MEL CEPSTRUM & ANN OVA: THE DIFFICULT DIALOG BETWEEN MIR AND MUSIC COGNITION

Jean-Julien Aucouturier & Emmanuel Bigand
LEAD/CNRS UMR 5022, University of Burgundy, Dijon, France.
aucouturier@gmail.com; bigand@u-bourgogne.fr

Mel is a MIR researcher (the audio type) who's always been convinced that his field of research had something to contribute to the study of music cognition. His feeling, however, hasn't been much shared by the reviewers of the many psychology journals he tried submitting his views to. Their critics, rejecting his data as irrelevant, have frustrated him - the more he tried to rebut, the more defensive both sides of the debate became. He was close to give up his hopes of interdisciplinary dialog when, in one final and desperate rejection letter, he sensed an unusual touch of interest in the editor's response. She, a cognitive psychologist named Ann, was clearly open to discussion. This was the opportunity that Mel had always hoped for: clarifying what psychologists really think of audio MIR, correcting misconceptions that he himself made about cognition, and maybe, developing a vision of how both fields could work together. The following is the imaginary dialog that ensued. Meet Dr Mel Cepstrum, the MIR researcher, and Prof. Ann Ova, the psychologist.

1. ON AUDIO FEATURES

Ann Ova: Let me start with a tentative definition of what we, music cognition researchers, are interested in. To me, cognition is like digestion: a chain of transformations affecting a stimulus (e.g. a piece of music reaching the ears), transforming it, breaking it into blocks and eventually metabolizing it to produce a behavior (an emotional reaction, recognition, learning, etc.). As researchers, we are seeking to understand this mechanism of "stimulus digestion": what in the signal triggers it, how it is activated, what brain/mind functions are required.

Mel Cepstrum: When I hear this, I form the impression that your collective goal is not very different from ours in Music Information Retrieval. First, we study the same behaviors: the recognition of music into melodies, artists, styles, genres, or the prediction of emotional reactions. Second, we too are looking for mechanisms, which we prefer to call algorithms, and we conceptualize them using similar steps: sensory transformations first (we'd call this the signal processing front-end or feature extraction), then linking to memory and learning (we'd say databases and statistical models). It is therefore surprising to me that a lot of work in music cognition tends to rely on audio characteristics that can be extracted "by ear", thus ignoring much of our work in the past 10 to 15 years on musical signal processing. For instance, of the nearly 1,000 pages of the Handbook of Music and Emotion [14], not a single one is devoted to computerized signal analysis, but examples abound of research asking participants to subjectively evaluate a musical extract's tempo, complexity, height etc. on scales from 1 to 5, so these characteristics can be correlated with what you call "behavior". While I understand this may have been the only approach available to, say, Robert Francès in 1958 [10], surely you do realize that all of this (pitch extraction, beat tracking, etc.) can now be automated with computer algorithms? What's the superiority of doing it by hand?

A.O. This is true, much of what we study is analyzed by hand, or rather "by ear", by participants. I believe the advantage of doing so is that we only consider as possible acoustic correlates of a given behavior constructs that can be cognitively assessed by the participants themselves. We want to use what they *really* hear, not what a computer thinks they hear, and the best way to do this is to simply ask them.

M.C. But you'll agree that there are unique advantages to automatic analysis: it's fast and cheap, you can process a large number of stimuli in just minutes, while it would take a large number of participants to do the same by ear.

A.O. I understand this is an important criteria in your field - certainly one does not want to index iTunes by hand, but this is not an important concern for us. If a particular experimental design is expensive in terms of experimenter and participant time, but it is the design of choice, so be it.

M.C. Right - but isn't automatic signal analysis also more objective? It can extract physical properties from the signal, e.g. the *root-mean-square* that qualifies its physical energy or the *zero-crossing-rate* which describes the noisiness of the waveform - without mediating these by cognitive judgments. It can also realistically simulate the audio processing chain of the peripheral auditory system. For instance, *Mel-Frequency Cepstrum Coefficients*, a mathematical construct derived from the signal's Fourier transform, are designed to reproduce the non-linear tonotopic scale of the cochlea and the dynamical response of the hair cells of the basilar membrane.

A.O. This is only partly correct, you see. If you look at MFCCs closely (take Logan [23], say), you see that parts of the algorithm were designed to improve their computational value for machine learning, and not at all to improve their cognitive relevance. That final discrete cosine transform, for instance, is used to reduce correlations between coefficients, which would make their statistical modeling more complex. Now, one could argue I guess

that the brain uses a similar computational trick - authors like Lewicki [19] are thinking along these lines, I suppose - but you'll agree that the responsibility rests on us, researchers, to prove that correct. Until then, MFCCs are maths. Useful maths for you perhaps, but irrelevant to our concerns.

M.C. Wait, that's a bit harsh. How about that study by Terasawa, Slaney and their colleagues at Stanford [31]: they resynthesized sounds from MFCCs and showed that human timbre dissimilarity ratings between sounds correlated exactly with the MFCCs. Doesn't that prove something?

A.O. Good one. This is indeed an important study, perhaps the first to tackle this problem diligently. But what does this prove, you ask? That an algorithmic construction, the MFCC, closely predicts a cognitive judgement. Should we conclude the brain implements a discrete cosine transform? Probably not. Just like fitting reactors on an airplane and seeing it take off should not lead us to conclude anything about how birds fly. Don't you think?

M.C. You're killing me. Are you seriously rejecting 10 years' worth of results as mere coincidences? Our findings that, say, taking the derivative of MFCCs improve genre classification by 10%, or that periodicities in the range 1-10 seconds (the *rhythm fluctuation patterns* of Pampalk [27]) are enough to account for timbre similarity, shouldn't that, at least, give you some sort of *intuition* about how these behaviors are cognitively produced?

A.O. Sorry if I sounded dismissive. In theory, you're right, and actually, we have been paying attention to your work (initiative like the MIRToolbox [17] have helped). But in practice, it's been really difficult to use your work, or to derive useful intuitions from it. Let me show you what I mean here, precisely. I'm looking at this data produced in my lab, a series of emotional valence and arousal judgements produced by participants listening to very short musical extracts (shorter than 500ms). At this level, it is unlikely that emotional reactions result from a cognitive analysis of say, melody or harmony, because the extracts are too short to even include a single note. The question for us is therefore to understand what low-level features of the raw sound are responsible for the emotion. It's the classical Gjerdingen & Perrott scenario [12], isn't it? This, if I understand correctly, is the ideal use-case for MIR features: a quasi-stationary signal, mostly important by its timbre quality. Well, let's have a look.

Table 1 reproduces the results of a multivariate regression we computed between the stimuli's valence and arousal and the whole batch of features offered by the MIRToolbox [17]. Let's see what intuition I, the cognition researcher, should derive from this. We see stimulus valence is very well explained by, let me get this right, the *entropy of the period of the magnitude of the maximum peak detected every 50ms in the signal's chromagram* (a chromagram, as you know, gives at each successive time position the energy observed at the frequency corresponding to each note - c, c#, d, etc., of each octave). Similarly, stimulus arousal seems to result from the *standard deviation of the 6th MFCC* and the *mean of the 3rd*, and - mind you - not the opposite.

Table 1: Top MIR features in a regression of valence and arousal emotional judgements

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

M.C. Hmm. This seems a bit too complicated maybe?

A.O. See what I mean? That surely fits well to the data, but I'm sure you realize it does not actually *explain* anything. Even if we took it literally, this would be a formidable mix-bag of an explanation. We have here an emotional reaction, valence, of which neuroscience tells us it is at least partly pre-attentive and subcortical, and which we explain here with constructs requiring memory and statistical learning ("entropy"), rhythmic entrainment ("period"), temporal integration ("maximum peak"), harmonic analysis ("chromagram") and arguably a participant's musical training in a western culture (because the chromagram relies on the 12-tone western pitch system).

M.C. Well, you got a point. But isn't this exactly the same problem when psychologists rely on features evaluated subjectively by their participants? When they study cultural differences between western and Indian classical music, Balkwill & Thompson [2] argue emotions are related to their stimuli's musical complexity, which they measure by asking participants, I quote, "*to evaluate how much was going on melodically in the except - was there a lot of repetition as in "Mary had a little lamb?*". Now, isn't that carrying a lot of assumptions too? The construct of "*being like Mary had a little lamb?*" is probably only derivable at a cortical level, using a lot of cognitive functions such as memory, melodic representations, etc. and certainly presupposes the participants know of that song in the first place. Are these assumptions realistic knowing what we know of emotions?

A.O. Well, you're probably right. And I guess one could even add that MIR has the advantage of not hiding these assumptions under their apparent lexical simplicity. But still, you have to admit that the logics behind your typical MIR signal feature is difficult for us to follow. If we want to use it to prove anything, it is crucially important for us to know *what* we're dealing with: a physical measure? a cognitive model? Let's have a look. In the MIR bestiary, we find, first, features deriving from traditional psychoacoustics: for instance, the *spectral centroid* which is the traditional correlate for the first perceptual dimension identified by MDS studies of timbre [13] or the *log attack time*, which correlates with the second most-important dimension; then, your field offers quite a lot of mathematical variants of these same characteristics, which seem to be justified only by the fact that they are concep-

tually close (for instance, *spectral skewness*, the 3rd spectral moment, which seems to be included because of the special status of the first moment, the above-mentioned *centroid*) or even that they are easy enough to compute (*spectral entropy*, obtained by multiplying the Fourier spectrum with its logarithm); other features seem to start their career as intermediary steps in the processing chain of another feature, gain special status and then a name of their own (for instance, the "*fluctuation pattern*" you mentioned earlier [27], which was originally an intermediary step in a tempo extraction algorithm); or even, as by-products of other algorithms, like some measures of *pulse clarity* [16] which are in fact the error estimation in the output of a beat-tracking algorithm. And the list goes on, growing every year: the sole MIRToolbox library offers more than 300 features, very few of which having a clear epistemological status. Now, I do not doubt they serve your purpose well, but I hope you see it is unclear whether they can serve ours.

2. ON PRECISION AND GROUNDTRUTH

M.C. I do. And I have to admit it never occurred to me that our drive to optimize our features for precision (deriving such features, selecting variants that work, recombining them, etc.) had taken us so far from the cognitive reality. Still, isn't it paradoxical that this same process is taking us closer and closer to the actual phenomenon, in terms of percentage of precision? I mean, work like Liu & Zhang [21] simulate with more than 95% precision the human judgements of "depression" and "contentement" made when listening to more than 250 extracts of music, by combining features describing timbre (e.g. spectral centroid), rhythm (e.g. average autocorrelation peak) and intensity (e.g. rms). Even if their algorithm has no pretense of being a cognitive model, the fact that it agrees with humans 95% of the time on hundreds of stimuli can only make us think it captures a large share of the physical and sensory features used by human cognition - right?

A.O. Let me ask a question. Isn't this definition of precision, relative to a so-called ground truth, a bit illusionary? Does everybody, in every culture, have the same exact definition of what is, say, "rock music", or of 2 songs that "sound the same"?

M.C. I see where you're going with this. We, MIR researchers, have always been uneasy about this point, to be honest. We're stuck between 2 research traditions: one, machine learning, which is interested in the capacity of algorithms to learn from a set of examples, whatever these examples are. For this tradition, whether the ground truth is meaningful or not is irrelevant. It is just taken as a temporary gold standard, relative to which different algorithms can be compared. Whether "rock" is indeed "rock" or "jazz" does not matter - actually, we want algorithms that have the flexibility to also learn that "jazz" is "rock" if we like them to. However, we also have the second research goal of being useful to electronic music distribution systems. Now, in this world, defining a unique ground truth is suddenly very relevant, but you soon realize it is also close to impossible: we have plenty of examples where what some call "rock", others will call "pop" or "jazz" and so on. I guess that's what you would call individual variations. Most of our recent research tries to address this paradox: for instance, how tags learned on one dataset generalize to other datasets [24], how to personalize music recommendations [5] or even letting uses define their own personal categories in interaction with the system [26]. But one cannot just rule out the idea of precision. After all, you psychologists also have to rely on the same concept: take the psychoacoustics of musical timbre. What these studies do is, similarly, consider average similarity ratings over many users (not that many, incidentally, compared with the thousands of samples we are routinely dealing with in MIR), and select features that explain the best percentage of the data's variance - finding, for instance, that the spectral centroid of an extract correlates at 93% with the first principal component of the human-constructed timbre space. But why should we accept spectral centroid as an important "psychologically-validated" characteristics of timbre, and simultaneously reject, say, Liu & Zhang's *average autocorrelation peak* [21] (or Alluri & Toiviainen's 6th *band spectral flux [1]*) when it allows to classify emotions at 95%? Sometimes, I wonder if you have a bit of a "not-invented-here" bias...

A.O. You may be right. Perhaps we have been disregarding advances in signal processing just because they look complicated and we can't be bothered to follow what you've been doing. Signal features produced by recent MIR research could and probably should be integrated to modern psychoacoustics, especially for those problems that could not yet be solved, such as dissonance [28], and you'll have to teach us on that. However, your using the example of psychoacoustics is interesting. I don't know if you realize that the psychoacoustics methodology is designed to investigate percepts, i.e. the immediate psychological gestalts corresponding to those few physical characteristics that define an auditory object, regardless of the listener and its culture. A musical sound *has* pitch, loudness and timbre. These are percepts. The same sound, however, does not *have* genre or emotion - these are constructed cognitively; their value could change (e.g. if you start calling the sound "pop" instead of "rock") without changing the physical definition of a sound. Now, to be honest, the frontier between what's a percept and what's a cognitive construction has been challenged in recent years, with the realization that action and perception are intertwined, but still most cognition researchers would agree a fundamental difference remains between the two. I'm worried you're applying the psychoacoustics metaphor to behaviors (genres, emotions, etc.) for which it does not apply.

M.C. This is fascinating. I realize just now that, all these years, I have been using the terms psychoacoustics and music cognition is a nearly interchangeable way. The more I think about it, the more I realize that indeed MIR takes a psychoacoustics approach to, as you say, genres and emotions, treating these as if they were a set of physical properties of the sound. What's surprising is that it

works so well. In fact, we're not capturing "rock" or "sad" music, we're capturing *things that sound like* "rock", or things that sound like a "sad song". Because music is a structured human activity, there are a lot of regularities there: most "sad" music indeed sounds the same (dark timbre, low pitch, what have you). But these features do not *make* the music sad -

A.O. - or at least, you're not testing whether they do -

M.C. right. We can potentially find music that *is* sad without exhibiting any of these features.

A.O. Take, say, that Dixieland upbeat tune they play at funerals in New Orleans.

M.C. Exactly. For these songs, our models will fail completely. But because such songs are rare (or at least they're rare in our test databases), say there are maybe 5% of them, we can still reach 95% performance without actually modeling anything specific about how, say, genre is cognitively constructed.

3. ON PHYSICAL AND COGNITIVE MODELS

A.O. It's a possibility, indeed. But you make it sound worse than it is, I think. It's not that your approach is better or worse than ours, but it's important that we understand the difference, and how we can be complementary. You're interested in the result, and how much algorithms and humans agree on it. In music cognition, I think we're less interested in the result than we are in the process. If we were to design computer algorithms to do maths, say, we're not interested in building machines that can multiply numbers as well and as fast as humans, but rather in doing them in such a way that multiplying 8*7 is more difficult than 3*4, as it is for humans.

M.C. This is indeed a true difference between our disciplines. We're happy when we see our algorithms duly classify as "rock" certain songs that are clearly on the border of that definition (Queen's Bohemian Rapsody, say) ...

A.O. ... whereas we would rather understand what makes a song more prototypically "rock" than another, or how much "rock" does one have to listen to form a stable representation of what that genre is.

M.C. But your problem in that case is how to measure prototypicality, because if you ask the same participants to judge it subjectively then your argument becomes completely tautological...

A.O. You're right

M.C. ... whereas MIR gives you a tool to do just this: a measure, let's say a physical measure, of the "rockness" of a song. How much it sounds like rock.

A.O. This, what you just said, is really interesting. The key word here is "physical". I believe that music cognition would gain a lot indeed if it had a more complete and powerful arsenal of tools to control stimuli physically. Tools that do not have the pretense of infringing into cognitive thinking, just purely, state-of-the-art physical modeling. If we start seeing MIR in this way, a lot of research avenues open I think.

M.C. In sum, in order to be useful to cognition, we should stop trying to do any ourselves.

4. ON FOLK PSYCHOLOGY

A.O. I can sense the irony, you know. This said, if I can make a small request, and I'm saying this in part jokingly but not solely, it would help indeed if you guys could at least stop using the word "semantics".

M.C. Wait... What?

A.O. "Semantics" - as in "a semantic model" of genre classification, "mixing acoustics with semantic" information, "semantic gap". Just, what do you mean by this?

M.C. Well, I suppose we take it as the "high-level" meaning of music, like saying "rock" is semantically related to youth and rebellion, electric guitars, all that linguistic and social knowledge around music. All which is where perception stops and, err..., cognition kicks in? Activating the semantic networks of musical concepts, err...

A.O. See: that. We hate it where you do that. Folk psychology. Like there is a box in our head somewhere with a knowledge base, and some kind of process that activates this or that depending on the input. You lose us instantly with that kind of thinking. If you browse the psychology literature, you will not find a single cognitive model which uses a "semantic" layer. That single word, let alone your using it assuming that it will appeal to us, does probably more harm to the dialogue between our disciplines than the mathematical complexity of your work. The "entropy of the period of the magnitude", I can deal with; "semantics", I sincerely have no idea. It literally drives me away.

M.C. Interesting - that certainly explains some reviewer reactions when I tried to communicate MIR results in psychological journals! Now, on the question of musical genre, you have to admit, conversely, that research in cognition does not have much to say about the links between social, lexical, sensory categories - all that we wrongly call "semantics". Neuroscience research has shown for instance that Wagner operas could prime recognition on such words as heroism, courage, etc. [15]. This has probably profound implications for everyday music perception. How come music cognition research is not studying this?

A.O. You're right. Most of us would consider that musical genre, as an object of study, is too complex, i.e. we know in advance that studying it won't help us isolate experimentally any particular process that could constitute it. For instance, if one wants to understand the sensory process by which a rock song is recognized as rock, it is simpler, more elementary if you will, to study the same process in the case of the recognition of environmental sounds. This latter case is less plagued by cultural learning, ambiguity, subjectivity that musical genre.

M.C. I see. Unfortunately, we in MIR don't have that luxury. If iTunes users want rock music, we cannot easily justify to study hammer noises instead.

A.O. Naturally. Once again, your discipline is interested in the result, and we are interested in the process.

5. INSPIRING EXAMPLES

M.C. But I'd like to backtrack a bit to your argument that MIR could contribute to cognition as a tool for physical modeling. This sounded promising.

A.O. Yes, I believe there is room to invent a methodology to use MIR tools to build a scientific proof in cognition. Precisely, MIR can be used, I think, as a physical measure of the information available for human cognition in the musical signal for a given task. And this measure can be used to control our stimuli and separate what's in the signal from what's constructed out of it by cognition.

M.C. I think there is work that already goes in the direction. In the speech domain, de Boer & Kuhl [8] for instance have shown that speech recognition algorithms (hidden Markov models) have better word recognition performance when they are trained and tested on infant-directed speech (IDS, or "motherese") than adult speech, which they claim validates the argument that the developmental value of IDS is to bootstrap language learning.

A.O. It's a lovely result, and indeed a very good example of how to integrate a physical, holistic recognition algorithm into a cognitive argument. What's important here is that the algorithm is not presented as a cognitive model: nobody here is pretending that the human brain implements a hidden Markov model. It only gives a proof of feasibility: from a purely physical point of view, the information exists in the IDS signal to allow for an easier treatment than adult speech. It would be very difficult to replace the machine by a human measure in this argument - computer modeling was really the clever thing to use.

M.C. There are a few other examples. For instance, Kaplan and colleagues [25] show that machine learning can classify dog barks into contexts like being afraid, playful, etc. This was taken to indicate, for the first time, that dog vocalizations contain communicative features. Like you said, from a purely physical, objective point of view, the point is to show that the information exists in the signal to allow for a potential communicative use.

A.O. I think we have discovered a design pattern here: one could probably imagine a similar application with music.

M.C. Let's see. Could we show for instance with a machine's good recognition performance that there exist enough harmonic information in, say, Indian classical music to explain the good performance (e.g. [2]) of western listeners when they are asked to classify emotions in raags (even though they are not familiar with this musical tradition)?

A.O. I confirm: this is exactly the type of question that's interesting for us, music cognition researchers, and indeed, I wouldn't know how to prove this without MIR. Provided the algorithm uses a representation of "harmony" which is both plausible biologically and agnostic culturally, of course. Some low-level measure of consonance/dissonance rather than a chromagram, perhaps?

M.C. Studying this would be pretty interesting from our point of view too. These characteristics found "enough to explain" a behavior would allow us to improve the performance of our algorithms in cross-cultural contexts, which are becoming a key issue in MIR. A recent international project, CompMusic [29], has even brought forth the question whether our most trusted algorithms have a western bias, because they were "evolved" with corpuses composed mostly of western music. By the way, you see, we're perhaps less naive than you first thought...

6. A SECOND LOOK AT BIOLOGICAL PLAUSIBILITY

A.O. I agree a lot of questions overlap between our 2 disciplines, perhaps more than I had first assumed. On the other hand, you'll also have to recognize that we are less hermetic to computational modeling than you think. I already mentioned the MIRToolbox, which is gaining interest among music psychologists. But most of our recent work include some element of computational modeling, often inspired by neuroscience. For instance, recent studies in harmonic priming [3] rely on fairly advanced computational models of auditory short-term memory [18]. Recently, human performance in tempo tracking was even explained in Journal of Experimental Psychology by a non-linear oscillation model [22].

M.C. This is true. Curiously, we MIR researchers are aware of these algorithms (auditory models by Leman [18], Large [22] or Cariani [4]), but for some reason they are not widely used. The criteria here is, again, precision. In our experience, when we first try them, "cognitively plausible" algorithms tend to work less well than brute-force engineer solutions, so they quickly drift into oblivion before we spend much time with them. One example was Lidy and Rauber's study [20] applying a wide range of "psychoacoustical" optimizations for genre recognition and finding very little improvement, if any. So we conclude, a bit hastingly maybe, that they're below our standards. But we've already discussed the difference between optimizing precision and modeling an underlying cognitive process.

A.O. Indeed. But even if you consider precision alone, the claim that cognitively plausible models will always perform poorly is not necessarily true, I think. In the image processing community, models which follow biological constraints radically are now giving comparable performance to their non-biologically-plausible alternatives (e.g. Serre, Poggio and colleagues [30]), and even faster learning rates. Now, you could argue that more is known in the psychophysiology of the visual cortex than for the auditory cortex, and it is therefore logical that machine vision should be ahead, but it is less and less the case, I reckon. We now have a good understanding of the re-

sponse patterns of neurons throughout the auditory pathway (see e.g. the idea of spectro-temporal receptive fields [11]), and computational models even exist to model them [6].

M.C. That's right. I have seen one application of these models to instrument timbre classification, but this was by researchers outside the MIR community [9], and I don't think we have really picked it up. I guess we should look into these more seriously.

A.O. I think we should indeed. The potential is not only better precision, but better interdisciplinary dialogue. Again, let's turn to machine vision for an example. The visual cognition community has now started to take inspiration from models like Serre's [30] to explain experimental results. For instance, I'm looking at a recent paper by Crouzet, Thorpe and colleagues [7], finding that humans are capable of ultra-fast face categorization. The authors write: "*Our ability to initiate directed saccades toward faces as early as 100–110 ms after stimulus onset clearly leaves little time for anything other than a feedforward pass. [Conveniently,] there is recent evidence (Serre et al. 2007) that such a purely feed-forward hierarchical processing mechanisms may be sufficient to account for at least some forms of rapid categorization*". In terms of interdisciplinary collaboration, I look at this with envy.

M.C. This is inspiring indeed! Let's work together so that, in a few years' time, we can write similar arguments in a similar article, linking a MIR model with some aspect of music cognition to derive a common scientific conclusion.

7. REFERENCES

[1] Alluri, V. & Toiviainen, P. (2010). Exploring perceptual and acoustic correlates of polyphonic timbre. Music Perception, 27(3), 223-241.

[2] Balkwill, L.& Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. Music Perception, 17, 43-64

[3] Bigand, E. & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. Cognition, 100(1), 100-130

[4] Cariani, P. (2001). Temporal Codes, Timing Nets, and Music Perception. Journal of New Music Research, 30(2), 107-136.

[5] Celma, O. & Lamere, P. (2011). If You Like Radiohead, You Might Like This Article, AI Magazine, 32(3), 57-66

[6] Chi, T., Ru, P. & Shamma, S. (2005) Multi-resolution spectrotemporal analysis of complex sounds. Journal of Acoustical Society of America, 118(2), 887-906

[7] Crouzet, S., Kirchner, H. & Thorpe, S. (2010). Fast saccades toward faces: Face detection in just 100 ms, Journal of Vision, 10(4):16.1-17

[8] De Boer, B. & Kuhl, P. (2003) Investigating the role of infant-directed speech with a computer model. Acoustics Research Letters On-line 4(4), 129–134

[9] Elhilali, M, Shamma, SA, Thorpe, SJ & Pressnitzer, D (2007) Models of timbre using spectro-temporal receptive fields: investigation of coding strategies, in proc. 19th International Congress on Acoustics, Madrid, Spain.

[10] Francès, R. (1958) La perception de la musique, Paris: Vrin.

[11] Ghazanfar, A. & Nicolelis, M. (2001). The Structure & Function of Dynamic Cortical & Thalamic Receptive Fields, Cerebral Cortex, 11(3):183-93.

[12] Gjerdingen, R. & Perrott, D. (2008) Scanning the Dial: The Rapid Recognition of Music Genres, Journal of *New Music Research,* 37*(2),* 93–100

[13] Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. Journal of the Acoustical Society of America, 61, 1270–1277.

[14] Juslin, P. & Sloboda, J. (2010) Handbook of Music and Emotion. Oxford University Press, USA

[15] Koelsch, S., Kasper, E, Sammler, D., Schulze, K., Gunter, T. & Friederici, A. (2004) Music, Language and Meaning: Brain Signatures of Semantic Processing, Nature Neuroscience 7, 302 - 307.

[16] Lartillot, O., Eerola, T., Toiviainen, P. & Fornari, J. (2008) , Multi-feature modeling of pulse clarity: Design, validation, and optimization, in proc 9th Int. Conference on Music Information Retrieval, Philadelphia PA, USA

[17] Lartillot, O. & Toiviainen, P. (2007) A Matlab Toolbox for Musical Feature Extraction From Audio, in proc. 10th Int. Conference on Digital Audio Effects, Bordeaux, France.

[18] Leman, M. (2000). An Auditory Model of the Role of Short-term Memory in Probe-tone Ratings. Music Perception, 17(4), 481-510.

[19] Lewicki, M.S. (2002). Efficient coding of natural sounds, Nature Neuroscience, 5 (4), 356-363.

[20] Lidy, T. & Rauber, A. (2005). Evaluation of Feature Extractors and Psycho-Acoustic Transformations for Music Genre Classification. in proc. 6th Int. Conference on Music Information Retrieval, London, UK.

[21] Liu, D. & Zhang, H.-J. (2006) Automatic mood detection and tracking of music audio signal, IEEE Transactions on Speech and Audio processing, 14(1), 5-18

[22] Loehr, J., Large, E. & Palmer, C. (2011). Temporal coordination in music performance: Adaptation to tempo change. Journal of Experimental Psychology: Human Perception and Performance, 37 (4), 1292–1309

[23] Logan, B. (2000) Mel Frequency Cepstral Coefficients for Music Modeling. in Proc. 1st Int. Conf. on Music Information Retrieval, Plymouth, MA, USA.

[24] Marques G., Domingues M., Langlois T. & Gouyon F. (2011). Three Current Issues in Music Autotagging. in proc. 12th Int. Conf. on Music Information Retrieval, Miami, FL, USA.

[25] Molnár, C., Kaplan, F., Roy, P., Pachet, F., Pongrácz, P., Dóka, A. & Miklósi, Á. (2008) Classification of dog barks: a machine learning approach. Animal Cognition, 11(3):389-400.

[26] Pachet, F (2008). The future of content is in ourselves. Computers in Entertainment,6(3), 2008.

[27] Pampalk, E., Flexer, A. & Widmer, G. (2005). Improvements of Audio-Based Music Similarity and Genre Classification, in proc. 6th Int. Conf. on Music Information Retrieval, London, UK.

[28] Peretz, I. (2008). The need to consider underlying mechanisms: A response from dissonance. Behavioral and Brain Sciences, 31, 590-591

[29] Serra, X. (2011). A Multicultural Approach in Music Information Research, in proc. 12th Int. Conf. on Music Information Retrieval - see also *http://compmusic.upf.edu/node/71*

[30] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M. & Poggio, T. (2007). Object recognition with cortex-like mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(3), 411-426

[31] Terasawa, H., Slaney, M. & Berger, J. (2005) . The Thirteen Colors of Timbre, in proc. IEEE Workshop on Appl. of Signal Proc. to Audio and Acoustics, New Paltz, NY, USA.