

How Much Audition Involved In Everyday Categorization of Music?

Jean-Julien Aucouturier^{a,*},

^a*Ikegami Lab, Grad. School of Arts and Sciences, The University of Tokyo, Japan.*

Francois Pachet^b s

^b*SONY Computer Science Laboratory Paris, France.*

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 a possible infinity of similar metaphors. Such descriptions are generally considered as well-defined sensory constructs, strongly coupled to the acoustic properties of the corresponding musical stimuli, in a way that can be studied with the tools of psychophysical investigation. This paper describes a computational study of the grounding of a large and heterogeneous set of such high-level musical symbols, for which we seek a mapping to intrinsic sound properties of the corresponding musical stimuli. The study is computational insofar as we don’t rely on direct psychophysical investigation, but rather simulate human perception algorithmically with computer pattern recognition. We show that, surprisingly, the typical sensory mappings of high-level musical descriptions to acoustic properties of the corresponding musical stimuli are extremely weak and ambiguous. It is therefore likely that typical categorical processes of everyday polyphonic music require more symbolic processing than generally assumed: when we categorize a piece of music as being “piano”, how much do we *hear* piano, and how much do we *know* or *infer* it?

Key words: Symbol Grounding, Music, Semantics, Acoustic similarity, Pattern Recognition, Collaborative Tagging

* Corresponding author. Address: Ikegami Laboratory, Department of General Systems Studies, Graduate School of Arts and Sciences, The University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153-8902, Japan. Tel:+81-3-5454-4378. Fax : +81-3-5454-6541.

Email addresses: aucouturier@gmail.com (Jean-Julien Aucouturier),

1 “Sounds Like Teen Spirit”

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 probable infinity 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)¹, 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 and 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). Results 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”. This intuition is also reflected in scientific investiga-

pachet@csl.sony.fr (Francois Pachet).

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

tions in both experimental psychology and computer modelling.

1.1 Psychophysical studies

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, 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]), using a common methodology: dissimilarity judgments are collected from a set of human subjects (usually musicians) for a set of individual, natural instrumental sounds. The similarity ratings are then analysed with Multidimensional scaling ([Borg and Groenen, 1997]) in order to find a low-dimensional spatial arrangement of the stimuli in a euclidean space such that the distances between data points are optimally respected. The dimensions of this projection space are then given perceptual or acoustic interpretations. In the case of timbre, for instance, the 2 most important dimensions have usually been consensually correlated to the centroid of the spectral envelope (which measures the spectral energy distribution in the steady state portion of a tone, and corresponds to perceived “brightness”), and the log of the attack-time, i.e. the time between the onset and the instant of maximal amplitude.

Traditionally few psychophysical studies have targeted the high-level, symbolic constructs we’re concerned with here, due to two notable difficulties:

- High-level descriptions “à la genre” are typically developed over the time scale of a few seconds, e.g. in musical phrases/textures. In this context, psycho-physic mappings involve holistic aggregates of physical attributes, which are difficult to extract computationally from audio signals and manipulate in synthetic stimuli.
- As noted earlier, such descriptions are not necessarily consensual among subjects, hence difficult to investigate in the framework of experimental psychology.

To circumvent these difficulties, some recent works have focused on “niche” problems where consensus is indeed found, only in a limited community of expert listeners (e.g. [Berger and Fales, 2003] on the acoustical correlates of the “heaviness” of electric guitar sounds in *Heavy Metal* music), or have relied on very involved and specialized experimental set-ups, 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.

1.2 Computer Pattern Recognition

Following the same intuition, research in computer pattern recognition² seeks computational methods able to simulate 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 consists 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 and 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 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 description (as manifest e.g. in the past ISMIR conferences³). This 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 and Rodet, 2003]), singing language ([Tsai and Wang, 2003]), potential for commercial success ([Dhanaraj and Logan, 2005]), etc.

1.3 A computational study

This paper propose to re-evaluate the common-sense take on our everyday musical symbols that assumes a necessarily strong sensory coupling between mental representations and acoustical properties of the musical entities to which they refer. We describe here a computational study of the grounding of a large and heterogeneous set of high-level musical symbols, for which we seek a mapping to intrinsic sound properties of the corresponding auditory stimuli. We investigate whether there are direct acoustical correlates to part or totality of such symbols, and what are the typical properties of the mappings involved.

² 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.

³ International Conference on Music Information Retrieval, [ISM, 2005]

This study is computational insofar as we don't rely on direct psychophysical investigation, but rather simulate human perception algorithmically. More precisely, we make use of the two following opportunities:

- Computer models of timbre similarity: Recent tools developed in the domain of Computer Pattern Recognition are now able to computationally simulate human judgments of the acoustic similarity of 2 polyphonic sound textures⁴, with reasonable accuracy [Aucouturier et al., 2005]. These algorithms typically model the audio signal as the long-term statistical distribution of their local spectral features, a prototypical implementation of which being Gaussian Mixture Models of Mel-Frequency Cepstrum Coefficients. This paper uses an implementation of this technique to evaluate the perceptual similarity of the “global sound” of pieces of music, and analyze its agreement with the corresponding high-level descriptions. If a high correlation is found between a given high-level description and the acoustic similarity of the corresponding songs, then we can reasonably conclude that the description symbol has a consensual, univocal grounding in a stereotypical “sound” of music.
- Collaborative music tagging: Collaborative Tagging describes the process by which many users organize shared data by freely assigning keywords (or tags) to items, and sharing these tags with other users. Convergence to a shared vocabulary of tags can be often observed [Golder and Huberman, 2006], as a result of self-organization, although the categories are not predefined. In the context of music, collaborative tagging of a shared database of music pieces is therefore an efficient way to harvest vast amounts of high-level human judgments on very large quantities of musical data, and yet maintain a high degree of consensus. Possible sources for such data are public websites such as Last.fm⁵ and business initiatives relying on the same mechanisms (such as Pandora⁶ or Moodlogic⁷). This paper uses proprietary Collaborative Tagging data of the latter kind, made available through research contracts in Sony Computer Science Laboratory.

Our approach provides an alternative to traditional psychophysical investigations, by focusing on high-level attributes with a built-in consensus which emerged through a human collaborative tagging process (see Section 2.1.1). It is also unique with respect to previous studies by its scope (more than 800 attributes for each of 5,000 songs), which is a consequence both of Collaborative Tagging (which is typically a large and distributed process) and algorithmic simulation of acoustic similarity (which can process amounts of

⁴ what one usually denotes as “this *sounds like* ... (Beethoven's 9th Symphony, The Beatles, etc.)”

⁵ <http://www.last.fm/>

⁶ <http://www.pandora.com/>

⁷ <http://www.moodlogic.com/>

data well beyond the scope of psychological experimentation).

There are several limitations to our approach, notably:

- Precision of computational approximations: Although widely accepted in the research communities in which they originated, there are open questions about the precision and suitability of the two computational techniques used in this paper. The Collaborative Tagging data on the one hand should be validated in terms of consistency, compared to more traditional manually-controlled psychological tests. On the other hand, the precision of the computer simulation of human judgments of acoustic similarity can also be an issue. We comment on these 2 aspects in Section 3.2 and 2.2.1 respectively.
- Grounding to timbre only: The measure of acoustic similarity that we use here quantifies differences in the holistic “sound” of musical textures, which could be viewed as a multi-note, polyphonic extension of musical “timbre” ([Aucouturier et al., 2005]). The measure does not model explicitly other aspects of musical signals, such as pitch, harmony or rhythm, which nevertheless have clear relevance to a number of high-level musical descriptions. For instance, the harmonic mode (major, minor, etc.) and tempo of a melody were found to influence the judgment of its mood among “happy” or “sad” [Gagnon and Peretz, 2003]. Notably, this means that musical descriptions which we find bearing little correlation to intrinsic sound properties in the following could be found to have clearer coupling to unexamined aspects such as harmony or tempo. We do not expect such effects to have potential to dramatically change our present conclusions though: humans have been reported able to issue categorical judgments of musical genres and styles with good precision using as little as 200ms of audio [Perrott and Gjerdingen, 1999]. Such time scales indicate that sufficient information can be found in short-term, surface features of the audio signals, compared to longer-term analytic constructs such as rhythm (which require periodicity analysis at the time scale of a musical phrase). Moreover, these latter aspects can be studied with the same methodology, especially as computational techniques exist to simulate human perception of e.g. harmony [Gomez, 2006] or rhythm [Gouyon, 2005].

2 Experiment

In this section, we report on our experiment to evaluate the typical strength of the grounding/mapping of a large set of high-level musical descriptions (such as genre or mood), on the acoustic properties of the corresponding musical stimuli.

2.1 Methods

2.1.1 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 authors by research partnerships. The database contains 4936 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	Nb 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 project⁸. Each song in the database was annotated by several persons, and results agglomerated by thresholding techniques⁹ (“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.

2.1.2 Computer Measure of Acoustic Similarity

In this study, we 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. See Appendix A for a technical description of the algorithm. The measure was found to approximate human judgments with reasonable precision (see e.g. [Aucouturier et al., 2007a] for a technical discussion of the algorithm’s limitations).

2.1.3 Evaluation of Mapping Strength

We study here the strength of the mapping between a given high-level description (or attribute) and the acoustic “sound” or “timbre” property of the corresponding signals. To do so, we propose to evaluate the precision of an inference mechanism based on the timbre similarity described above, in which a test song \mathcal{S} is given an attribute value $\mathcal{A}(\mathcal{S})$ (e.g. “mood violent =true”) if a sufficient amount of songs which sound similar to \mathcal{S} have the same value (e.g.

⁸ <http://www.pandora.com/>

⁹ the exact details of the agglomerative process were not made available to the authors

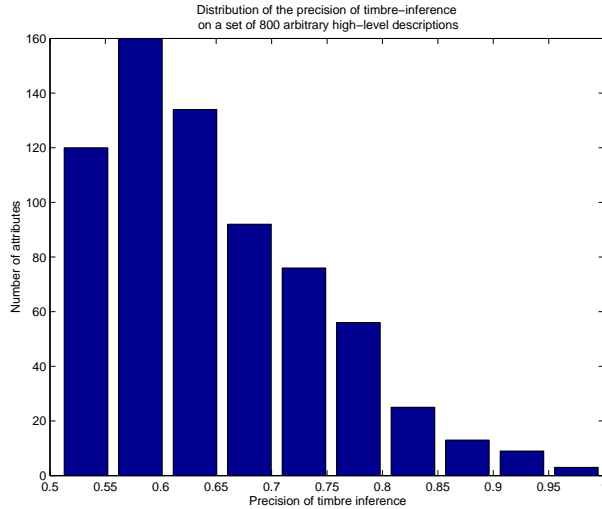


Fig. 1. Distribution of the precision of timbre inference on the set of 800 attributes.

“a song is violent if it sounds like a lot of violent songs”). A formalized description of the algorithm can be found in Appendix B. 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 \mathcal{A} and the “timbre” property of the associated musical signal as the *precision* of nearest neighbor classification when used to infer values of \mathcal{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 statement¹⁰.

2.2 Results

2.2.1 Validation of the similarity algorithm

Figure 1 shows the distribution of the precision of the nearest neighbor inference process described above on the set of the 801 boolean attributes in the database. It appears that some attributes are very correlated to timbre similarity (in the sense defined above), achieving precisions sometimes higher than 95%.

¹⁰ contrary to the case of “4-letter words”, maybe...

Table 2 shows the 10 attributes (out of 801) that were the most precisely inferred using timbre similarity. Most attributes correspond to “Styles” which described music with quite extreme observable features of the signal (e.g. saturated guitar, distorted vocals, high percussivity).

Table 2

Best inferred attributes. P_+ (resp. P_-) is the ratio of the number of correctly inferred *true* (resp. *false*) values over the total number of *true* (resp. *false*) values.

Attribute	P_+	P_-	Mean Precision
Style Techno (minimal)	0.93	0.99	0.96
Style Rave	0.94	0.98	0.96
Genre Lullaby/Nursery Rhyme	0.91	0.99	0.95
Style Hard Trance	0.92	0.95	0.93
Language Native American	0.88	0.98	0.93
Style Garage	0.87	0.99	0.93
Style Happy Hardcore	0.86	0.98	0.92
Style Metal	0.88	0.96	0.92
Style Hardcore	0.86	0.98	0.92
Style Grunge	0.87	0.95	0.91

The saliency of such results (with near perfect precision) can be proposed as an experimental validation of the suitability of the computational measure of similarity used in this work.

2.2.2 Surprisingly few attributes are correlated with timbre

While some attributes are very strong acoustic correlates, we observe from Figure 1 that there are surprisingly few of such near-perfect correspondences. 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 than 65% precision¹¹. While this is a statistically significant deviation from a random binary choice (50%), this is still significantly below human performance on similar tasks

¹¹ It should be mentioned that the nearest neighbor algorithm defined above is not a particularly flexible classification algorithm, and notably can have difficulties with strongly multimodal distributions (“Songs of category X sound either like A, or like B, or like C”). Therefore, these results are not quantitatively typical of the state-of-art pattern recognition techniques for music classification (as we reviewed in Section 1.2). However, recent results suggest better algorithms would not yield qualitatively different conclusions [Aucouturier and Pachet, 2004]

[Perrott and Gjerdingen, 1999, Dalla Bella and Peretz, 2005]. This indicates that very few typical high-level music descriptions have a consensual, univocal definition in terms of a prototypical “timbre” or “sound”.

2.2.3 Not all taxons of a given category behave similarly

Table 3 shows the distribution of the categories of the attributes which are inferred with more than 75% precision, while Table 4 shows the distribution of the attributes which are inferred with less than 55% precision.

Table 3

Categories of the best inferred attributes

Category	Nb attributes	Sample Attributes
Style	48	Jazz (Trad.), Hard Rock
Genre	10	Unplugged, Nightclub Music
Main Instrument	10	Guitar (distortion), Vocals (Spoken; Rap)
Aera/Epoch	9	1950-1960, 1960-1970
Musical Setup	7	Big Band, Rock band
Character	6	Metallic, Warm
Country	6	Cuba, Jamaica
Variant	3	Aggressive, Metallic
Situation	3	Computer Animation, Middle Ages
Mood	2	Aggressive, Negative
Language	1	Native American

From these two tables, we observe that many categories include both timbrally-related and unrelated attributes. “Genre unplugged”, “Style Hard Rock” (found in Table 3) are strong timbre 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” (found in Table 4) are poor timbre 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, 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”).

Table 3 and 4 also confirm that categories like “TextCategory”, “Situation” or “Mood” mostly capture cultural and subjective information which are

Table 4
Categories of the worst inferred attributes

Category	Nb attributes	Sample Attributes
TextCategory	35	Irony, Play on Words
Main Instrument	18	Clarinet, Body Percussion
Style	17	Underground, Electronica
Situation	10	Hollywood, Winter
Variant	6	Thin, Wrong / Amateurish
Mood	6	Maritime, Funny
Country	5	International, USA
Musical Setup	4	Duo, Girlgroup
Character	3	Child-oriented, Vibrating
Metric	2	6/8, 7/8
Dynamics	1	Decreasing
Genre	1	Jingle/Link
Language	1	African Languages
Tempo	1	Alternating

poorly described with acoustic similarity inference. Nevertheless, taxons like “TextCategory Explicit” or “Mood Violent” are very good timbre-correlates.

What appears from these results is that only a few taxons, wide-spread over many diverse categories, can reliably be inferred with acoustic similarity.

2.2.4 *Even instruments can't always be predicted*

Table 3 and 4 further shows that attributes of the “Main Instrument” category are not particularly well modeled by acoustic similarity of “global sound”, which yet 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* 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.

3 Discussion

3.1 *Motivated rather than predicted representations*

Our study shows that the typical sensory mappings of high-level musical descriptions to acoustic properties of the corresponding musical stimuli 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 means that 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 “timbre” or “sound”. This appears to apply even for descriptions generally expected to be intrinsic to the audio stimuli, such as instrumentations.

This contradicts common-sense intuition that our everyday musical categories are sensory constructs with well-defined physical correlates that can be studied with the tools of psychophysical investigation. The fact that such strong physical coupling is not picked up by state-of-art computer models of auditory processing (e.g. timbre similarity as used here) questions the respective proportion of auditory versus symbolic processing in human music perception. When we categorize a piece of music as being “piano”, how much do we *hear* piano, and how much do we *know* or *infer* it ?

These results do not imply of course that music categorization do 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 [Chai and B., 2001], which indicates there exist low-level perceptual features that can be indexical of the type of musical categories we investigate 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 and 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, 1987].

3.2 *Grounding through inter-symbolic associations*

What our results could indicate however is that 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 analysed 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 associate support between symbolic token and object in the world (which may result of 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 and Harnad, 2000].

Such inter-symbolic associations are easily observed in the dataset used in this study. Table 5 shows a selection of pairs of musical symbol tokens which were found to particularly fail a Pearson’s χ^2 -test ([Freedman et al., 1997]) of statistical independence. χ^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 5 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). Note that the consistency of such logical links is remarkable in the context of massive manual categorization. This contributes to validating data obtained by Collaborative Tagging, which therefore appears to be a promising method to generate artificial “niche” societies (such as the Heavy Metal fans discussed in [Berger and Fales, 2003]) where sufficient consensus can be found to provide a basis for psychological investigation.

Table 5 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”. 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”.

Table 5

Selected pairs of musical metadata with their Φ score (χ^2 normalized to the size of the population), between 0 (corresponding to statistical independence between the variables) and 1 (complete deterministic association)

Attribute1	Attribute2	Φ
Tautologic associations		
Language Finish	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

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

3.3 *Toward semantic priming effects on auditory perception*

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 lexical, semantic 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 when people may “hear” things which are not statistically present in the physical stimuli (e.g. “piano” which is not picked up by models of timbre similarity). Such paradoxes seem likely to occur mainly in the context of complex polyphonic music, which seems to require more complex auditory processing power 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.).

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 this study) as a

bootstrap for correlation analysis. First, we use timbre-inference to estimate the values of a few timbre-correlated attributes, and then exploit correlations at the symbolic-level (with a technique called decision trees) to make further predictions of weak timbre-correlated symbols on the basis of the pool of timbre-correlated attributes. The algorithm is able to substantially improve the precision of timbre-only estimations (sometimes by more than 15%), especially for descriptions such as “Situation Sailing” or “Situation Love” whose initial timbre estimate was poor.

This operational model shows that there is a critical mass effect in the number of high-level descriptions considered jointly. Figure 2 shows the mean improvement of precision for classifiers when reinforcing timbre-inference by correlation analysis (compared to timbre-only strategy). It appears that the more descriptions are considered, the better weak and ambiguous sensory mappings can be compensated. Symbolic complexity, as is arguably the case with polyphonic popular music, has a positive effect insofar as it provides denser networks of semantic associations.

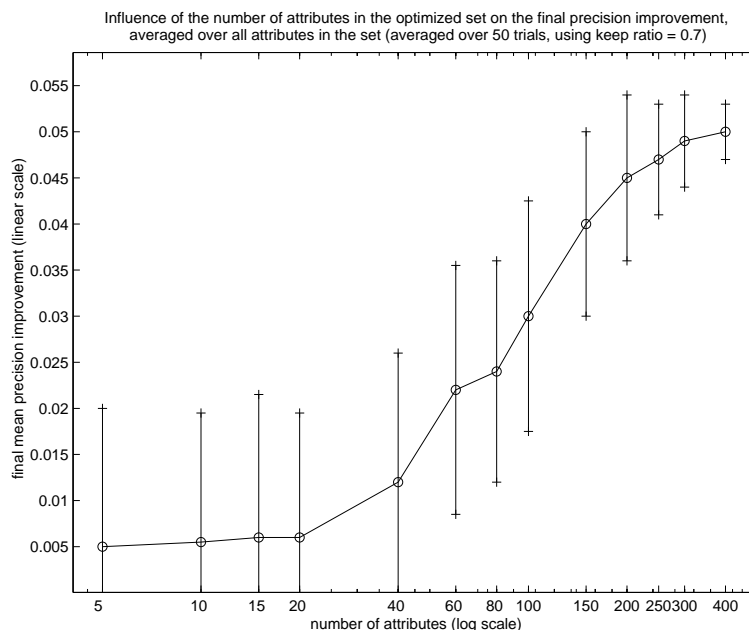


Fig. 2. Influence of number of symbolic attributes on the mean precision improvement over timbre-only inference

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 timbre similarity used here) could prove 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. What our present results show so far is that this proportion is likely to be more balanced than

generally assumed.

Acknowledgments

The authors wish to thank Anthony Beurivé and Pierre Roy for their support running the computational experiments described here. J.-J. A. also wishes to thank Petter Johansson and Ryoko Uno for their detailed comments and suggestions.

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

We give here a summary of the algorithm described in [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 $\log \mathcal{S}$.

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

We call mel-cepstrum the cepstrum computed after a non-linear frequency warping onto a perceptual frequency scale, the Mel-frequency scale ([Rabiner and 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 A.1 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 \mathcal{M} simpler Gaussian densities, called components or states of the mixture. ([Bishop, 1995]):

$$p(x_t) = \sum_{m=1}^{m=\mathcal{M}} \pi_m \mathcal{N}(x_t, \mu_m, \Sigma_m) \quad (\text{A.2})$$

where x_t is the feature vector observed at time t , \mathcal{N} is a Gaussian probability density function with mean μ_m , covariance matrix Σ_m , and π_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 (A.2)) is defined by :

$$d(A, B) = \int p_A(x) \log \frac{p_B(x)}{p_A(x)} dx \quad (\text{A.3})$$

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

$$d(\widetilde{A}, B) = \frac{1}{n} \sum_{i=1}^n \log \frac{p_B(x_i)}{p_A(x_i)} \quad (\text{A.4})$$

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

$$\lim_{n \rightarrow \infty} \left(\frac{1}{n} \sum_{i=1}^n X_i - \mathcal{E}(X) \right) = \frac{1}{\sqrt{n}} \mathcal{N}(0, \sigma^2) \quad (\text{A.5})$$

where X is the random variable $\log \frac{p_B(x)}{p_A(x)}$, 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 σ^2 , the variance of X .

B Appendix B: Nearest Neighbor classification

We give here a formalized description of the inference mechanism we use in Section 2 to quantify the acoustic grounding of high-level musical symbols.

We propose to infer the value of a given attribute \mathcal{A} for a given song \mathcal{S} by looking at the values of \mathcal{A} for songs that are timbrally similar to \mathcal{S} . More precisely, we define as our observation $\mathcal{O}_{\mathcal{S}}$ the number of songs among the set $\mathcal{N}_{\mathcal{S}}$ of the 10 nearest neighbors of \mathcal{S} for which \mathcal{A} is *true*, i.e.

$$\mathcal{O}_{\mathcal{S}} = \text{card}\{\mathcal{S}_i \setminus \mathcal{S}_i \in \mathcal{N}_{\mathcal{S}} \wedge \mathcal{A}(\mathcal{S}_i)\} \quad (\text{B.1})$$

If the attribute \mathcal{A} is correlated with timbre, large values of $\mathcal{O}_{\mathcal{S}}$ are a good indicator that $\mathcal{A}(\mathcal{S})$ is *true*. For instance, if 9 out of the 10 nearest neighbors

of a given song are “Hard Rock” songs, then it is very likely that the seed song be a “Hard Rock” song itself. However, we need to compensate this decision by the fact that attributes are not uniformly distributed in our database subset. For instance, “Genre Dance Music” is *true* for 4123 songs out of 4936, while “Main Instrument Bandoneon” has only one positive example. We thus define $P(\mathcal{A}(\mathcal{S})/\mathcal{O}_S)$ the probability that \mathcal{A} be true for \mathcal{S} given the observation \mathcal{O}_S of a given number of true values in the set of nearest neighbors, and $P(\overline{\mathcal{A}(\mathcal{S})}/\mathcal{O}_S)$ the probability that \mathcal{A} be false given the same observation. According to Bayes’ law,

$$p(\mathcal{A}(\mathcal{S})/\mathcal{O}_S) = p(\mathcal{O}_S/\mathcal{A}(\mathcal{S})) \frac{P(\mathcal{A}(\mathcal{S}))}{P(\mathcal{O}_S)} \quad (\text{B.2})$$

The likelihood distribution $p(\mathcal{O}_S/\mathcal{A}(\mathcal{S}))$ can easily be estimated by the histogram of the empirical frequencies of the number of positive neighbors for all songs having $\mathcal{A}(\mathcal{S}) = \textit{true}$ (similarly for $P(\overline{\mathcal{A}(\mathcal{S})}/\mathcal{O}_S)$). Figure B.1 shows 2 examples of such likelihood distributions computed for attributes “Character Calm” and “Genre Club/Discotheque”. If we assume a flat prior

$$P(\mathcal{A}(\mathcal{S})) = P(\overline{\mathcal{A}(\mathcal{S})}) = 0.5 \quad (\text{B.3})$$

we can estimate $\mathcal{A}(\mathcal{S})$ using the maximum likelihood criteria :

$$\mathcal{A}(\mathcal{S}) = p(\mathcal{O}_S/\mathcal{A}(\mathcal{S})) > p(\mathcal{O}_S/\overline{\mathcal{A}(\mathcal{S})}) \quad (\text{B.4})$$

Using the example given in figure B.1, we see that under the observation that 4 nearest neighbors out of 10 have “Character Calm”, we estimate that the seed song has “Character Calm”. However, under the same observation that 4 nearest neighbors out of 10 have “Genre Club/Discotheque”, we estimate that the seed song does not have “Genre Club/Discotheque”, because this observation is in fact surprisingly small given the large number of songs of “Genre Club/Discotheque” present in the whole database.

References

- [ISM, 2005] (2005). *Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005)*.
- [Aucouturier et al., 2007a] Aucouturier, J.-J., Defreville, B., and 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).

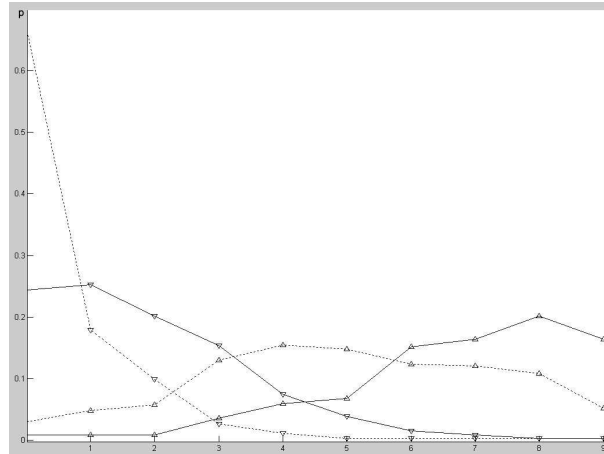


Fig. B.1. Likelihood distributions for 2 attributes “Character Calm” (downward pointing triangle) and “Genre Club/Discotheque” (upward pointing triangle). Positive likelihood $p(\mathcal{O}_S/\mathcal{A}(\mathcal{S}) = true)$ appear in solid line, and negative likelihood $p(\mathcal{O}_S/\mathcal{A}(\mathcal{S}) = false)$ in dashed line. The x-axis corresponds to the number of songs having $\mathcal{A}(\mathcal{S}) = true$ observed in the first 10 nearest neighbor of the seed song.

[Aucouturier and Pachet, 2004] Aucouturier, J.-J. and Pachet, F. (2004). Improving timbre similarity: How high’s the sky ? *Journal of Negative Results in Speech and Audio Sciences*, 1(1).

[Aucouturier et al., 2007b] Aucouturier, J.-J., Pachet, F., Roy, P., and Beuriv e, A. (2007b). Signal + context = better classification. In *Proceedings of the International Conference on Music Information Retrieval (submitted)*.

[Aucouturier et al., 2005] Aucouturier, J.-J., Pachet, F., and 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 et al., 2002] Bainbridge, D., Cunningham, S. J., and Downie, J. S. (2002). How people describe their music information needs: A grounded theory analysis of music queries. In *Proceedings, 3rd International Conference on Music Information Retrieval*, Paris, France.

[Berger and Fales, 2003] Berger, H. and Fales, C. (2003). The match of perceptual and acoustic features of timbre over time: “heaviness” in the perception of heavy metal guitar textures. Retrieved 10/01/2006 from www.indiana.edu/~savail/workingpapers/heavy.html.

[Bishop, 1995] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford Press.

[Borg and Groenen, 1997] Borg, I. and Groenen, P. (1997). *Modern Multidimensional Scaling*. Springer-Verlag, New York.

[Cangelosi and Harnad, 2000] Cangelosi, A. and Harnad, S. (2000). The adaptive advantage of symbolic theft over sensorimotor toil: Grounding language in perceptual categories. *Evolution of Communication*, 4(1):117–142.

- [Chai and B., 2001] Chai, W. and B., V. (2001). Folk music classification using hidden markov models. In *Proceedings of the International Conference on Artificial Intelligence*.
- [Dalla Bella and Peretz, 2005] Dalla Bella, S. and Peretz, I. (2005). Fine differentiation and ordering of classical music requires little learning but rhythm. *Elsevier Cognition*, 96.
- [Deacon, 1997] Deacon, T. (1997). *The Symbolic Species: The Coevolution of Language and Human Brain*. New-York: W. W. Norton.
- [Dhanaraj and Logan, 2005] Dhanaraj, R. and Logan, B. (2005). Automatic prediction of hit songs. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, London, UK.
- [Dubnov et al., 2006] Dubnov, S., McAdams, S., and 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 et al., 1997] Freedman, D., Pisani, R., and Purves, R. (1997). *Statistics, 3rd edition*. W.W. Norton and Co., New York.
- [Gagnon and Peretz, 2003] Gagnon, L. and Peretz, I. (2003). Mode and tempo relative contributions to “happy - sad” judgments in equitone melodies. *Cognition and Emotion*, 17:25–40.
- [Golder and Huberman, 2006] Golder, S. and Huberman, B. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 322:198–208.
- [Gomez, 2006] Gomez, E. (2006). Tonal description of polyphonic audio for music content processing. *INFORMS Journal on Computing, Special Cluster on Computation in Music*, 18(3).
- [Gouyon, 2005] Gouyon, Fabien and Dixon, S. (2005). A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–54.
- [Grey, 1977] Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *Journal of the Acoustical Society of America*, 61:1270–1277.
- [Guernsey, 1928] Guernsey, M. (1928). The role of consonance and dissonance in music. *American Journal of Psychology*, 40:173–204.
- [Janata, 2007] Janata, P. (2007). Timbre and semantics. Keynote Presentation, Journées fondatrices Perception Sonore, Lyon (France), January 2007. Available: <http://www.sfa.asso.fr/fr/gps>.
- [Koelsch et al., 2004] Koelsch, S., Kasper, E., Sammler, D., Schulze, K., Gunter, T., and Friederici, A. D. (2004). Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience*, 7:302 – 307.
- [Lakoff, 1987] Lakoff, G. (1987). *Women, Fire, and Dangerous Things*. Chicago, IL: University of Chicago Press.

- [Liu et al., 2003] Liu, D., Lu, L., and Zhang, H.-J. (2003). Automatic mood detection from acoustic music data. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR), Baltimore, Maryland, USA*.
- [McAdams et al., 1995] McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58:177–192.
- [Pachet and Cazaly, 2000] Pachet, F. and Cazaly, D. (2000). A taxonomy of musical genres. In *Proceedings RIAO*.
- [Perrott and Gjerdingen, 1999] Perrott, D. and Gjerdingen, R. (1999). Scanning the dial : an exploration of factors in the identification of musical style. In *Proceedings of the 1999 Society for Music Perception and Cognition*.
- [Rabiner and Juang, 1993] Rabiner, L. and Juang, B. (1993). *Fundamentals of speech recognition*. Prentice-Hall.
- [Tsai and Wang, 2003] Tsai, W.-H. and Wang, H.-M. (2003). Towards automatic identification of singing language in popular music recordings. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain*.
- [Tzanetakis et al., 2001] Tzanetakis, G., Essl, G., and Cook, P. (2001). Automatic musical genre classification of audio signals. In *proceedings ISMIR*.
- [Velmans, 1991] Velmans, M. (1991). Is human information processing conscious? *Behavioral and Brain Sciences*, 14:651–726.
- [Vincent and Rodet, 2003] Vincent, E. and Rodet, X. (2003). Instrument identification in solo and ensemble music using independent subspace analysis. In *Proceedings of the International Conference on Music Information Retrieval (ISMIR), Barcelona, Spain*.