because they improve its explanatory power (a posteriori), but because they correspond to a theory or hypothesis (a priori) that is being tested. Such hypotheses are possible because theories are built from the bottom up, with intermediate validations of lower-level variables first [44]. Transcribed to MIR, this means a feature should not be incorporated into an algorithm before its individual predicting power for the problem being modelled is tested and validated. If we can't validate that spectral flux has any sensory or perceptual relevance to at least a part of emotion classification (if such a psychological process can be defined, see Section 3), then there is no *scientific* reason to consider it (although there may be considerable *statistical* and *pragmatic* reasons to do so).

There are considerable difficulties in this endeavour, however. First, it is unclear what subpart of the problem each feature is really addressing. Clearly, a low-level (and debatable) model of the peripheral auditory system such as the MFCCs is unlikely to do justice to such a highly cognitive process as genre recognition or emotion evaluation. There are very many intermediate steps/features that the human brain computes and integrates in its information processing workflow before reaching the auditory representation upon which such judgements are made, and one has to identify the right level of cognitive digestion of the stimuli against which to test a specific feature. The classical argument brought forth by MIR, again, is that we know very little about such "intermediate representations", and therefore it is more efficient to start with the end result in a data-driven manner [25]. This statement was accurate in the early years of the field; nowadays however, it is delusive: in the past ten years, the cognitive neurosciences have made much progress in clarifying this workflow for a large class of problems relevant to MIR, both in terms of the neural pathways that are involved and the specific properties of the audio signal which they seem to process. For instance, in the special case of the emotional evaluation of speech prosody, it is now fairly well-documented that a pathway exists (Figure 1) that can be traced back to the superior temporal gyrus (STG) in the temporal lobe, where generic auditory features are formed, with short analysis timescales in the left hemisphere and long timescales in the right [63], then projects to the right superior temporal sulcus (STS), where specific auditory representations for expressive voice are formed, and then to the right Inferior Frontal Gyrus (IFG), where these features are evaluated and confronted to linguistic/pragmatic information [51]. We contend that it is at this level of granularity

that the importance of MIR features should be evaluated: MFCCs could be tested on whether they are a good model of, say, the auditory representation assembled in the right STG, rather than a model of the final conscious evaluations formed at the end of the chain in the right IFG. A second problem then, would be how to build a groundtruth for human evaluations "made" at the level of the STG - which are not available to consciousness. *Impossible*, shrugs the MIR researcher, turning back to his/her usual groundtruth of last.fm emotion tags. Again, delusive: cognitive psychology has assembled an impressive armada of experimental paradigms that allow to quantify a large variety of auditory subprocesses. For instance, the paradigm of mismatch negativity (MMN) elicits pre-attentive auditory responses from the primary auditory cortex, in reaction to oddball sounds presented in a sequence of otherwise similar sounds [36]. The amplitude of the MMN response (measured by EEG) has been related to the strength of the sensory differences between the stimuli, which means it can be correlated with a computerized representation of the sound [59]. MMNs have been demonstrated for differences in instrument timbre [23] or emotional speech [10]. Such responses are an appropriately low level of representation to build groundtruths to evaluate and improve MIR features.

3 Tasks

Problem 4: The misguided metaphor of psychoacoustics

In first approximation, the research methodology of MIR resembles psychophysics. On the one hand, typical psychoacoustical studies, e.g. of musical timbre, proceed by collecting average similarity ratings over many users, then selecting features that explain the best percentage of the data's variance - finding, for instance, that the spectral centroid of an extract correlates at 93% with the first principal component of the human-constructed timbre space [24]. MIR, on the other hand, has shown that features such as Liu & Zhang's average autocorrelation peak [31] (or Alluri & Toiviainen's 6th band spectral flux [1]) allows to predict or classify emotions at 95% precision - a similar process, it seems, and arguably even more refined than classical psychoacoustics because of more sophisticated mapping algorithms (e.g. computing optimal embedding with SVMs instead of PCA), features and even more data (psychophysics rarely use more than a hundred stimuli, while MIR studies routinely use tens of thousands). It therefore seems incomprehensible, or even prejudiced, that the natural sciences of music should accept the outcome from the former process (e.g. the spectral centroid for timbre) as "psychologically validated" while ignoring or even rejecting results from the latter process.

In truth, for most of the behaviours investigated by MIR, there are valid reasons for doing so: the methodology of psychoacoustics is designed to investigate percepts, i.e. the immediate psychological gestalts corresponding to those few physical characteristics that define an auditory object, regardless of the listener and its culture. A musical sound has pitch, loudness and timbre, and these can be subjected to psychoacoustics (see [45], [64] and [24] respectively). The same sound, however, does not *have* genre or emotion - these are constructed cognitively; their value could change (e.g. a song may be described as "pop" instead of "rock") without changing the physical definition of a sound. Even if recent results on

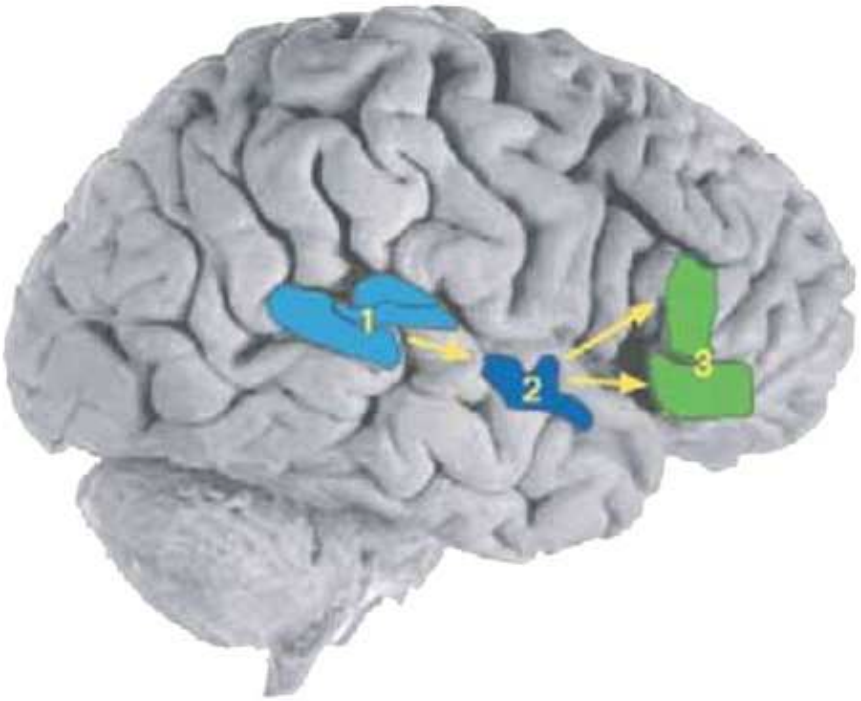

Fig. 1 Processing of emotional speech is known to activate a network of right frontal and temporal areas, including (1) the Superior Temporal Gyrus, (2) the right Superior Temporal Sulcus where auditory representations for emotions are formed and (3) the right Inferior Front Gyrus, where emotions are evaluated and (possibly) made available to consciousness. It seems more likely that the features used by MIR should correspond to earlier stages of the neural workflow, against which they should be validated, rather than to the final conscious evaluations formed at the end of the chain, and which are nevertheless routinely used as groundtruth in MIR systems. Figure adapted from Schirmer & Kotz, 2006 [51]

action-perception have begun challenging the frontier between what's a percept and what's a cognitive construction (see e.g. [39] on the question of emotion perception), most in cognition would still agree that a fundamental difference remains between the two. For them, MIR research is applying the psychoacoustics metaphor, i.e. searching for acoustic correlates, to behaviours (genres, emotions, etc.) to which it does not apply.

A better principled way for MIR to describe the insights of its typical genre or mood classification algorithms would be to recognize that these are not capturing "rock" or "sad" music, but rather *things that sound like* "rock", or *things that sound like* a "sad song". Good performance for such systems, then, may be less revealing of human cognitive processes than of a sociological reality: because music is a structured human activity with a bias for regularity [53], most "sad" music indeed sounds the same (dark timbre, low pitch, etc.). These features do

not necessarily make the music sad: one can find music that is experienced as sad without exhibiting any of these features (for a particularly extreme example, see [9]). But because such songs are rare (or at least they're rare in typical test datasets), e.g. there are maybe 10% of them, MIR systems can still reach 90% performance without actually modeling anything specific about how, e.g. genre is cognitively constructed. Unlike true psychoacoustics, the high precision values found in MIR do not presume anything has been learned about music cognition.

Problem 5: Wrong[6] incentives: tasks that are too far away from perception

It follows that, for MIR research to bridge the gap between its audio signal processing expertise and the concerns of e.g. cognitive psychology, tasks like genre and mood classification, or even broadly defined music similarity, which form a large share of the annual MIREX evaluation campaign [43], may be counter-productive. Most psychologists would consider that e.g. musical genre, as an object of study, is too complex, i.e. it is known in advance that studying it won't help isolate experimentally any particular process that could constitute it. For instance, if one wants to understand the sensory process by which a rock song is recognized as rock, it is simpler and more elementary to study the same process for environmental sounds. This latter case is less plagued by cultural learning, ambiguity and subjectivity than musical genre. For the MIR researcher too, any optimization of features or machine learning over such problems is unlikely to translate into novel sensory/cognitive insights, but rather result in ever finer adaptation, in the best case, to the existing socio-statistical patterns in music production and consumption[7] and, in the worst case, to the fallacious specificities of the field's evaluation metrics [60]. Yet, we have seen a trend in recent years to both neglect the development of new features and to favour the emergence of new tasks and new evaluation corpuses. The 2012 MIREX competition now includes 19 tasks, in which all competing implementations are typically using the same features, with minute parameter changes [43]. Tasks as defined by the MIREX initiative are powerful incentives that shape where the discipline is going, year after year, and therefore should not be taken lightly.

From this perspective, it may be appropriate for the field to declare a temporary "moratorium"[8] on such emblematic tasks as genre or mood classification or music similarity, and refocus on a core set of better-defined tasks, of a lower-level nature,

[6] By "wrong", we do not mean that the typical attitude of MIR research is mistaken or flawed. If anything, the tasks and methodologies discussed here are largely to credit for the many and important technological successes achieved in the MIR community. By calling these "wrong", we propose however that these attitudes, while arguably beneficial for the engineering purposes of MIR, are also harmful to the interdisciplinary dialog between MIR and the natural sciences of music. These are "the wrong things" for a MIR practionner to do when addressing a psychologist. See also Problem 6 below.

[7] a valuable topic of investigation in its own right, see e.g. [49]

[8] a strong initiative not without precedent, e.g. when in 2009 the Python community stopped all changes to the language's syntax for a period of two years from the release of Python 3.1., in order to let non-CPython implementations "catch up" to the core implementation of the language [11]

more likely to generate insights about human perception of music, and for which well-defined measure paradigms already exist, e.g.

- musical instrument recognition (for which neuroscience has already started producing valid computational models [41])
- detection of dissonance in series of chords [46]
- pre-attentive responses to deviants in e.g. timbre or prosodic sequences [23].
- any of the individual mechanisms for emotion induction reviewed by Juslin & Västfjäll (2008) [27], to be measured by implicit paradigms [22] or indirect cognitive effects [35].

While some of these may temporarily lead us away from the immediately utilitarian goals of MIR (genre recognition for carps, anyone? [12]), features and systems developed with such goals will not only benefit the field extrinsically (better interaction with the natural sciences, and broader recognition), but also intrinsically. A comparison with the neighboring discipline of computer vision may help here: in the same time MIR has been increasing its scope and the number of its evaluation tasks, computer vision has refocused its effort on a small set of extremely narrowly-defined problems: object segmentation, object detection and object recognition, each with its own standardized testbed and a wealth of neuroscientific data to build on [56]. This has led to the development of computer models that follow biological constraints radically and which overperform their previous, non-biologically-plausible alternatives [54]. Furthermore, such models are now routinely used in the vision cognition community to explain experimental results [14].

4 Purpose

Problem 6: Wrong attitude to limit cases

While both communities are trying to model and simulate processes of music perception, MIR and the natural sciences of music have fundamentally different purposes. MIR is interested in the result of its simulation, and how much it matches the outcome of human behaviour. A science like music cognition is less interested in outcomes than in processes[9]. If its interest lies in designing computer algorithms to e.g. do maths, it would be less interested in building machines that can multiply numbers as well and as fast as humans, but rather in building them in such a way that multiplying $8x7$ is more difficult than $3x4$, as it is for humans [16]. This bias is evident in the way MIR treats the limit cases of the evaluation of its algorithms. Training data is assigned to classes, which are decided once and for all (the "groundtruth"), and then serve as evaluation criteria regardless of their largely varying degrees of stereotypicality. MIR algorithm designers are happy e.g.

[9] One reviewer of this article went even further to say that MIR is merely a category of vocational training, and therefore does not belong to the scientific community and should not be expected to be capable to generate scientific questions. While this view is historically accurate, we believe recent years have seen increasing academic migration between these both extremes, with MIR researchers turning to traditional scientific disciplines such as psychology or neuroscience, and conversely, graduates from such fields coming to MIR in their postgraduate or postdoctoral days. In any case, one is forced to consider that much of the obstacles identified in this article could be addressed by more systematic training for MIR students in the methods of empirical sciences, incl. experimental design and hypothesis testing.

when they see their algorithms duly classify as "rock" certain songs that are clearly on the border of that definition (e.g. a song like Queen's Bohemian Rapsody), and they would be just as well if the groundtruth had arbitrarily labelled them "pop" instead. Algorithm optimization then purposes to "fix" as many false negatives as possible, at all cost, even when there are good sensory reasons for a given "rock" song to be classified as something else. Psychologists, on the other hand, would rather understand what makes a song more prototypically "rock" than another, or to how much "rock" one has to listen to form a stable representation of that genre.

Flattening out limit cases, as MIR routinely does, has multiple bad consequences. First, typical MIR algorithms tend to err very little, but when they do, they do in very ungraceful ways: they are not optimized to be solid on stereotypical cases and flexible elsewhere. More often than not, genre recognition algorithms accept "Bohemian Rhapsody' where they reject "We will rock you" [55]; nearest neighbors that occur most frequently in music similarity systems are also those that have least perceptual similarity with the target [18]. Second, averaged groundtruths tend to produce average algorithms, in terms of how they apply to specific populations. While most human subjects tend to agree on most stereotypical instances, there are important inter-individual differences in the margins of the class distributions - for instance, in the case of emotion recognition, based on culture [4], musical expertise [7], age [30] or even individual personality [62], all specificities that MIR algorithms and groundtruths are in effect neglecting [50] while they are some of the most central objects of study for the natural sciences of music.

Problem 7: Trying too hard: MIR as a physical model

We already discussed the confusion that ensues when publicizing MIR as sophisticated, generalized psychoacoustics (Section 3). Working on problems that are non purely perceptual, MIR is sentenced to only produce incidental explanations, of *things that sound like* a rock or sad song, but no definite model of what could make them so. There is solace in this situation, however, because music cognition research has precisely been missing this capacity to measure "how things sound", in order to control its experimental stimuli and separate what's in the physical signal from what's constructed out of it by cognition. It is ironical that MIR may be most useful to cognition when it provides tools which do not have the pretense of infringing into cognitive thinking, but instead just pure, state-of-the-art physical modeling [52].

More precisely, MIR can be understood to provide a physical measure of the information available for human cognition in the musical signal for a given task. This measure can then be used to construct scientific proofs in a variety of situations. For instance, in the speech domain, de Boer & Kuhl [15] have shown that speech recognition algorithms (hidden Markov models - HMM) have better word recognition performance when they are trained and tested on infant-directed speech (IDS, or "motherese") than on adult speech. This can be taken to validate the argument that the developmental value of IDS is to bootstrap language learning. It is particularly notable in this example that the algorithm is not presented as a cognitive model: the authors are not pretending that the human brain im-

plements a HMM. Their result only gives a proof of feasibility, i.e. that, from a purely physical point of view, the information exists in the IDS signal to allow for an easier treatment than adult speech; it does not show that the information is actually used. Still, it would be very difficult to replace the machine by a human measure in this argument. A similar argument is developed in the work of Kaplan and colleagues [38], who show that machine learning can classify dog barks into contexts like being afraid, playful, etc. This was taken to indicate, for the first time, that dog vocalizations contain communicative features.

The difference between such type of proofs and the rejected claim of [2] is subtle but important. The latter attempted to prove a negative hypothesis: if MIR models based on specific audio features cannot easily classify this or that, then it is taken to indicate that, contrary to intuition, human classification cannot be based on these features. The fallacy there resides in the impossibility to prove that human cognition had absolutely no means to exploit such features in ways that algorithms could not. An equally valid (and quite possibly more likely) conclusion to the same "experiment" is therefore that the tested MIR models simply failed to process music in the same way as humans. The claims brought forth in works like [15] and [38] start from the opposite premise: they set to prove that, contrary to intuition, a set of features or processes are *enough* (instead of *not enough*) to allow correct classification. It is indeed an open experimental question whether, say, dog barks or infant cries, contain enough acoustical information to communicate context; after all, context may be decoded from other co-varying cues such as body posture, situation, etc. It is experimentally very difficult to control for such other cues, and restrict human subjects to use only acoustic information: even if they fail to detect context in, e.g., audio-only recordings, it can be taken to indicate that the task is not ecological, and that with proper training, maybe they could. The only "clean" way to test for this hypothesis is arguably to construct a naive machine able to exploit auditory-only information, which is precisely what MIR allows to do. While such studies cannot prove that humans actually exploit this information, it does prove that exploiting it would be enough.

Modern psychology is particularly found of such parsimonious explanations [21,57]. This type of proof could be applied to a variety of important cognitive problems in music, including, e.g., how listeners from one culture identify emotions in music from other cultures; how non-musicians detect deviations in sequences of chords despite no formal knowledge of harmony; how one forms a workable representation of a musical genre after listening to only a couple of examples, etc.

5 Conclusion

In summary, while the opening example of our rejected submission was attempting to cast a pattern recognition result into a cognitive result, and failed under the fallacy of negative hypothesis testing, we propose that there are ways to construct scientific proofs with MIR that address central problems in the natural sciences of music. By using MIR as a tool for physical, rather than cognitive, modeling, proofs of feasibility can be made for parsimonious mechanisms of information processing that would otherwise be difficult to demonstrate; by examining musical behaviours at a lower level than e.g. genre, similarity or mood judgements, for instance in pre-attentive or unconscious contexts, MIR features can be used to test hypotheses

on what type of signal characteristics are important for early auditory processing; by insisting on using features with clear, uncompromised neural modularity (e.g. spectro-temporal receptive fields), MIR has an opportunity to bring unprecedented signal-processing sophistication to cognitive neuroscience and psychology. Adapting MIR's body of expertise to the needs and purposes of the natural sciences such as experimental psychology or cognitive neuroscience may require a leap of faith, and a willingness to challenge some of the field's favorite tools (MFCCs...), some of its most emblematic tasks (genres...), and some of its most foundational application scenarios (massive, one-size-fit-all classification). But such may be the cost to pay for reaching broader recognition outside of the computer science/engineering community, for overcoming the recent stalling in feature and model development, and for transitioning into the field's second decade of activity with a renewed sense of purpose.

Acknowledgements We wish to credit Gert Lanckriet (UCSD), Juan Bello (NYU) and Geoffroy Peeters (IRCAM) for an animated discussion at ISMIR 2012 leading to the idea of a moratorium on all non-essential MIR tasks, from which we borrowed some of the thinking in the present Section 3.

References

1. Alluri, V., Toiviainen, P.: Exploring perceptual and acoustic correlates of polyphonic timbre. Music Perception **27(3)**, 223–241 (2010)
2. Aucouturier, J.J.: Sounds like teen spirit: Computational insights into the grounding of everyday musical terms. In: J. Minett, W. Wang (eds.) Language, Evolution and the Brain. Frontiers in Linguistics Series (2009)
3. Aucouturier, J.J., Bigand, E.: Mel cepstrum and ann ova: the difficult dialogue between mir and cognitive psychology. In: Proc. of the 13th International Conference on Music Information Retrieval, Porto, Portugal (2012)
4. Balkwill, L., Thompson, W.F.: A cross-cultural investigation of the perception of emotion in music: Psycho-physical and cultural cues. Music Perception **17**, 43–64 (1999)
5. Bertin-Mahieux, T., Eck, D., Maillet, F., Lamere, P.: Autotagger: a model for predicting social tags from acoustic features on large music databases. Journal of New Music Research **37(2)**, 151165
6. Bigand, E., Delbé, C., Tillmann, B., Gérard, Y.: Categorisation of extremely brief auditory stimuli: Domain-specific or domain-general processes? PLoS ONE **6(10)** (2011)
7. Bigand, E., Vieillard, S., Madurel, F., Marozeau, J., Dacquet, A.: Multidimensional scaling of emotions responses to music: effect of musical expertise and duration. Cognition & Emotion **19**, 1113–39 (2005)
8. Birmingham, W.P., Meek, C.J.: A comprehensive trainable error model for sung music queries. Journal Of Artificial Intelligence Research **22**, 57–91 (2004)
9. Bonini, F.: All the pain and joy of the world in a single melody: A transylvanian case study on musical emotion. Music Perception **26(3)**, 257–261 (2009)
10. Bostanov, V., Kotchoubey, B.: Recognition of affective prosody: Continuous wavelet measures of event-related brain potentials to emotional exclamations. Psychophysiology **41**, 259–268 (2004)
11. Cannon, B., Noller, J., van Rossum, G.: Python language moratorium. Python Enhancement Proposals (PEPs) 3003, available: http://www.python.org/dev/peps/pep-3003 (2009)
12. Chase, A.R.: Music discriminations by carp (cyprinus carpio). Animal Learning & behavior **29(4)**, 336–353 (2001)
13. Chi, T., Ru, P., Shamma, S.: Multi-resolution spectrotemporal analysis of complex sounds. Journal of Acoustical Society of America **118(2)**, 887–906 (2005)
14. Crouzet, S., Kirchner, H., Thorpe, S.: Fast saccades toward faces: Face detection in just 100 ms. Journal of Vision **10(4)-16**, 1–17 (2010)

15. De Boer, B., Kuhl, P.: Investigating the role of infant-directed speech with a computer model. Acoustics Research Letters Online **4(4)**, 129–134 (2003)
16. Dehaene, S.: Varieties of numerical abilities. Cognition **44**, 1–42 (1992)
17. Fiebrink, R., Fujinaga, I.: Feature selection pitfalls and music classification. In: Proc. International Conference on Music Information Retrieval (2006)
18. Flexer, A., Schnitzer, D., Schlueter, J.: A mirex meta-analysis of hubness in audio music similarity. In: Proc. 13th International Conference on Music Information Retrieval, Porto, Portugal (2012)
19. Fodor, J.: Modularity of mind: an essay on faculty psychology. Cambridge, Mass.: MIT Press (1983)
20. Ghazanfar, A., Nicolelis, M.: The structure and function of dynamic cortical and thalamic receptive fields. Cerebral Cortex **11(3)**, 183–93 (2001)
21. Gigerenzer, G., Todd, P.M.: Simple heuristics that make us smart. New York: Oxford University Press (1999)
22. Goerlich, K., Witteman, J., Schiller, N., Van Heuven, V., Aleman, A., Martens, S.: The nature of affective priming in music and speech. J. Cogn. Neurosci. **24(8)**, 1725–41 (2012)
23. Goydke, K., Altenmller, E., Möller, J., Münte, T.: Changes in emotional tone and instrumental timbre are reflected by the mismatch negativity. Cognitive Brain Research **21(3)**, 351–9 (2004)
24. Grey, J.M.: Multidimensional perceptual scaling of musical timbres. Journal of the Acoustical Society of America **61**, 1270–1277 (1977)
25. Humphrey, E.J., Bello, J.P., LeCun, Y.: Moving beyond feature design: Deep architectures and automatic feature learning in music informatics. In: Proc. 13th International Conference on Music Information Retrieval, Porto, Portugal (2012)
26. Juslin, P., Sloboda, J.: Handbook of Music and Emotion. Oxford University Press (2010)
27. Juslin, P., Vstfjll, D.: Emotional responses to music: the need to consider underlying mechanisms. Behavioural and Brain Sciences **31**, 559–621 (2008)
28. Lartillot, O., Toiviainen, P.: A matlab toolbox for musical feature extraction from audio. In: Proceedings of the 10th Int. Conference on Digital Audio Effects, Bordeaux, France (2007)
29. Lewicki, M.: Efficient coding of natural sounds. Nature Neuroscience **5(4)**, 356–363 (2002)
30. Lima, C.F., Castro, S.L.: Emotion recognition in music changes across the adult life span. Cognition and Emotion **25(4)**, 585–598 (2011)
31. Liu, D., Zhang, H.J.: Automatic mood detection and tracking of music audio signal. IEEE Transactions on Speech and Audio processing **14(1)**, 5–18 (2006)
32. Logan, B.: Mel frequency cepstral coefficients for music modeling. In: Proc. 1st Int. Conf. on Music Information Retrieval, Plymouth, MA, USA (2000)
33. MacCallum, B., Mauch, M., Burt, A., Leroi, A.M.: Evolution of music by public choice. Proceedings of the National Academy of Sciences (2012)
34. Mannes, E.: The Power of Music: Pioneering Discoveries in the new science of song. Walker & Co (2011)
35. Masataka, N., Perlovsky, L.: The efficacy of musical emotions provoked by mozart's music for the reconciliation of cognitive dissonance. Scientific Reports **2** (2012)
36. May, P.J.C., Tiitinen, H.: Mismatch negativity (mmn), the deviance-elicited auditory deflection, explained. Psychophysiology **47**, 66122 (2010)
37. Mithen, S.: The Singing Neanderthal: The Origins of Music, Language, Mind, and Body. Cambridge: Harvard University Press (2007)
38. Molnár, C., Kaplan, F., Roy, P., Pachet, F., Pongrácz, P., Dóka, A., Miklósi, .: Classification of dog barks: a machine learning approach. Animal Cognition **11(3)**, 389–400 (2008)
39. Niedenthal, P.M.: Embodying emotion. Science **316(5827)**, 1002–1005 (2007)
40. Pachet, F., Roy, P.: Analytical features: a knowledge-based approach to audio feature generation. EURASIP Journal on Audio, Speech, and Music Processing (2009(1))
41. Patil, K., Pressnitzer, D., Shamma, S., Elhilali, M.: Music in our ears: The biological bases of musical timbre perception. PLoS Computational Biology **8(11)** (2012)
42. Peeters, G., McAdams, S., Herrera, P.: Instrument sound description in the context of mpeg-7. In: Proceedings of the International Computer Music Conference, Berlin, Germany (2000)
43. Peeters, G., Urbano, J., Jones, G.J.F.: Notes from the ismir 2012 late-breaking session on evaluation in music information retrieval. In: Proc. 13th International Conference on Music Information Retrieval, Porto, Portugal (2012)

44. Platt, J.R.: Strong inference. Science **146(3642)**, 347–353 (1964)
45. Pollack, I.: Decoupling of auditory pitch and stimulus frequency: The shepard demonstration revisited. Journal of the Acoustical Society of America **63** (1978)
46. Poulin-Charronnat, B., Bigand, E., Koelsch, S.: Processing of musical syntax tonic versus subdominant: An event-related potential study. Journal of Cognitive Neuroscience pp. 1545–1554 (2006)
47. Rabiner, L.R., Juang, B.H.: Fundamentals of Speech Recognition. Prentice-Hall (1993)
48. Sacks, O.: Musicophilia: Tales of Music and the Brain (2008)
49. Salganik, M.J., Dodds, P., Watts, D.J.: Experimental study of inequality and unpredictability in an artificial cultural market. Science **311(5762)**, 854856
50. Schedl, M., Flexer, A.: Putting the user in the center of music information retrieval. In: Proc. 13th International Conference on Music Information Retrieval, Porto, Portugal (2012)
51. Schirmer, A., Kotz, S.: Beyond the right hemisphere: Brain mechanisms mediating vocal emotional processing. Trends in Cognitive Sciences **10**, 24–30 (2006)
52. Serra, J.: Is pattern recognition a physical science? In: 15th International Conference on Pattern Recognition, Barcelona, Spain (2000)
53. Serra, J., Corral, A., Boguna, M., Haro, M., Arcos, J.L.: Measuring the evolution of contemporary western popular music. Scientific Reports **2** (2012)
54. Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T.: Object recognition with cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence **29(3)**, 411–426 (2007)
55. Sturm, B.: Classification accuracy is not enough: On the analysis of music genre recognition systems. Journal of Intelligent Information Systems, this issue. (2013)
56. Szeliski, R.: Computer Vision: Algorithms and Applications (2011)
57. Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J.B., Bonatti, L.L.: Pure reasoning in 12-month-old infants as probabilistic inference. Science **332**, 1054–1059 (2011)
58. Terasawa, H., Slaney, M., Berger, J.: The thirteen colors of timbre. In: Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA (2005)
59. Toiviainen, P., Tervaniemi, M., Louhivuori, J., Saher, M., Huotilainen, M., N^ätänen, R.: Timbre similarity: convergence of neural, behavioral and computational approaches. Music Perception **16**, 223–241 (1998)
60. Urbano, J., Downie, J.S., McFee, B., Schedl, M.: How significant is statistically significant? the case of audio music similarity and retrieval. In: Proceedings of 13th International Conference on Music Information Retrieval, Porto, Portugal (2012)
61. Volk, A., Honingh, A.: Mathematical and computational approaches to music: Three methodological reflections. Journal of Mathematics and Music **6(2)** (2011)
62. Vuoskoski, J.K., Eerola, T.: The role of mood and personality in the perception of emotions represented by music. Cortex **47(9)**, 1099 (2011)
63. Zatorre, R., Belin, P.: Spectral and temporal processing in human auditory cortex. Cerebral Cortex **11**, 946953 (2001)
64. Zwicker, E.: Procedure for calculating loudness of temporally variable sounds. Journal of the Acoustical Society of America **62**, 675 (1977)