

# Sounds Like Teen Spirit: Computational Insights into the Grounding of Everyday Musical Terms

Jean-Julien Aucouturier

*Ikegami Laboratory, Graduate School of Arts and Sciences,  
The University of Tokyo, Japan.*

## Abstract

The diversity of symbolic dimensions along which we think about music in our everyday listening experience is puzzling. Songs are commonly said to be “energetic”, to make us “sad” or “nostalgic”, to sound “like film music”, to be perfect to “drive a car on the highway” among many other similar metaphors. Such descriptions are generally considered as well-defined sensory constructs, strongly coupled to the acoustic properties of the corresponding musical stimuli. Yet 10 years of computer pattern recognition research for musical audio signals have failed to reliably characterize such mappings. This either means that pattern recognition needs to develop richer models of human auditory processing, or that the words we use to talk about music are less heavily based on acoustic similarity than we usually think. In this chapter, we pursue the latter hypothesis. We show that even if their acoustic mapping is weak, typical audio categories can be reliably predicted when considering inter-symbolic associations. We propose a computational model to re-integrate pattern recognition techniques into this larger vision, and discuss its implications for the evolution and learnability of human symbolic systems.

## 1. The Many Words of Music

There are a variety of dimensions along which human listeners appreciate music, and a variety of terms with which these can be communicated. A popular song like “The Beatles–Yesterday” could be described as e.g. pop-rock, acoustic, mellow, nostalgic, cheesy—among a large number of other possible taxons. Such descriptions can be organized in many semantic dimensions, such as musical genre, instrumentation, mood, lyric content, appreciation of quality, etc. which translate the very diverse and multi-layered meanings that music evokes in human listeners. Musical descriptions typically have varied degrees of consensus among listeners. Appreciations of timbre, rhythm or harmony are generally considered as objective constructs (related to the underlying generative and physical process),<sup>1</sup> while other descriptions such as musical genre or mood can be understood as social constructs which are semiotically related to a particular musical object category without being structurally intrinsic to that category, and will typically depend on individual subjectivity, culture and experience.

However subjective and local to some community of listeners, many of these descriptions are extremely manifest in our culture, and central to the ways we think and interact about music. For instance, symbols such as musical genres strongly resist objective definition. Pachet & Cazaly (2000) compare 3 genre taxonomies used in online music services on the WWW: allmusic.com (531 genres), amazon.com (719 genres) and mp3.com (430 genres), and show that there is no consensus in the labels used in these classifications: only 70 words are common to the three taxonomies. More importantly, categories with the same label do not have the same definition in extension: even largely used labels like Rock or Pop do not denote the same set of songs. Nevertheless, library studies (Bainbridge et al., 2002) show that genre is systematically the most common metadata used for musical searches, when exact bibliographic information (such as author and title) is not available. Genre-like musical symbols are undoubtedly useful, even when we don’t agree on them.

The prevalence of such high-level descriptions in our everyday experience of music suggests, from a common-sense perspective,

that these descriptions are well-defined sensory constructs, strongly coupled to the acoustic properties of the corresponding musical stimuli. Surely there is music which “sounds like” rock, jazz or techno; “something” in certain sounds which feels “smooth”, “metallic” or “stressful”.

A considerable amount of psycho-physical research has been done to find the acoustical correlates to the psychological sensations involved with music listening. For methodological reasons however, most effort has focused on the atomic properties of single instrumental tones, namely pitch, loudness and timbre (for timbre, see e.g. Grey, 1977 or McAdams et al., 1995). Few psychophysical studies<sup>2</sup> have targeted the high-level, symbolic constructs we’re concerned with here: High-level descriptions “à la genre” are typically developed over the time scale of a few seconds and expressed in polyphonic settings, which makes the manual exploration of potentially correlated physical attributes considerably more difficult than for single tones.

An alternative approach to psychoacoustics is to rely on computer pattern recognition<sup>3</sup> to try to reproduce human categorical judgments of music, based on the sole acoustic content of the audio signals. Typically, musical audio signals are cut into short overlapping frames (typically 50ms with a 50% overlap), and for each frame, a mathematical encoding (aka a set of features) of the sample values is computed. Features usually consist of a generic, all-purpose spectral representation such as Mel Frequency Cepstrum Coefficients, a particular encoding of the spectral envelope also used for automatic speech recognition (Rabiner & Juang, 1993). All feature vectors in the stimuli are then fed to a classification algorithm which models the global statistical distribution of the features of signals corresponding to each class (e.g. rock or jazz in the case of a genre classification system). Global distributions for each class can then be used to compute decision boundaries between classes. A new, unobserved signal is classified by computing its feature vectors, finding the most probable class for each of them, and taking the overall most represented class for the whole signal.

This approach is a widely adopted paradigm in the research community concerned with automatic music recognition (as manifest e.g. in the past ISMIR conferences<sup>4</sup>). It has been used to model a

very large spectrum of musical classifications, many of which are relevant for our present study: genre (Tzanetakis et al., 2001), mood (Liu et al., 2003), instrument (Vincent & Rodet, 2003), potential for commercial success (Dhanaraj & Logan, 2005), etc. However, recent research increasingly suggests that this approach is intrinsically bounded to moderate performance:

- **Glass ceiling:** Many years of quasi-exhaustive search in the space of algorithmic variants (features and models) and parameters have failed to improve the precision of such systems above an empirical glass-ceiling (Aucouturier & Pachet, 2004), around 70% precision (although this of course should be defined precisely and depends on tasks, databases, etc.).
- **The more plausible biologically, the worse:** State-of-the-art algorithms typically do not take any account of temporal ordering: a piece of music where all successive bits of 100ms are shuffled in random order will generate the same representation as the original one—which contradicts elementary perceptual evidence. Yet, traditional means to integrate temporal ordering into these models (e.g. delta-coefficients or hidden Markov models) typically do not provide any improvement over the best static models for real-world, complex polyphonic textures of several seconds length (Aucouturier & Pachet, 2006; Scaringella & Zoia, 2005). Similarly, attempts to integrate biologically-plausible models of the human auditory system (such as spectral mapping between neighboring frequency components) have been found to slightly degrade the performance (Lidy & Rauber, 2005).

### *1.1. The Nature of Musical Categories*

One way to look at the relative failure of the above approaches is to state, simply, that more work is needed. The fact that state-of-the-art pattern recognition algorithms should fail to capture the physical correlates of given musical genre or mood categories does not rule

out the possibility that such mappings exist, only in more intricate manifolds than currently analysed. These may require, for example, more specific musical features (a precise note-to-note transcription of the underlying musical score, for instance) and better data representations (human cognition is notoriously better at singling out simple structures in high dimensional data than our best artificial neural networks).

But there is an alternative point of view on the same observation. What if the current pattern recognition models were perfect? What if 70% were in effect all that there is out there to find, i.e. that the words we use to talk about music were less heavily based on acoustic similarity that we usually think? When we categorize a piece of music as being “rock music”, how much do we hear “rock”, and how much do we know or infer it from information that is not directly available in the musical signal?

It is this point of view that we wish to pursue here. We will show how imperfect acoustic-only models can still be used to explain the grounding of musical categories, thus reformulating the goal of pattern recognition into a larger “cognitive” framework. To this aim, we will use data from a recent computational study of a large and heterogeneous set of high-level musical symbols, made available through research partnerships: 5,000 songs, each described with more than 800 categories. First, we will show that, indeed, only very few of the typical descriptions we routinely make about music can be univocally correlated to a stereotypical “sound” of the corresponding audio stimuli (as analysed by state-of-the-art pattern recognition). Second, we will observe that there is nevertheless a wealth of extrinsic information available for learning, in the network of correspondences between musical categories: however it sounds, “rock” music uses “guitar”, sounds “exciting” and talks about “youth”. We will then present a model (the so-called “bootstrap model”) explaining how musical symbols can be grounded to musical signals using such inter-symbolic correlations, validate its relevance as an engineering tool, and make a number of cognitive predictions regarding the learnability for the resulting symbolic system.

Note that, for the sake of brevity, only minimal space will be devoted here to the technical description of the algorithms on

which we base our discussion (next section). The interested reader is suggested to refer to e.g. (Aucouturier et al., 2007b; Aucouturier et al., 2007a; Aucouturier et al., 2005).

## 2. Methods

### 2.1 Data

We base our study on a large set of human-made judgments of high-level musical descriptions, collected for a large quantity of commercial music pieces. The data is proprietary, and made available to the author by research partnerships. The database contains 4,936 songs, each described by a set of 801 Boolean attributes (e.g. “Mood happy” = true). These attributes are grouped in 18 categories, which can be found in Table 1.

**Table 1. Categories of the attributes used in the database.**

Category	Number of attributes	Example attribute
Aera/Epoch	16	1970–1980
Affiliate	5	Germany
Character	39	Child-oriented
Country	31	Brazil
Dynamics	4	Decreasing
Genre	36	Jazz Standard
Language	15	Spanish
Main Instrument	107	Contra Bass (pizz.)
Metric	14	3/4
Mood	58	Aggressive
Musical Setup	25	String Ensemble
Rhythmics	10	Groovy
Situation	82	City By Night
Special Creative Period	3	Early
Style	176	Bebop
Tempo	8	Slow – Adagio
Text Category	123	Forgiveness
Variant	46	Natural / Acoustic

Attribute values were filled in manually by human listeners, under a process related to Collaborative Tagging, in a business initiative comparable to the Pandora project.<sup>5</sup> Each song in the database was annotated by several persons, and results agglomerated by thresholding techniques (“rock” songs are songs which were significantly often tagged as “rock”). The high-level descriptions found in the database are very diverse. Some attributes directly describe some physical property of the sound (“Main Instrument”, “Dynamics”), while others seem to result from a more cultural view on the music object (“Genre”, “Mood”, “Situation”).

One should note that such category taxonomies are not intended for universality: the definition of attributes such as “Style Alternative Rock” and how they differ from, say, “Style Rock-Pop” is a convention which is only local to the tagging community, and may not be made explicit easily. In many collaborative tagging systems, tags are not primarily intended for direct informative display, but rather for creating a mid-level representation which can be used for matching and recommendation (“if two items share the same set of tags, then they are probably similar”). For all these reasons, we propose here to analyze this set of attributes as an arbitrary ontology only defined by the values taken on the database, and not to consider any exterior musical assumption of what a “Genre” or “Style” should be.

## 2.2 Computer Measure of Acoustic Similarity

We propose to simulate human judgments of the acoustic similarity between musical stimuli by using a computational measure previously introduced in the pattern recognition community (Aucouturier et al., 2005). The algorithm takes two audio signals as input, and outputs a numerical value (in the range  $[0, \infty]$  which quantifies their acoustic (dis)similarity. It considers music holistically, as a succession (or distribution) of short sound samples (or frames) which are described in a very generic manner (some encoding of their Fourier spectrum, like MFCCs). Hence, the measure doesn’t make any assumption about analytical quantities that may be important in the musical signal. It captures a bit of some generalized notion

of “timbre” (i.e. statistics of spectral shapes), but also accounts for some “rhythm” (e.g. the density of percussive frames over time, as these have particular spectral signatures), and some “harmony” (again, embodied in the specific spectral signatures of e.g. single tones, complex chords or non-harmonic sounds). See Appendix 1 for a technical description of the algorithm. The measure was found to approximate human judgments with near-perfect precision for natural sound environments (e.g. urban soundscapes, park, street, boulevard, etc.) (Aucouturier et al., 2007a), and with reasonable precision for music—it captures the holistic sensation that one usually denotes as “this sounds like ... (Beethoven’s 9th Symphony, The Beatles, etc.)” See e.g. (Aucouturier et al., 2007a) for a technical discussion of the algorithm’s performance and limitations.

### ***2.3 Evaluation of Mapping Strength***

We propose to study here the strength of the mapping between a given high-level description (or attribute) and the acoustic quality of the corresponding signals. To do so, we propose to evaluate the precision of an inference mechanism based on the above acoustic similarity, in which a test song  $S$  is given an attribute value  $A(S)$  (e.g. “mood violent = true”) if a sufficient number of songs which sound similar to  $S$  have the same value (e.g. “a song is violent if it sounds like a lot of violent songs”). A formalized description of the algorithm can be found in (Aucouturier et al., 2007b). This mechanism is usually called nearest neighbor classification in the pattern recognition community (Bishop, 1995).

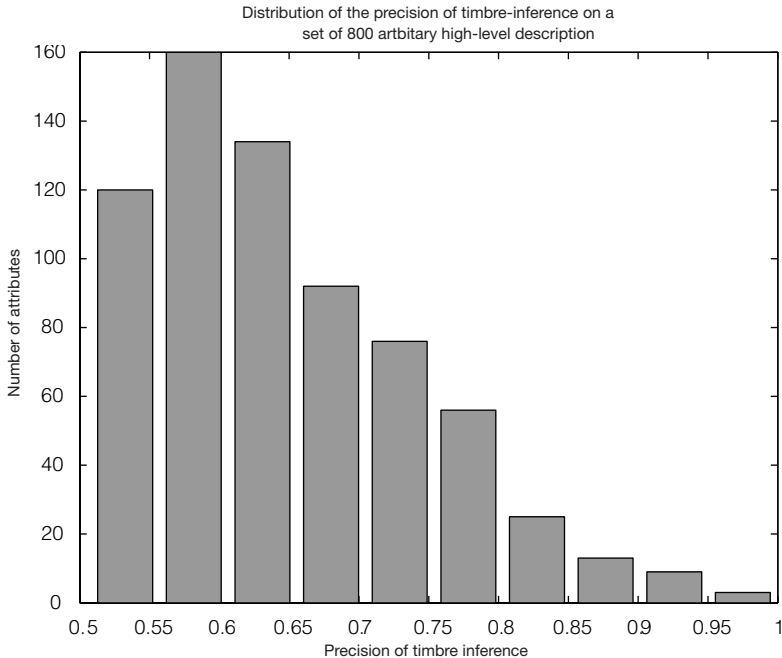
We measure the strength of the mapping between a given attribute  $A$  and the “sound” property of the associated musical signal as the precision of nearest neighbor classification when used to infer values of  $A$ . If there is a strong acoustic grounding for a given symbol, then nearest neighbor classification is expected to work with high precision. On the contrary, if a high-level description has no correlation to acoustic properties (e.g. whether the song title writes with more or less than 8 letters), then the precision of the inference mechanism should be low: the fact that there is a certain proportion of attribute values in the acoustic neighborhood of the test song does

not convey any useful information to infer the attribute value for the test song. In other words, “this sounds like a 7-letter song” is a pretty absurd statement.<sup>6</sup>

### 3. Musical Terms Are Motivated Rather Than Predicted

We observe from Figure 1 that there are surprisingly few categories in the large set we consider that can be well predicted from our simulated measure of acoustic similarity. Only 6% of the attributes in the database are estimated with more than 80% precision, and more than a half of the database’s attributes are estimated with less

**Figure 1. Distribution of the precision of acoustic inference on the set of 800 attributes this is a statistically significant deviation from a random binary choice (50%), this is still significantly below human performance on similar tasks (Skowronek et al., 2006; Dalla Bella & Peretz, 2005).**



that 65% precision. While this is a statistically significant deviation from a random binary choice (50%), this is still significantly below human performance on similar tasks (Skowronek et al., 2006; Dalla Bella & Peretz, 2005).

Closer inspection further shows that not all taxons of a given category behave similarly. For instance, “Genre unplugged”, “Style Hard Rock” are strong acoustic correlates (in the sense defined above), probably because the instances found in the database are very prototypical, and timbrally consistent (e.g. salient saturated guitar and strong percussions in Hard Rock), while “Genre Jingle” and “Style Electronica” are poor acoustic correlates, possibly because they are very heterogeneous. “Electronica” for instance spans possibly everything from HardCore Techno—solely percussive, electronic pop (artists like Emilie Simon or Air) where voice is predominant—to Intelligent Techno—which uses concrete sound recordings and electronic blips—and even margin artists like Craig Armstrong (whose music could be also be described as “symphonic”).

Even attributes of the “Main Instrument” category are not particularly well modeled by acoustic similarity of “global sound”, which nevertheless seems a very related concept. This may be explained by the fact that instruments described by such attributes are usually not salient throughout the song, if salient at all. For instance, a given song by Elton John (a pianist) may be labeled as “piano” music, even though one can barely hear any piano sound on careful inspection, e.g. because it is very distant in a mix with predominant strings and synthetic pads, or because it is heavily processed with audio effects such as flangers and delays.

This shows that the typical sensory mappings of high-level musical descriptions to acoustic properties of the corresponding musical stimuli (as analysed with computer algorithms) are weak and ambiguous. Except for a few very stereotypical categories (like “heavy metal” or “aggressive mood”), most descriptions cannot be inferred from acoustic nearest neighbors with good precision.

This may mean one of two things:

- Either our measure of acoustic similarity fails to capture important aspects of human perception—indeed the

algorithm makes a number of critical simplifications, such as its representation of time ordering. It could be that, some day, a better algorithm be found that can capture some or all of the categories we consider here with significantly better precision.

- The other possible explanation for these results is cognitive: very few of the typical high-level symbolic descriptions that we manipulate in everyday music listening have a consensual, univocal definition in terms of a prototypical “sound”. When we categorize a piece of music as being “rock”, how much do we hear “rock”, and how much do we know or infer it?

Discriminating one of these two alternatives is surprisingly difficult. Disproving the second one would require finding a good enough mapping for each category in our set—a task that 10 years of musical pattern recognition research so far have had trouble with. Disproving the first hypothesis would require to look at all possible implementations of computational models of audition—which seems equally difficult. In fact, proving the second statement is logically untenable, as we would need to prove that our model of acoustic similarity is perfect in order to use the fact that it is not. We are facing a conceptual mismatch between the abilities of machines (which are auditory-only chimera) and that of humans, which possibly incorporate many more (semantic inference, culture, etc.).

While rigorous scientific arguments are difficult to put forward in favor of one or the other alternative, we feel that there is much to be gained by considering the second hypothesis. First, the engineering reality is such that no solution seems to be easily found to go beyond the “glass-ceiling” in machine performance. Investigating how humans and machine can build on imperfect acoustic representations (whether true by nature or by technical insufficiency) therefore can pave the way to novel technical solutions, regardless of their cognitive plausibility. Moreover, conceptual elaboration on what it means to ground musical categories on insufficient acoustical phenomenon may produce models and predictions supporting scientific investigation, further down the line. Hence:

Hence, **Working hypothesis:** Most of our everyday musical categories are not sensory constructs with well-defined physical correlates. There is no such thing that sounds like “rock” (nor “teen spirit”<sup>7</sup>).

The hypothesis of weak sensory mapping does not imply of course that music categorization does not need any auditory input, and can be derived arbitrarily of the corresponding physical stimuli, nor does it imply that acoustic stimuli of different categories cannot be distinguished. Recent research shows that even lower vertebrates are capable of discriminating musical styles (Chase, 2001), which indicates there exist low-level perceptual features that can be indexical of the type of musical categories we investigate here. Moreover, our results do not mean that the musical descriptions studied here are not amenable to traditional psychophysical investigation (using acoustic properties not considered here). However, our experiments are concerned with categorization rather than discrimination. Even if people or animals can make fine and consistent (dis)similarity judgments of musical stimuli (Dalla Bella & Peretz, 2005), it is unclear whether our use of high-level musical symbols can be predicted, from a strict auditory processing point of view, from the corresponding audio signals. In a related study described in (Janata, 2007), subjects were asked to rate the similarity between pairs of 60 sounds and 60 lexical descriptions thereof. The study concludes that there is no immediately obvious correspondence between single acoustic attributes and single semantic dimensions, and go as far as suggesting that the sound/word similarity judgment is a forced comparison (“to what extent would a sound spontaneously evoke the concepts that it is judged to be similar to?”). In that respect, the type of musical symbols considered here appear as motivated rather than predicted constructs, in the sense of (Lakoff, 2007).

#### **4. Grounding through Inter-symbolic Associations**

Then, how does this work at all? Most likely, the mechanisms for grounding high-level musical symbols on entities in the world

(acoustic signals) are no different than those involved in general language. Symbolic references, as analyzed in e.g. (Deacon, 1997), not only involve indexical relations between a symbol and an object, but also between a symbol and other symbols. Such inter-symbolic associations have the power to compensate a lack of associative support between a symbolic token and an object in the world (which may result from e.g. difficult sensory processing from the latter to the former or low co-occurrence of the two tokens in the same context), by recruiting a large number of other associations through token-token relationships. Such symbolic-level processing has been argued to be the basis of the evolutionary advantage of language (Cangelosi & Harnad, 2000).

Such inter-symbolic associations are easily observed in the dataset used in this study. Table 2 shows a selection of pairs of musical symbol tokens which were found to particularly fail a Pearson's  $\chi^2$ -test (Freedman et al., 1997) of statistical independence.  $\chi^2$  tests the hypothesis that the relative frequencies of occurrence of observed events follow a flat random distribution (e.g. that hard rock songs are not significantly more likely to talk about violence than non hard-rock songs).

We observe in Table 2 that a number of such correlations translate trivial word-to-word associations between attributes, such as "TextCategory Christmas" and "Situation Christmas", as well as logical links of mutual exclusion: a single song can't at the same time have varying and steady dynamics, or be both vocal and instrumental (i.e. non-vocal).

Table 2 further shows a number of dictionary-like associations, which have little to do with the actual musical usage of the words. For instance, the analysis reveals common-sense relations such as "Christmas" and "Special occasions", "Well-known" and "Popular", "Strong" and "Powerful". This shows that the process of categorizing music is consistent with psycholinguistics evidences of semantic associations, and that the specific usage of words that describe music is largely consistent with their generic usage: it is difficult to think of music that is both "strong" and not "powerful".

Finally, we also observe associations which are not intrinsic properties of the words used to describe music, but which are

**Table 2. Selected pairs of musical metadata with their  $\Phi$  score ( $\Phi^2$  normalized to the size of the population), between 0 (corresponding to statistical independence between the variables) and 1 (complete deterministic association).**

Attribute 1	Attribute 2	$\Phi$
Tautologic associations		
Language Finnish	Country Finland	0.93
Textcategory Christmas	Situation Christmas	0.81
Dynamics dynamic (up+down)	Dynamics steady	0.80
Mood aggressive	Variant aggressive	0.70
Main instruments male	Main instruments female	0.70
Dictionary associations		
Textcategory Christmas	Genre Special Occasions	0.89
Mood strong	Character powerful	0.68
Mood harmonious	Character well-balanced	0.60
Character robotic	Mood technical	0.55
Mood negative	Character mean	0.51
Encyclopedic associations		
Main instruments spoken vocals	Style Rap	0.75
Style Reggae	Country Jamaica	0.62
Musical Setup Rock Band	Main Instruments Guitar (distortion)	0.54
Character Mean	Style Metal	0.53
Musical Setup Big Band	Aera/Epoch 1940–1950	0.52
Main Instruments transverse flute	Character Warm	0.51

extrinsic properties of the music domain being described (in an encyclopedia way), e.g.:

- between musical genres: “Rap” and “Hip hop”
- between genres and countries: “Bossa Nova” and “Brazil”
- between genres and instruments: “Hip Hop” and “Spoken vocals”
- between genres and period: “Rag time” and “1930’s”

- between setups and instruments: “Rock band” and “Electric guitar”
- between genres and mood/character: “Jazz” and “Warm”, “Metal” and “Mean”
- between instruments and countries: “Tabla” and “India”
- between instruments and mood: “Transverse flute” and “Warm”

Some of these relations capture historical (“ragtime is music from the 1930’s”) or cultural knowledge (“rock uses guitars”), but also more subjective aspects linked to perception of timbre (“flute sounds warm”, “saturated guitar sounds aggressive”).

The existence of such correlations suggests that the perception of a given piece of music creates links to other music pieces one already knows, with potential to associate/reactivate common symbols in a metaphoric manner. It is known since (Koelsch et al., 2004) that music perception can activate brain mechanisms related to semantic processing, and can prime the meaning of a word in a similar way language can. Our observation that typical symbolic descriptions of music have only weak coupling to physical attributes of the musical signal opens the possibility for even deeper paradoxes, where auditory processing is only secondary to inter-symbolic inference. Just as language priming effects can backtrack to sensory indexical relations involving prime words (such as increase in galvanic skin response upon hearing words semantically related to words previously presented along with a mild electric shock) (Velmans, 1991), semantic relations may create auditory illusions in which people may “hear” things which are not statistically present in the physical stimuli (e.g. “piano” which is not picked up by models of acoustic similarity). A given song by Elton John may not offer strong auditory support for being categorized as “piano music”, but the knowledge that Elton John is a pianist, or that this particular song bears some similarity with another piano song, enables some kind of “automatic completion” of what is perceived onto what is thought to be perceived. Such paradoxes seem likely to occur mainly in the context of complex polyphonic music, which appears to require more complex auditory processing than other audio signals such as natural

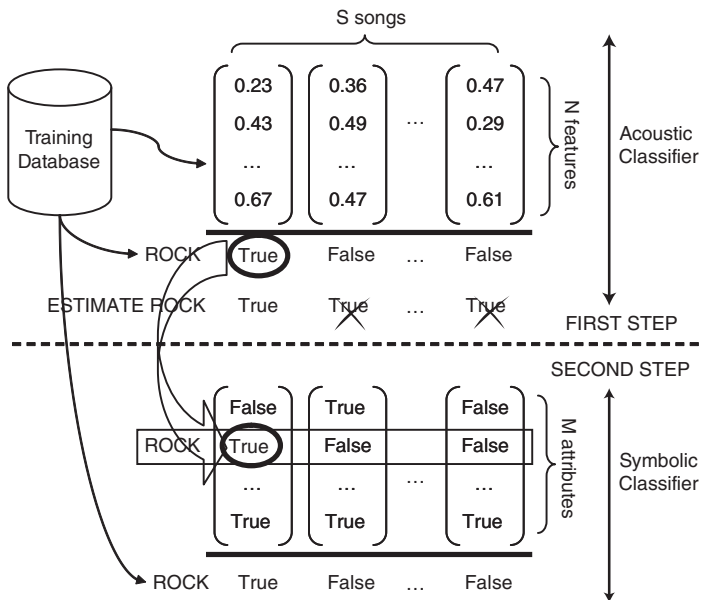
soundscapes (Aucouturier et al., 2007a), while at the same time generates a wealth of semantic affordances by e.g. mapping to the space of music pieces a particular listener already knows (e.g. “Elton John” finding room in my musical universe among other British pop singers, but also other pop pianists like Paolo Conte, etc.).

## 5. The Bootstrap Model

In (Aucouturier et al., 2007b), we proposed an operational (not cognitive) model that implements such a mechanism in a systematic way. The system uses nearest-neighbor inference with acoustic similarity (as in Section 2) as a bootstrap for correlation analysis.

The model has the two-step architecture presented in Figure 2. In a first step, we use acoustic inference to estimate the values of

**Figure 2. Architecture of the bootstrap model: acoustically-based classifiers (such as nearest-neighbor inference with acoustic similarity) are used as a bootstrap for symbolically-based classifiers exploiting inter-symbolic correlations.**



the musical attributes, given a song. Audio signals from a training database are encoded with audio features (namely MFCCs), and a classifier based on these features (namely, nearest neighbor with acoustic similarity) is built for each of the  $M$  musical terms (or attributes) in the corpus, using examples from the training database. Results from this first step only are poor, as demonstrated in Section 3: only a few acoustically correlated descriptions can be predicted with good precision.

In a second step, the same songs are described using features which are now taken to be the output of the classifiers from step 1, namely Boolean values for each of the  $M$  attributes. We train a second type of classifier (decision trees, see e.g. Quinlan, 1993) on these attribute values, to predict the values of the same attributes, i.e. we learn the statistical regularities in the training database showing e.g. that all music from Brazil (“Country Brazil = true”) with guitar (“Instrument Guitar = true”) but no drums (“Instrument drums = false”) is likely to be Bossa-Nova (“Genre = Bossa Nova”). From a pattern recognition point-of-view, step-2 classifiers are trained using the true values of the attributes, as found in the training database. This can be seen as prior knowledge, or musical common sense learnt from previous exposure to music. However, when processing a new song (i.e. in “testing”) mode), decision trees can only be fed with the estimated values for these same attributes, i.e. the output of step-1 audio classifiers, which may be corrupted. We know “Bossa-Nova” uses “Guitar” (knowledge embodied in step-2 predictor for Bossa-Nova), but we may be wrong on predicting whether a given song has “Guitar” (i.e. the step-1 predictor for Guitar may be wrong). The possible discrepancy between “true context” learnt from symbolic data and possibly wrong entry points into this context (estimated from signal) has important consequences for the predictions of the model (Section 6).

The model described in (Aucouturier et al., 2007b) is an iterative generalization of the above architecture: the outputs of step-2 classifiers are in turn used as symbolic features for a third step of classification, and so forth. This is conceptually similar, and was shown to further improve the precision of the resulting estimates. Table 3 shows the test performance of the above algorithm on a set

of 45 randomly chosen attributes. 30 out of the 45 attributes see their classification precision improved by the multi-step process (the remaining 15 do not appear in the table). We observe that, for 10 classifiers, the precision improves by more than 10% (absolute), and that 15 classifiers have a final precision greater than 70%. Cultural attributes such as “Situation Sailing” or “Situation Love” can be estimated with reasonable precision, whereas their initial acoustic-only estimate was poor. It also appears that two “Main Instrument” attributes (guitar and choir), that were surprisingly bad acoustic correlates, have been refined using inter-symbolic correlations. Many of the resulting classifiers reach performances that are comparable to human performance on similar tasks (Skowronek et al., 2006; Dalla Bella & Peretz, 2005).

**Table 3. Set optimization of 45 attribute estimates.**

Attribute	$p(\widetilde{A}_k^0)$	$p(\widetilde{A}_k^1)$	$i_t$	$\Delta(p)$
Situation Sailing	0.48	0.71	10	0.23
Situation Flying	0.49	0.64	3	0.15
Situation Rain	0.50	0.64	9	0.14
Instrument Guitar	0.60	0.69	4	0.09
Situation Sex	0.59	0.68	11	0.09
Situation Love	0.63	0.70	3	0.07
Lyrics Love	0.61	0.67	11	0.06
Situation Party	0.60	0.66	6	0.06
Tempo medium	0.59	0.64	4	0.05
Character slick	0.65	0.69	11	0.04
Aera/Epoch 90s	0.71	0.75	13	0.04
Character harmony	0.62	0.66	6	0.04
Rhythmics rhythmic	0.64	0.68	4	0.04
Genre Dancemusic	0.65	0.68	12	0.03
Mood dreamy	0.64	0.67	2	0.03
Style Pop	0.71	0.74	6	0.03

*Continued on next page*

Table 3—Continued

Mood positive	0.58	0.61	6	0.03
Mood harmonious	0.62	0.65	4	0.03
Instrument Choir	0.60	0.63	13	0.03
Dynamics up+down	0.61	0.63	5	0.02
Lyrics Associations	0.57	0.59	10	0.02
Variant expressive	0.62	0.64	2	0.02
Setup Pop Band	0.72	0.74	7	0.02
Lyrics Poetry	0.57	0.59	10	0.02
Character friendly	0.65	0.67	6	0.02
Character repeating	0.63	0.64	9	0.01
Rhythmics groovy	0.63	0.64	4	0.01
Mood romantic	0.69	0.70	9	0.01
Lyrics Wisdom	0.58	0.59	4	0.01
Lyrics Romantics	0.65	0.66	14	0.01

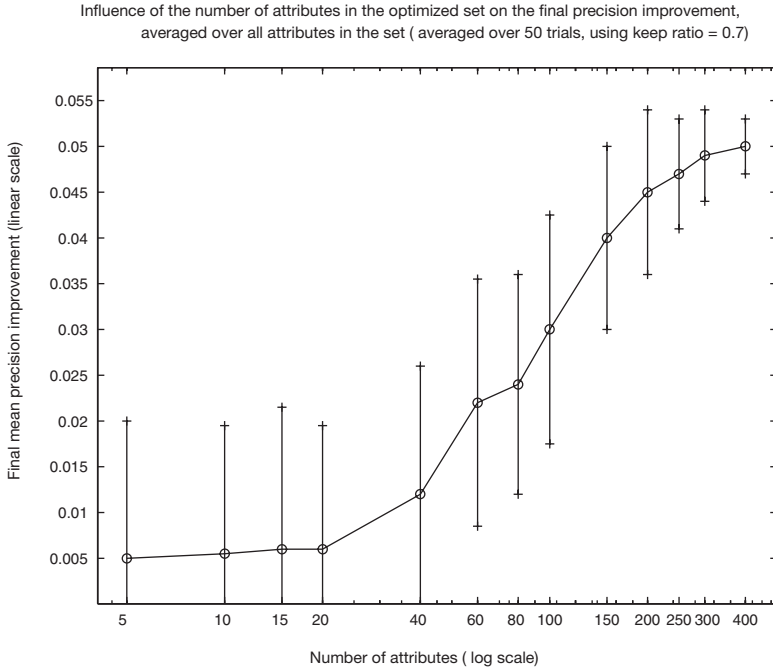
## 6. Predictions

The bootstrap model is mostly an engineering proposal, which can be validated in terms of added precision for pattern recognition. However, from this model, we are able to make a number of predictions on properties of a potentially underlying cognitive process.

### 6.1 Positive influence of symbolic complexity

The analysis of the bootstrap model shows that there is a critical mass effect in the number of high-level descriptions considered jointly. Figure 3 shows the mean improvement of precision for classifiers when reinforcing acoustic-inference by correlation analysis (compared to acoustic-only strategy). It appears that the more descriptions are considered, the better weak and ambiguous sensory mappings can be compensated. Symbolic complexity has a positive effect insofar as it provides denser networks of semantic associations.

**Figure 3. Influence of number of symbolic attributes on the mean precision improvement over timbre-only inference.**



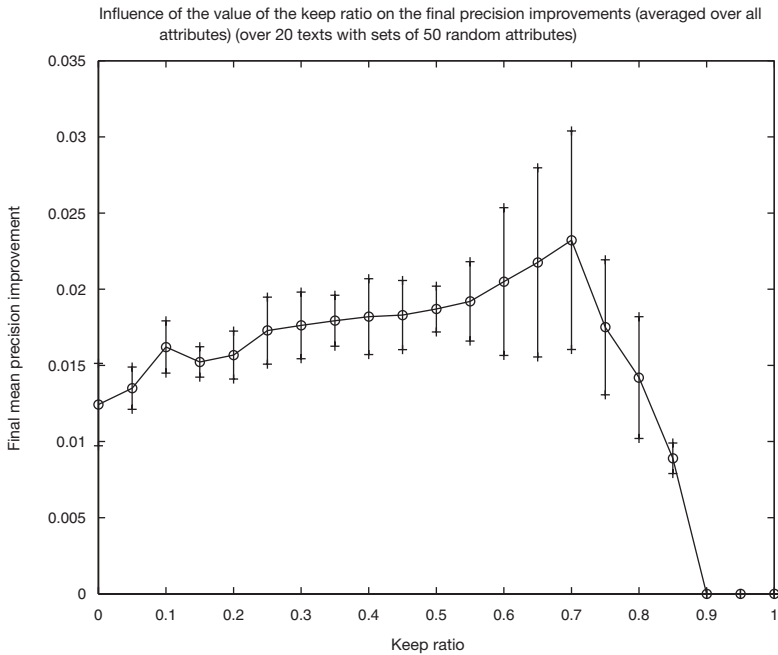
## 6.2 Sensibility on bootstrap precision

Figure 3 also shows that larger sets improve the stability of the results: the performance on small sets depends critically on the quality of the original acoustic estimates, which are used for bootstrap.

To further investigate the influence of the precision of the initial acoustically-based predictors, which are used as entry points into the network of correlations, we introduce a filtering parameter  $\theta$ . Classifiers' output are only included as features in subsequent steps if their precision is greater than  $\theta$ .

Parameter  $\theta$  is therefore a trade-off between quantity and quality of the correlations to be exploited in decision trees. Figure 4 shows the influence of this parameter. The curve has an inverted-U

**Figure 4. Influence of precision threshold on the average improvement of precision parameter. The curve has an inverted-U shape: small values lead to selecting too many bad classifiers, whereas large values constrain the system to use only high-quality features, which are ultimately too few to bootstrap correlation analysis. The optimal value is found around 70% precision, which is consistent with the empirical upper-bound found with signal-only approaches (so-called “glass ceiling”) (Aucouturier & Pachet, 2004). This predicts that contextual inference should be based on rules that mostly involve constructs that can be well-mapped to acoustic parameters of the signals.**



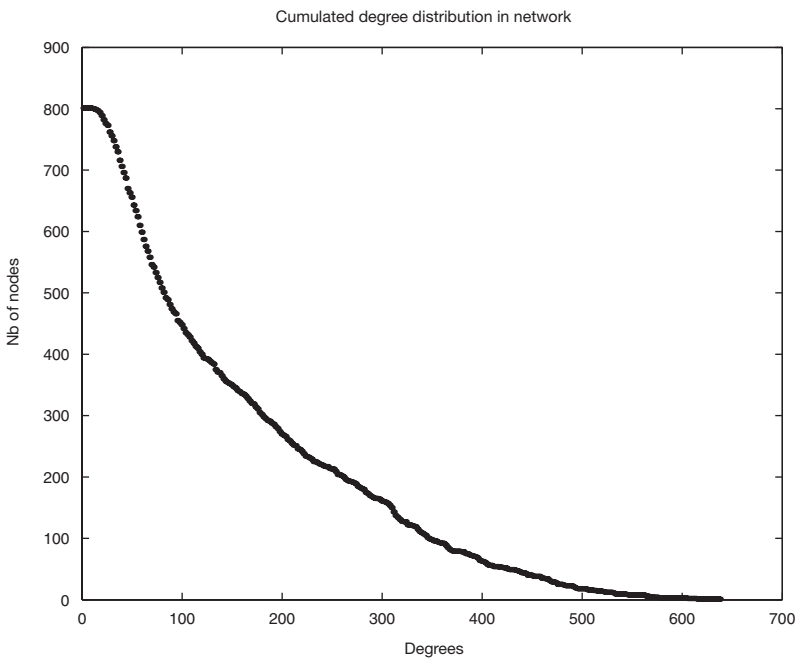
shape: small values lead to selecting too many bad classifiers, whereas large values constrain the system to use only high-quality features, which are ultimately too few for bootstrap correlation analysis. The optimal value is found around 70% precision, which is consistent with the empirical upper-bound found with signal-only approaches (so-called “glass ceiling”) (Aucouturier & Pachet, 2004). This predicts that contextual inference should be based on rules that mostly involve constructs that can be well-mapped to acoustic parameters of the signals.

### 6.3 Some entry-points are better than others

Not all symbols in the complex network of correlations illustrated in Section 4 are equally connected. If we take each of the 801 attributes in our musical corpus as a node, and establish a link between two nodes if they are statistically correlated with Pearson's  $r > 50$  (a total of 33,884 links), we find that the resulting network has an exponential degree distribution<sup>8</sup> (Figure 5). There are nodes with more symbolic weight than others: very many poorly connected symbols and a few very connected ones. Exponential degree distributions are not uncommon in networks representing musical concepts. Cano et al. (2006) observed that networks of artists linked by recommendations of online music websites (like Yahoo or AMG) are also exponential.<sup>9</sup>

The degree of a musical description is difficult to predict: it results of a compromise between discriminative power and popularity. If too popular (e.g. “music”), a symbol has no significant

**Figure 5. Degree Distribution**

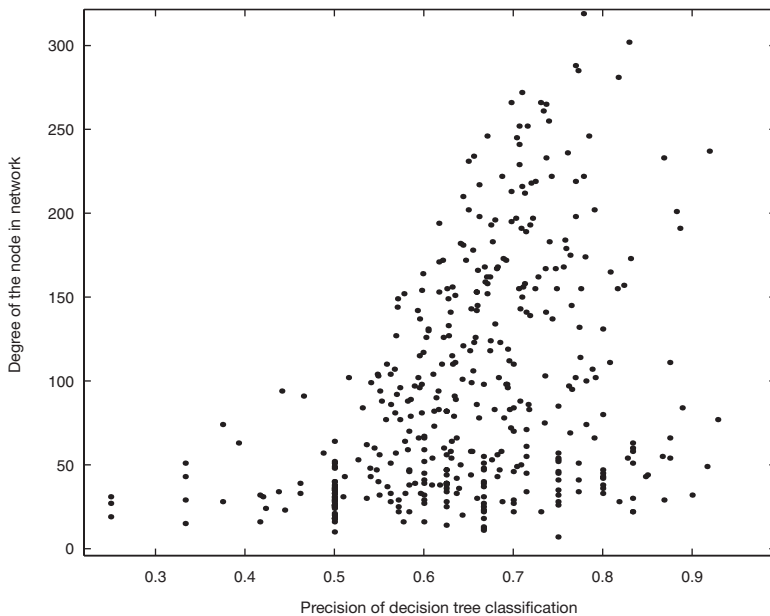


correlation to other symbols. If too discriminative (e.g. “ukulele”), then the corpus doesn’t include enough songs to observe meaningful correlations.

Figure 6 shows the relationship between the network degree of the attributes used for bootstrap and the resulting mean precision of the classifiers obtained by the 2-step procedure of the bootstrap model. We observe an inclusion relationship: if bootstrap attributes (i.e. those that happen to be well-grounded in purely physical parameters of the audio signals) are also high-degree nodes in the network, then high precision is achieved for other attributes via the correlation network. This suggests that a high positive correlation between the perceptual grounding (i.e. precision of acoustic-only estimation) and the degree of the symbol in the semantic network is a desirable feature to learn a network of perceptual categories. A network of symbols should be easier to learn if the hubs of the semantic network coincide with well-defined sensory constructs.

**Figure 6. Degree Distribution**

Joint plot of precision of decision tree influence against degree of the node in the network



## 7. Conclusion

From a common-sense perspective, the very varied lexical descriptions we routinely use to communicate about music are well-defined sensory constructs, strongly coupled to the acoustic properties of the corresponding musical stimuli. Surely there is music which “sounds like” rock, jazz or techno; “something” in certain sounds which feels “smooth”, “metallic” or “stressful”. However, attempts to model such physical mappings with computer pattern recognition have (arguably) failed, so far. This could mean one of two things. Either that pattern recognition is an underachieving paradigm, or that the words we use to talk about music were less heavily based on acoustic similarity that we usually think. In this paper, we elaborated on the second hypothesis.

We showed that even if their mapping to the physical parameters of audio stimuli is weak, typical audio categories can be predicted with reasonable precision when considering associations at the symbolic level. In this respect, the subset of language we use to describe music is not different than language as a whole. Inter-symbolic associations have the power to compensate a lack of associative support between symbolic token and object in the world, by recruiting a large number of other associations through token-token relationships. We have proposed here a engineering model (the so-called “bootstrap model”) that shows how to re-integrate pattern recognition techniques in this larger vision: signal-based classifiers are used as entry points into a learned network of semantic relations. With this model, we are able to predict properties of the symbolic system, such as the positive influence of symbolic complexity, and the importance of bootstrapping with highly connected nodes of the semantic network.

Such predictions remain to be tested through psychological experiments. While it is difficult to arbitrarily manipulate symbolic systems in human subjects, some recent experiments have illustrated ways to let humans ground artificial representations, e.g. with sophisticated interactional tasks (Healey et al., 2007) or by studying schizophrenic patients (Kim et al., 2005).

On the whole, in the study of such semantic priming effects on auditory perception, the use of computational techniques that simulate human auditory processing (such as the acoustic similarity algorithm used here) appears to be a promising research methodology: computer algorithms are auditory-only chimeras, deprived of the higher-level semantic capabilities found in people. The comparison of their performance with humans on the same tasks could help clarifying the proportion of auditory vs symbolic processing in music perception. If we were “auditory-only”, how could we ground the words we use to talk about music ? How could we learn such a vast symbolic system ? How can language evolve from purely indexical relations, and come to exist in parallel to the sensory world ?

## **Acknowledgments**

Much of the discussion found here is based on pattern recognition experiments originally co-authored with François Pachet, Pierre Roy and Anthony Beurivé at Sony Computer Science Labs, Paris (Aucouturier et al., 2007b). Access to the musical corpus used for experiments was granted in the framework of research partnerships within the Sony corporation, under the initiative and leadership of François Pachet. The author wishes to thank Don Byrd, Petter Johansson, Josh Tenenbaum and Ryoko Uno for their detailed comments and suggestions on the cognitive significance of these results.

## Appendix A: Description of the computer measure of acoustic similarity

We give here a technical description of the algorithm summarized in Section 2. For a longer account, please refer to (Aucouturier et al., 2005). The audio signal is first cut into frames. For each frame, we estimate the spectral envelope by computing a set of mel-frequency cepstrum coefficients (MFCCs). The cepstrum is the inverse Fourier transform of the logarithm of the Fourier spectrum  $\log S$ .

$$c_n = \frac{1}{2\pi} \int_{\omega=-\pi}^{\omega=\pi} \log S(\omega) \exp j\omega n d\omega \quad (\text{A1})$$

We call mel-cepstrum the cepstrum computed after a non-linear frequency warping onto a perceptual frequency scale, the Mel-frequency scale (Rabiner & Juang, 1993), which reproduces the non-linearity of the frequency resolution of the human auditory system (low Hertz frequencies are more easily discriminated than high Hertz frequencies). The  $c_n$  in Equation A1 are called Mel frequency cepstrum coefficients (MFCCs), of which we keep a given number  $N$ .

We then model the distribution of the MFCCs over all frames using a Gaussian mixture model (GMM). A GMM estimates a probability density as the weighted sum of  $M$  simpler Gaussian densities, called components or states of the mixture (Bishop, 1995):

$$p(x_t) = \sum_{m=1}^{m=M} \pi_m \mathcal{N}(x_t, \mu_m, \Sigma_m) \quad (\text{A2})$$

where  $x_t$  is the feature vector observed at time  $t$ ,  $\mathcal{N}$  is a Gaussian probability density function with mean  $\mu_m$ , covariance matrix  $\Sigma_m$ , and  $\pi_m$  is a mixture coefficient (also called state prior probability). The parameters of the GMM are learned with the classic E-M algorithm (Bishop, 1995).

We then compare the GMM models to match the timbre of different songs, which gives a similarity measure based on the audio content of the music. We use a Monte-Carlo approximation of the Kullback-Leibler (KL) distance between each duple of models A and

B. The KL-distance between 2 GMM probability distributions  $p_A$  and  $p_B$  (as defined in A2) is defined by:

$$d(A, B) = \int p_A(x) \log \frac{p_B(x)}{p_A(x)} dx \quad (A3)$$

The KL distance can thus be approximated by the empirical mean:

$$d(\widetilde{A}, B) = \frac{1}{n} \sum_{i=1}^n \log \frac{p_B(x_i)}{p_A(x_i)} \quad (A4)$$

(where  $n$  is the number of samples  $x_i$  drawn according to  $p_A$ ) by virtue of the central limit theorem:

$$\lim_{n \rightarrow \infty} \left( \frac{1}{n} \sum_{i=1}^n X_i - \mathcal{E}(X) \right) = \frac{1}{\sqrt{n}} \mathcal{N}(0, \sigma^2) \quad (A5)$$

where  $X$  is the random variable  $\log(p_A(x)/p_B(x))$ ,  $X_i$  a realization of  $X$ ,  $\mathcal{E}(X)$  the mean of  $X$  and  $\mathcal{N}(0, \sigma^2)$  a normal distribution of mean 0 and variance  $\sigma^2$ , the variance of  $X$ .

## End Notes

1. Note that there are nevertheless important and well-documented subjective effects in the perception of such dimensions. For instance, it is known since (Guernsey, 1928) that the sensation of consonance for pitch intervals depends on the listener's musical expertise.
2. There are exceptions, which have relied on very specialized experimental setups, such as (Dubnov et al., 2006) who asked listeners to rate their sensation of "Emotional Force" using a specially designed apparatus, while hearing two specially commissioned versions of a musical piece, in a live-concert setting.
3. In this work, we describe as Computer Pattern Recognition the sub-domain of Artificial Intelligence which tries to simulate human perceptive processes

(classification, similarity, etc.) for natural objects (text, image, sound) with computational techniques, without necessary concern for their biological plausibility.

4. International Conference on Music Information Retrieval (ISMIR, 2007).
5. <http://www.pandora.com/>
6. Contrary to the case of “4-letter words”, maybe...
7. “Smells Like Teen Spirit” is a song by the American rock band Nirvana, written by Kurt Cobain, Krist Novoselic, and Dave Grohl, on 1991 album *Nevermind*.
8. The degree of a node tells how many links the node has to other nodes.
9. These can also be scale-free, though.

## References

- ISMIR (2007). *Proceedings of the Eighth International Conference on Music Information Retrieval*. ISMIR 2007.
- Aucouturier, J.-J., Defreville, B., & Pachet, F. (2007a). The bag-of-frame approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *Journal of the Acoustical Society of America* (accepted).
- Aucouturier, J.-J. & Pachet, F. (2004). Improving timbre similarity: How high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1).
- Aucouturier, J.-J. & Pachet, F. (2006). The influence of polyphony on the dynamical modelling of musical timbre. *Pattern Recognition Letters* (in press).
- Aucouturier, J.-J., Pachet, F., Roy, P., & Beurivé, A. (2007b). Signal + context = better classification. *Proceedings of the International Conference on Music Information Retrieval*. Vienna (Austria).
- Aucouturier, J.-J., Pachet, F., & Sandler, M. (2005). The way it sounds: Timbre models for analysis and retrieval of polyphonic music signals. *IEEE Transactions on Multimedia*, 7(6), 1028–1035.
- Bainbridge, D., Cunningham, S. J., & Downie, J. S. (2002). How people describe their music information needs: A grounded theory analysis of music queries. *Proceedings, 3rd International Conference on Music Information Retrieval*. Paris, France.
- Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford Press.

- Cangelosi, A. & Harnad, S. (2000). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, 4(1), 117–142.
- Cano, P., Celma, O., Koppenberger, M., & Buldu, J. (2006). Topology of music recommendation networks. *Chaos*, 16(1).
- Chase, A. R. (2001). Music discrimination by carp (*Cyprinus carpio*). *Animal Learning & Behavior*, 29(4), 336–353.
- Dalla Bella, S. & Peretz, I. (2005). Fine differentiation and ordering of classical music requires little learning but rhythm. *Cognition*, 96.
- Deacon, T. (1997). *The symbolic species: The coevolution of language and human brain*. New-York: W. W. Norton.
- Dhanaraj, R. & Logan, B. (2005). Automatic prediction of hit songs. *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, UK.
- Dubnov, S., McAdams, S., & Reynolds, R. (2006). Structural and affective aspects of music from statistical audio signal analysis. *Journal of the American Society for Information Science and Technology*, 57(11), 1526–1536.
- Freedman, D., Pisani, R., & Purves, R. (1997). *Statistics* (3rd ed.). W. W. Norton & Co., New York.
- Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61, 1270–1277.
- Guernsey, M. (1928). The role of consonance and dissonance in music. *American Journal of Psychology*, 40, 173–204.
- Healey, P. G. T., Swoboda, N., Umata, I., & King, J. (2007). Graphical language games: Interactional constraints on representational form. *Cognitive Science*, 31, 285–309.
- Janata, P. (2007). Timbre and semantics. Keynote Presentation, *Journées Fondatrices Perception Sonore*. Lyon (France), January 2007. Available at <http://www.sfa.asso.fr/fr/gps>.
- Kim, J.-J., Ho Seok, J., Park, H.-J., Soo Lee, D., Chul Lee, M., & Soo Kwon, J. (2005). Functional disconnection of the semantic networks in schizophrenia. *Neuroreport*, 16(4), 355–359.
- Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., & Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience*, 7, 302–307.
- Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago, IL: University of Chicago Press.
- Lidy, T., & Rauber, A. (2005). Evaluation of feature extractors and psychoacoustic transformations for music genre classification. *Proceedings*

- of the 6th International Conference on Music Information Retrieval.* London, UK.
- Liu, D., Lu, L., & Zhang, H.-J. (2003). Automatic mood detection from acoustic music data. *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*. Baltimore, Maryland, USA.
- McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58, 177–192.
- Pachet, F. & Cazaly, D. (2000). A taxonomy of musical genres. *Proceedings Recherche d'Information Assisté par Ordinateur (RIAO)*. Paris (France).
- Quinlan, J. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufman.
- Rabiner, L. & Juang, B. (1993). *Fundamentals of speech recognition*. Prentice-Hall.
- Scaringella, N. & Zoia, G. (2005). On the modelling of time information for automatic genre recognition systems in audio signals. *Proc. International Symposium on Music Information Retrieval*. London (UK).
- Skowronek, J., Van de Par, S., & McKinney, M. (2006). Groundtruth for automatic music mood classification. *Proceedings of the 7th International Conference on Music Information Retrieval*. Victoria BC (Canada).
- Tzanetakis, G., Essl, G., & Cook, P. (2001). Automatic musical genre classification of audio signals. *Proceedings ISMIR*.
- Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences*, 14, 651–726.
- Vincent, E. & Rodet, X. (2003). Instrument identification in solo and ensemble music using independent subspace analysis. *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*. Barcelona, Spain.