Getting Closer to the Essence of Music: The *Con Espressione* Manifesto

GERHARD WIDMER, Johannes Kepler University Linz, Austria

This text offers a personal and very subjective view on the current situation of Music Information Research (MIR). Motivated by the desire to build systems with a somewhat deeper understanding of music than the ones we currently have, I try to sketch a number of challenges for the next decade of MIR research, grouped around six simple truths about music that are probably generally agreed on, but often ignored in everyday research.

CCS Concepts:•**Applied computing** → **Sound and music computing;**

General Terms: AI, Machine Learning, Music

Additional Key Words and Phrases: MIR, music perception, musical expressivity.

ACM Reference Format:
Gerhard Widmer, 2016. Getting Closer to the Essence of Music: The Con Espressione Manifesto. *ACM Trans. Intell. Syst. Tech.* 00, 00, Article 1 (March 2016), 14 pages.
DOI: 0000001.0000001

1. ABOUT THIS MANIFESTO

This text wishes to point our research community to a set of open problems and challenges in Music Information Research (MIR) [Serra et al. 2013],[1] and to initiate new research that will hopefully lead to a qualitative leap in musically intelligent systems. As a *manifesto*, it presents the author's personal views[2] on the strengths and limitations of current MIR research, on what is missing, and on what some of the next research steps could be. The discussion is based on the conviction that musically gratifying interaction with computers at a high level of musical quality – which I take to be the ultimate goal of our research community – will only be possible if computers (and, in the process, we) achieve a much deeper understanding of the very *essence* of music, namely, how music is perceived by, and affects, human listeners. This is also the personal manifesto of a music lover who is somewhat dissatisfied with the level of musical sophistication exhibited by current MIR systems.

The specific motivation for publishing the manifesto at this point in time is the start of a large research project called *Con Espressione* (www.cp.jku.at/research/ConEspressione), generously supported by the European Research Council (ERC),

[1]The audience addressed here is all researchers whose goal is to develop computer systems that deal and interact with musical contents, or musicians, in ways that are musically meaningful and useful. That includes fields like Music Information Research (MIR), Sound and Music Computing (SMC), but also parts of neighbouring fields like musical informatics, computational music theory, performance research, and even music cognition and psychology. For notational convenience, I will keep referring to my target field as MIR, with the implicit understanding that it really encompasses a broader set of research communities.

[2]To emphasise the highly subjective character of this discussion, I have chosen to write the article in the first person, rather than using the generic "we" or the passive voice.

Author's addresses: G. Widmer, Dept. of Computational Perception, Johannes Kepler University Linz, Austria; and Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria. gerhard.widmer@jku.at

DOI: 0000001.0000001

which hopes to address some of the challenges, with a particular focus on the *expressive* aspects of music. But the problems discussed here go way beyond what a single research project can tackle. Thus, this is a call to the entire MIR community to consider some of the issues in designing their research agendas.

2. THE NEED FOR A DEEPER UNDERSTANDING OF MUSIC IN COMPUTERS

The field of MIR (and related fields) has made a lot of progress, has achieved some spectacular results, and has produced – and keeps producing – highly useful applications in the commercial world of digital music. The contributions in this volume are evidence of this. Computers can detect music in complex audio, can identify, track, classify, and tag songs; they can extract many structural elements from music signals, such as onsets, beats, rhythm patterns, metrical structure, melodic lines, harmonies; they can segment musical pieces based on sound similarity, local changes, or repetition, and use this for music summarisation and many other useful services. A good overview of the technical state of the art can be found in [Müller 2015].

Here are a few things our computers can *not* (yet) do:

(1) Distinguish between songs that I might find boring or interesting.[3]
(2) Return (among other pieces) Beethoven's piano sonata op.81a (*Les Adieux*) when asked for a piece of classical music with a surprise at the beginning.
(3) Classify Tom Jobim's or João Gilberto's rendition of *Garota de Ipanema* as more relaxed and 'flowing' than Frank Sinatra's.[4]
(4) Play along with human musicians (e.g., accompany a soloist in a piano concerto) in a musically sympathetic way, recognising, anticipating, and adapting to the musicians' expressive way of playing (dramatic, lyrical, sober, ...).

For item (1), the computer would have to have an idea of redundancy vs. unpredictability, structure vs. randomness, and the role of expectation in human listening; for (2), it would have to have (learned) a model of classical music style (in this particular case, harmony); for (3), it would have to analyse subtle aspects of performance (timing, intonation, articulation, etc.), and understand how they contribute to the musical and expressive character of a performance; and for (4), it would have to have all of the above capabilities (plus the ability to recognise and decide in real time). None of these aspects, I claim, has yet received sufficient attention in the MIR community, and all of these (and more) would be needed to bring musical computers closer to understanding the essence of music: how it affects human listeners.

One of the big, overarching goals for the next decade would thus be to *equip computers with a deeper 'understanding' of music, its qualities, and how they are perceived by humans, in order to support a new generation of music systems and services at a new level of quality*.[5] That is the central tenet of this manifesto. An important consequence

[3] Of course – as for all the other examples I am presenting here – there is no 'true' answer. There will be differences in judgment between listeners, depending on musical preferences, experience, mood, historical period, even social context. However, there are fundamental information-theoretic principles at work in (aesthetic) perception [Meyer 1956; Moles 1966]. A song consisting of one single chord repeated for three minutes, or a melody stepping up and down the major scale ten times in a row, will be perceived as monotonous, if not boring, by almost anyone. (Note that this is not an aesthetic judgment: monotony may be intended and understood as an element of style, even as an artistic or political statement; think of genres like Minimal Music, Punk Rock, or *'Krautrock'*).

[4] ... or am I the only one to think so?

[5] Of course, producing human-level performance on some task in a machine does not necessarily require mimicking humans – think of chess programs, for instance. (Thanks to François Pachet for reminding me of this.) Then again, "deeper understanding of music" does not necessarily mean "human-like". What I mean is computational models that can identify or predict some of the same kinds of patterns in music that educated human listeners would perceive. Developing such models will require us, as researchers, to understand more

of adopting this is to put the listener into the center of our focus, but in the role of an *active listener* rather than the general user-as-consumer with 'listening habits' and 'entertainment needs' as discussed in [Schedl et al. 2013].

Before proceeding, a qualification is in order. This text focuses exclusively on *Western (tonal) music*. Likewise, when speaking about listeners, listeners' expectations, etc., I will mean persons who have grown up with Western music and thus have a shared (if implicit) understanding of music and stylistic norms. Of course, there are many other important musical traditions in the world, and the MIR community is increasingly addressing these as well (e.g., in the CompMusic project [Serra 2011]). But looking at different musical cultures is beyond the scope of the present paper.

3. SIX THESES ABOUT MUSIC

I would like to structure my critique of the current state of the art by proposing – as is expected of a *manifesto* – six *theses* about music and its effect on listeners. None of these is probably controversial in itself – in fact, the theses are rather trivial truisms –, but each of them points to a particular aspect that has, I believe, not received enough attention in the MIR community. Each of the theses, if taken seriously, has implications for future research, which I will try to work out in the following.

I. Music is a temporal construct / process

Music as played and heard is a process that unfolds in time. Even when viewed as a static object (represented, e.g., by the printed score), a musical composition is organised along an abstract time line, where the ordering of musical events, and their placement on the abstract time grid, is essential. Clearly, then – and that has been noted by many authors – the (still) predominant *bag-of-frames (BOF)* approach to many music classification tasks is inadequate. It is inadequate even as a summarising model of the perceived *sound* of a piece [Aucouturier and Pachet 2004], and definitely as a model of a piece as heard by listeners, or intended by the composers and performers. For instance, its limitations for emotion recognition have been convincingly demonstrated in [Huq et al. 2010].

The obvious alternative – *temporal models*, or at the very least, features that incorporate some contextual information – has been examined by several authors, with mixed success. For instance, [Flexer et al. 2005] showed that using Hidden Markov Models (HMMs) for timbre modelling increases the likelihood of the models (i.e., the fit on the data), but does not improve similarity-based genre classification. Other authors [Madsen et al. 2014; Vaizman et al. 2011] demonstrate some improvement in emotion detection from audio, using temporal models. In all these cases, however, very simple, low-level audio features (MFCCs) were used as a basis. But as [Aucouturier and Pachet 2004] already concluded in 2004: summarising statistics over low-level sound features will not permit our systems to surpass a certain level of performance (the much-cited *'glass ceiling'*), and this statement still seems valid.

The conclusion then must be that – temporal models or not – frame-based audio features are fundamentally the *wrong representation level*. Simple intuition tells us that much of what we consciously perceive or expect in a piece is at the level of *events* – notes, chords, etc. – not short windows of sound textures. Consider the Beethoven *Les Adieux* example from above. The three opening chords of the first movement are clearly heard, by any listener, as three distinct events, and the listener is suprised at the onset of the third chord, not "somewhere between audio frames 710 and 725, where the distribution of chroma vectors changes". This is even more important because most

about human music perception. Or to put it in [Herrera et al. 2009]'s words, "[w]e will only develop music understanding systems by means of understanding music understanding".

pianists tend to slightly delay the third chord, thus heightening the level of expectation in the listener and, concomitantly, the level of surprise as the third chord is not the expected harmony.[6] It is only by perceiving the passage as made up of discrete events that the delay of the third event can be registered at all, and have the effect that it does. Likewise, a prediction of whether a song would be perceived as 'boring' or 'interesting' (which, among other things, will have to rely on some estimate of the musical complexity of the song [Russell 1982]), will probably not be possible on the basis of variance measures at the short-term feature level.

My hope, then, is to see more research in the future on temporal modelling of music *at the level of musically meaningful (partly discrete) events and patterns*.[7] The fact that our current technologies for tasks like audio source separation, onset detection, note transcription, chord identification, or melodic/motivic pattern discovery from audio are still notoriously unreliable and brittle (though even the latter has recently been shown to be at least partly feasible [Collins et al. 2014]), makes this particularly challenging.

II. Music is fundamentally non-Markovian

To my knowledge, almost all temporal models used in MIR are either of a Markovian kind, assuming a strictly limited range of dependency of the musical present on the musical past (as in HMMs, but also, for instance, in auto-regressive features or 'dynamic textures' [Barrington et al. 2010]), or have a kind of decaying memory (as in simple Recurrent Neural Networks (RNNs)). On the other hand, it seems clear that music is of a fundamentally *non-Markovian* nature. Music is full of long-term dependencies: most pieces start and end in the same key, even if they modulate to other tonalities in between; themes return at regular or irregular intervals, after some intermittent material; harmonic rhythm (the rate of change of harmonies) is similar in similar passages; and so on. Moreover, musical units (segments, phrases, etc.) tend to have certain lengths in terms of number of bars (often a power of two), and listeners are used to expecting the return of a refrain after a certain number of bars. This means that we need the ability to *count*, which low-order Markov models are incapable of.[8]

What is needed, first of all, is to broadly acknowledge the non-Markovian nature of music and be critically aware of the fundamental limitations of HMMs and similar models in describing music. I do not always see that in the MIR literature. Second, we need more research on complex temporal models with variable degrees of memory. An example of recent work attempting to create a model that accounts for temporal dependencies in polyphonic music is [Boulanger-Lewandowski et al. 2012], which employed a complex, hybrid network made up of Restricted Boltzmann Machines (RBMs) and RNNs. However, evidence that the network does capture temporal dependencies is only indirect (via likelihoods and prediction accuracies), and in the end the authors had to conclude that long-term structure seems still out of the model's reach. Perhaps more promising are recent advances in RNNs with Long-Short-Term Memory (LSTM) units [Hochreiter and Schmidhuber 1997]. Actually, [Eck and Schmidhuber 2002] showed already in 2002 that LSTMs are capable of learning longer-term depen-

[6] For the uninitiated reader: the third chord is a surprising C minor instead of the Eb major that listeners will (consciously or unconsciously) expect. The resulting chord progression, notated as I-V-vi, is appropriately called a *deceptive cadence* in music theory. It is not too abundant in classical music, and extremely rare at the very beginning of pieces – and thus all the more unexpected and surprising here.

[7] This may also help avoid the *Clever Hans* effect recently identified in various MIR systems [Sturm 2014], which is clearly related to these systems focusing on features at musically irrelevant levels.

[8] Some of these problems are addressed in recent work on *Markov Constraint* models, such as [Roy and Pachet 2013], which proposes a solution for the counting problem in musical meter and, more recently, [Papadopoulos et al. 2015], which presents a method for sampling Markov sequences that satisfy some regular constraints (represented by an automaton).

dencies and structure in music. More recent work in MIR using LSTM networks has mostly focused on low- to mid-level tasks such as onset detection [Eyben et al. 2010], note transcription [Böck and Schedl 2012], or chord identification [Sigtia et al. 2015], where rather local context is sufficient. A lot more work is needed on models that can predict musical events and patterns over longer timespans.

Complementing this, it will be important to invest more research efforts into learning structural *abstractions*, over which temporal dependencies can then be modelled. Much of tonal music has a multi-level, often hierarchical organisation, with higher-level building blocks made up of smaller patterns (e.g., the ubiquitous ii-V-I chord progressions in Jazz). At a high level, one could then get by with low-order Markov dependencies. Hybrid architectures that can learn multi-level abstractions and temporal relations simultaneously (as, allegedly, *Hierarchical Temporal Memory (HTM)* [Hawkins and George 2006] can), would be particularly attractive.

III. Music is perceived by human listeners

In much of current MIR research, a recording is taken directly as a representation of a piece of music, from which computers then extract patterns such as beat, segment structure, etc. This pragmatic approach may be sufficient for practical applications such as music synchronisation or indexing, but when our goal is to predict more refined human categorisations (such as, e.g., emotions or interestingness), we need to remember that the ultimate place of music is in the *human mind* [Wiggins et al. 2010]: what we hear and how we respond to music is a product of an active *process of perception*, and only by understanding that process will we ultimately be able to predict some of music's effects.[9]

Human musical memory, and our conceptualisation of a piece of music, critically rely on abstraction and grouping. Humans are exceptionally good at segmenting the stream of musical events into meaningful units, on-line, while listening [Deutsch 2013]. In trying to explain this, music psychologists often appeal to the 'laws' of *Gestalt psychology* [Wertheimer 1938]. Lerdahl & Jackendoff's [1983] highly influential *Generative Theory of Tonal Music* derives various grouping rules from such Gestalt principles, in different structural dimensions (grouping, meter, hierarchical pitch abstraction, tension/release) with intuitively convincing structural predictions on selected classical music examples (though the Gestalt approach to music segmentation has also been challenged [Bod 2002]). Implicitly, some of these Gestalt concepts also play a role in current music segmentation algorithms [Paulus et al. 2010] (e.g., similarity of recurring segments as a grouping criterion, or local changes in some feature dimensions as indicators of boundaries), but the hierarchical abstraction ('time-span reduction') and tension-release ('prolongational reduction') models, which are particularly interesting from a music perception point of view, have not found their way into the world of practical MIR. One reason is that the rules as given are not free of ambiguities, contradictions, and cyclic dependencies, which so far has prevented researchers from fully implementing the theory even at the level of symbolic scores [Hamanaka et al. 2006]. I do believe it would be worthwhile to dig further into this (perhaps via probabilistic modeling, to address various inconsistencies). The ultimate challenge will be to apply similar principles to musical grouping at the audio level, and combine this with the best of current MIR audio segmentation algorithms.

[9]Actually, a full account of music perception would even go beyond the 'mind', acknowledging that the body of the listener and social interactions also play an important role. Current theories on *embodied and social cognition* (e.g., [Leman 2008]) are highly relevant, but to keep the presentation focused (and, admittedly, for a lack of conrete ideas on how to adequately address this aspect), I will leave these out of the present discussion.

A second aspect that is vital to our perception and appreciation of music is the *dynamics* of the listening process, the permanent ebb and flow of tension and release; of anticipation and realisation; of expectations emerging about what is to come next (and when), and their confirmation or denial. Authors like [Meyer 1956; Narmour 1992; Huron 2006] argue that this is a major source of the aesthetic and emotional effect of music, and the reason why we may be enthralled by a piece, or lose interest. That there is a correlation between the predictability of certain musical features, and the emotional response reported by listeners, has also been shown in [Dubnov et al. 2006]. In modelling this, it seems natural to take an information-theoretic approach, using notions like conditional entropy and information content to quantify the listener's uncertainty about what is to come next, and her surprise, or lack thereof, at what really comes next. Such models of musical expectancy have been advocated by several researchers in recent years [Abdallah and Plumbley 2008; Pearce and Wiggins 2012; Temperley 2007]. While the principal ideas are extremely elegant and appealing, there are severe problems in applying them to non-trivial kinds of music – in particular, how to model joint probability distributions over huge spaces of musical events. Current experimental models evade this by making strong Markov assumptions and restricting the experiments to strictly monophonic [Pearce and Wiggins 2012] – even isochronous [Abdallah and Plumbley 2008] – music. They are thus useful as theoretical models of musical learning and expectancy, but not yet for practical applications. Again, my conviction is that the (only) approach to making this tractable for complex music is by *abstraction*, i.e., modelling music in terms of higher-level patterns.

An interesting aspect of these information-theoretic models, as indicated by first experimental results, is that they can also help in predicting perceived grouping and segment boundaries [Pearce et al. 2010b]. I believe it would be extremely important to carry this kind of work further, towards more complex and realistic musical scenarios.

IV. Music perception and appreciation are learned

The question of which fundamental mechanisms – if any – of music perception are innate, and which ones are learned, is interesting but beyond the scope of this paper. Arguably, all of the higher-level patterns and the 'meanings' of music are learned, to a large extent simply through exposure [Patel 2008]. That gives us hope to also make substantial progress in machine understanding of music via massive unsupervised learning. The potential of statistical learning for explaining the emergence of musical expectation has been demonstrated in [Pearce et al. 2010a; Pearce and Wiggins 2012]. Ongoing advances in the field of *deep learning architectures* [Bengio 2009] now promise to provide a general basis for learning features and representations directly from raw data [Humphrey et al. 2013]. Deep learning models have shown promise in several MIR tasks, from speech and music detection [Schlüter and Sonnleitner 2012], to audio segmentation [Ullrich et al. 2014], but also predicting performers' expressive dynamics from scores in classical piano music via score features learned in this way [Grachten and Krebs 2014] (to cite just some of our own work). Particularly attractive are unsupervised learning scenarios, which promise to make it possible to exploit large musical datasets without the need for expensive manual annotation.

This is now the time for the MIR community to embark on massive feature / representation learning endeavours – much like the current trend in image analysis, which starts to produce quite spectacular results (e.g. [He et al. 2015]). Given the computational and data-related demands, the MIR community should join forces and pool its resources, efforts, and learned models (in cases where the training data itself cannot be shared) – and indeed, it has already begun to do so (see [Porter et al. 2015; Weyde et al. 2015] for two recent initiatives). Also inspiring is recent work on networks that learn to verbally describe the content of images [Vinyals et al. 2015]. Music search en-

gines that support description-based search without the need for manual annotation would be extremely useful not only in the general consumer music market, but also in specialised domains such as 'production music', where customers search for music with very specific properties.

Two aspects of music should again be kept in mind: the *multi-level* structure of music that ranges from timbre and sound through notes, chords, rhythmic patterns, harmonic patterns (e.g., cadences), melodic motifs, themes, sections, etc.; and the fact that different music-parametric dimensions (melody, harmony, rhythm) *interact* in complex ways. Learning useful and musically effective representations will benefit from a careful design of learning architectures, guided, wherever possible, by thoughtful analysis of the nature of music.

Long-term style learning would lead to building blocks like typical cadences, typical harmonic progressions, melodic clichés, accompaniment patterns, and the like. But interactive real-time systems (such as an automatic music accompanist) will also need learning at a different time scale: *short-term, on-line, intra-piece learning*, to induce some of the characteristics of the currently playing song. This is how we develop, during listening, very specific expectations about how a song is going to continue, and when to expect certain things (like the next chord change, or the return of the refrain). There has been relatively little research on this in the MIR community that I am aware of; most relevant is work on machine improvisation (e.g., [Assayag and Dubnov 2004; Nika and Chemillier 2012; Pachet 2003]), or on learning to anticipate the timing of events in expressive performance models [Raphael 2010; Arzt and Widmer 2010].

A big open problem is how to *integrate* long- and short-term learning. It is not at all obvious according to what principles that should be done. The *IDyOM* model of [Pearce and Wiggins 2012], which in its current form is a learning-based model of melodic prediction (at a symbolic level), simply predicts a probability distribution over the next pitch that is a weighted sum of the predictions of a long-term and a short-term learning model (weighted by the respective Shannon entropies). For more complex learning and prediction tasks, this will become more difficult. Unfortunately, there is precious little that music psychology and cognition research can tell us about this, in concrete terms. We may also see this as an opportunity: new computational or information-theoretic models we may come up with might serve as a source of inspiration to the music psychology world.

V. Music is (usually) performed

In almost all kinds of music, musical compositions are performed (translated into sound) by human musicians, and the details of the performance contribute much to the character of the music, and how it affects listeners. Expressive performance serves several functions [Palmer 1997], most importantly, to clarify the musical structure of a piece to the listener, and to highlight and communicate expressive and affective qualities of the music (see also item VI below).

Surprisingly, the aspect of performance has not seen too much attention in the MIR literature so far. That is a pity, as expressive performance can contribute much to the effect of music, and to qualities that one may like or dislike in a recording. Sophisticated music recommenders and other services should be aware of that. Moreover, interactive music systems, such as the automatic accompaniment system mentioned in Section 2 above, will need the ability to recognise and emulate different performance and expression styles, in real time (in addition to being able to anticipate expressive timing, as in [Raphael 2010]).

Many of the features we need to extract from a recording in order to account for performance aspects, are rather different from the audio features mainly in use in MIR. For instance, the 'flowing' character of João Gilberto's renditions of Bossa Nova songs –

in contrast to Frank Sinatra's singing or Stan Getz's saxophone playing – is due to the fact that he sings *against* or *'above'* the steady beat of the song, taking incredible and at the same time extremely natural-sounding liberties (a characteristic of Gilberto's art of Bossa Nova singing). Current beat tracking algorithms (e.g., [Krebs et al. 2015]) would readily recognise the steady 4/4 beat; but characterising this floating on top of the beat requires additional, in a sense 'orthogonal' features.

Depending on the instrument, there is a large variety of parameters that performers can control and shape, from tempo, timing, loudness, articulation to complex continuous aspects such as intonation, vibrato, timbral control of the singing voice[10], etc. We currently do not even have features that can quantify the degree of 'staccatoness' vs. 'legatoness', much less methods for exactly measuring micro-timing in chords, the sound balance between individual voices in a polyphonic piece, or qualities of singing. It would be desirable, at some point, to have such performance-related features included as a standard part of MIR feature extraction toolkits, alongside the current feature sets describing timbre (e.g., spectral centroid, MFCCs), rhythm (e.g., beat histograms), and the like.

Recognising and characterising performance-related aspects in music is one problem. Another is to build *predictive models* that can *produce* performances with certain musical qualities.[11] In the context of classical music, there has been quite some research on computational models of expressive music performance, as summarised in a 2004 survey [Widmer and Goebl 2004] – and the state of the art has not really improved a lot since then (though YQX has successfully performed Chopin in a RENCON computer piano performance contest [Widmer et al. 2009]). The group of researchers working on computational performance analysis and modelling has traditionally been quite small. Bringing the power of the full MIR community (with its interests that extend way beyond the narrow world of classical music) to bear on this class of problems would be extremely promising.

VI. Music is expressive and affects us

Approaches to automatic music recommendation have been rather superficial so far, evading the issue of what listeners actually hear, and *why* they might like a song. Typical music recommenders rely on indirect information such as timbral and rhythmic similarity between songs, expert-curated or web-crawled meta-data, user-provided tags, collaborative filtering, and/or features characterising the geographical and activity context of users [Song et al. 2012].

But music is more than that. It moves us; it affects us; a song may cause us to feel elated or sentimental; we may be touched by the mourning, solemn character of a funeral march.[12],[13] I strongly believe that MIR systems should be aware of this dimension, at least to the extent that they can find music for us that really has the potential to satisfy our musical and affective needs. The recent increase in emotion recognition research [Kim et al. 2010; Yang and Chen 2012] shows that the MIR community acknowledges the importance of that dimension.

[10]To get an impression of the richness of expressive possibilities in vocal art, listen to any recording of, say, Sarah Vaughan or Abbey Lincoln.

[11]Again, the automatic accompanist is a use case for this; quite another one would be systems that adapt the expressive character of music to dynamically changing situations, e.g., in video games or interactive movies.

[12]Be aware of the difference between the *arousal* of emotions, and the 'mere' *expression* or *communication* of emotions [Gabrielsson 2002] – a distinction that is not always clearly made in the MIR literature. In either case, however, the ability of music to express emotions and moods, or to *incite* [D. Cope, personal communication] listeners to construct musical and affective meaning, considerably adds to its importance and power as an art form.

[13]A lucid (and short) discussion of different philosophical views on this can be found in [London 2000].

However, I contend that the set of qualities that music can express is much broader than 'just' emotions or moods, and thus broader than the dimensions usually targeted in emotion detection. [Juslin 2013] distinguishes three levels of 'coding' of expressive messages in music, and through that, implicitly, different classes of expressive content: while basic emotions are communicated via rather direct cues like loudness, tempo, vocal qualities in singing – and are thus perhaps most directly accessible to MIR systems via audio features –, more complex and abstract qualities arise from the structure of the music itself, its implications of melodic/harmonic/rhythmic tension, release, realisation or denial (see above). Juslin calls this *intrinsic coding* and suggests that such factors, "[b]y contributing dynamically shifting levels of tension, arousal and stability, [...] may help to express more complex, time-dependent emotions, such as relief and hope" – to which one might add such qualities as uncertainty, determination, humour, but also power and physical motion and other things that I would not subsume under categories like emotions or moods. Juslin's third level of meaning assignment – *associative coding* – relates to expressive meanings that are purely conventional and socially learned, such as that a song that is presented to us as a national anthem conjures up images of patriotic pride or nationalism (as the case may be). Such meanings are not necessarily related to any specific properties of the music itself.

From the above follow several research challenges. First of all, there is a need for broad empirical investigations on what kinds of expressive qualities humans can (relatively) reliably and consistently recognise in music. Second, we need to categorise these, and define appropriate vocabularies or description frameworks. I believe that the popular categorisation models for emotions – e.g., valence-arousal space [Thayer 1989], 'circumplex model' [Russell 1980], Geneva Emotional Music Scale (GEMS) [Zentner et al. 2008] – cannot capture all the expressive qualities that music can convey. The set of categorical 'mood adjectives' used in the MIREX Mood Classification Task for popular music[14] contains a number of interesting concepts that actually go beyond moods proper (for instance, I would claim that 'rowdy', 'whimsical', or 'literate' are neither emotions nor moods). But again, I fail to see any systematic evidence that this set covers all the qualities we can and want to distinguish.

In designing algorithms that can recognise and classify expressive dimensions, different sources of expressivity must be considered. Simple *'surface properties'* like tempo, dynamics, timbre and chosen instruments, mode (major/minor), seem to most directly communicate (and partly even induce) basic emotions [Juslin 2013]. Then there is the *structure* of the composition itself, with its ups and downs, games of tension and release, and more or less dramatic twists and turns, which are more challenging to capture in terms of features (see above). *Culturally defined* meanings whose source is outside the music itself will only be accessible or inferrable to MIR systems from extra-musical sources – particularly, the Web (e.g., [Knees and Schedl 2013]). A final aspect that has been largely ignored so far in the MIR world is, again, *performance*. Especially in classical music, the specific way in which performers play a piece has a tremendous influence on the perceived character of the resulting music (see the little *Con Espressione Game* in the next section) – but this also goes for João Gilberto's or Sarah Vaughan's singing, or any other performed music (including, yes, *Kraftwerk*). The *Con Espressione* project will place a special focus on performance as a source of expressivity, but all of the above description levels will be needed, as performance cannot be seen independently of the piece itself and its structure [Gabrielsson and Lindström 2010].

[14] http://www.music-ir.org/mirex/wiki/2013:Audio_K-POP_Mood_Classification

4. THE *CON ESPRESSIONE* PROJECT

Con Espressione is a five year research endeavour (2016-2020) funded by the European Research Council (ERC) (http://www.cp.jku.at/research/projects/ConEspressione). Its goal is to lay the foundations for a new generation of music systems that are aware of, or can recognise and characterise, *expressive* aspects of music. The primary focus (owing to my research team's extensive experience and prior work) will be on classical music and expressivity as communicated via expressive *performance*. In approaching this, we will have to address some of the challenges discussed above. Specifically, we will

— investigate description frameworks for characterising and categorising (intended and perceived) expressive dimensions;
— advance research on extracting performance parameters (beyond timing and dynamics) from audio recordings and live performances;
— work on computational models of structure perception in music (at score and audio levels), combining recent MIR advances with information-theoretic approaches, unsupervised learning, and knowledge from musicology;
— investigate the relation between musical structure, expressive performance, and the communication of expressive characters;
— learn discriminative models that recognise intended expressive messages in performances;
— learn predictive models that generate or modify performances to express certain intended qualities.

All this will involve large-scale machine learning, using large curated corpora currently in preparation. One of the demonstrators we hope to deliver by the end of the project is the *Compassionate Music Companion*, an interactive system that plays along with human musicians in a musically sympathetic way, recognising, anticipating, and adapting to the musicians' expressive way of playing, and providing musical interaction at a gratifying level.

The Con Espressione Game

I would like to take the opportunity to invite the reader to the *Con Espressione Game*: the link `bird.cp.jku.at/con_espressione_game` will take you to a page that asks you to listen to excerpts from five Mozart sonata renditions by different pianists, and to enter words which, to you, best describe the perceived character of the recordings. This is a first small test to see to what extent (if at all) there is consensus in the perception of expressive qualities in classical piano performances. The collected responses will be analysed and the results announced to the MIR community in due course.

5. CONCLUSION

This little manifesto has reminded the reader of a few simple truths about music (rather pompously called 'theses' here), and what they might imply in terms of research challenges for our field:

i. Music is a temporal construct / process
ii. Music is fundamentally non-Markovian
iii. Music is perceived by human listeners
iv. Music perception and appreciation are learned
v. Music is (usually) performed
vi. Music is expressive and affects us

The ERC project *Con Espressione* will expressly try to address some of the questions that follow from these principles, focusing on the expressivity of music and music performance. My hope is that this manifesto will stimulate other research teams to join in this effort, so that future MIR systems will understand a bit more of the *essence* of music, and will be able to provide services at a new level of musical quality.

ACKNOWLEDGMENTS

The author gratefully acknowledges generous long-term financial support for this research by the Austrian Science Fund FWF (Wittgenstein Award 2009, project number Z159) and the European Research Council (ERC Advanced Grant, project number ERC-2014-AdG 670035). Thanks to David Cope, François Pachet, Perfecto Herrera, Arthur Flexer, Werner Goebl, Thomas Grill, André Holzapfel, Rainer Kelz, Peter Knees, Florian Krebs, Markus Schedl, Jan Schlüter, and Reinhard Sonnleitner, as well as two anonymous reviewers, for helpful comments and hints, with apologies for not having been able to incorporate all the (very appropriate and useful) comments in a short article like this.

REFERENCES

S. Abdallah and M. Plumbley. 2008. Information Dynamics: Patterns of Expectation and Surprise in the Perception of Music. *Connection Science* 21, 2-3 (2008), 89–117.

A. Arzt and G. Widmer. 2010. Simple Tempo Models for Real-time Music Tracking. In *Proceedings of the 7th Sound and Music Computing Conference (SMC 2010)*. Barcelona, Spain.

G. Assayag and S. Dubnov. 2004. Using Factor Oracles for Machine Improvisation. *Soft Computing* 8 (2004), 1–7.

J.J. Aucouturier and F. Pachet. 2004. Improving Timbre Similarity: How High is the Sky. *Journal of Negative Tesults in Speech and Audio Sciences* 1, 1 (2004), 1–13.

L. Barrington, A. Chan, and G. Lanckriet. 2010. Modeling Music as a Dynamic Texture. *IEEE Trans. on Audio, Speech, and Language Processing* 18, 3 (2010), 602–612.

Y. Bengio. 2009. Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning* 2, 1 (2009), 1–127.

S. Böck and M. Schedl. 2012. Polyphonic Piano Note Transcription with Recurrent Neural Networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*. Kyoto, Japan.

R. Bod. 2002. Memory-based Models of Melodic Analysis: Challenging the Gestalt Principles. *Journal of New Music Research* 30, 3 (2002), 27–36.

N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent. 2012. Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*. Edinburgh, Scotland.

T. Collins, S. Böck, F. Krebs, and G. Widmer. 2014. Bridging the Audio-Symbolic Gap: The Discovery of Repeated Note Content Directly from Polyphonic Music Audio. In *Proceedings of the 53rd AES Conference on Semantic Audio*. London, UK.

D. Deutsch. 2013. Grouping Mechanisms in Music. In *The Psychology of Music (3rd Ed.)*, D. Deutsch (Ed.). Academic Press.

S. Dubnov, S. McAdams, and R. Reynolds. 2006. Structural and Affective Aspects of Music from Statistical Audio Signal Analysis. *Journal of the Americal Society for Information Science and Technology* 57, 11 (2006), 1526–1536.

D. Eck and J. Schmidhuber. 2002. Learning the Long-Term Structure of the Blues. In *Artificial Neural Networks - ICANN 2002*. Springer Verlag, Berlin, 284–289.

F. Eyben, S. Böck, B. Schuller, and A. Graves. 2010. Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*.

Utrecht, The Netherlands.
A. Flexer, E. Pampalk, and G. Widmer. 2005. Hidden Markov Models for Spectral Similarity of Songs. In *Proceedings of the 8th International Conference on Digital Audio Effects (DAFx 2005)*. Madrid, Spain.
A. Gabrielsson. 2002. Emotion Perceived and Emotion Felt: Same or Different? *Musicae Scientiae* Special Issue 2001-2002 (2002), 123–147.
A. Gabrielsson and E. Lindström. 2010. The Role of Structure in the Musical Expression of Emotions. In *Handbook of Music and Emotion: Theory, Research, Applications*, P. Juslin and J. Sloboda (Eds.). Oxford University Press, New York, 367–400.
M. Grachten and F. Krebs. 2014. An Assessment of Learned Score Features for Modeling Expressive Dynamics in Music. *IEEE Transactions on Multimedia* 16, 5 (2014), 1211–1218.
M. Hamanaka, K. Hirata, and S. Tojo. 2006. Implementing A Generative Theory of Tonal Music. *Journal of New Music Research* 35, 4 (2006), 249–277.
J. Hawkins and D. George. 2006. *Hierarchical Temporal Memory: Concepts, Theory, and Terminology*. Numenta, technical report.
K. He, X. Zhang, S. Ren, and J. Sun. 2015. *Delving Deep into Rectifiers: Surpassing Human-level Performance on Imagenet Classification*. arxiv preprint arxiv:1502.01852 (2015).
P. Herrera, J. Serrà, C. Laurier, E. Guaus, E. Gómez, and X. Serra. 2009. The Discipline Formerly Known as MIR. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR 2010)*. Kobe, Japan.
S. Hochreiter and J. Schmidhuber. 1997. Long Short-term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
E. Humphrey, J. Bello, and Y. LeCun. 2013. Feature Learning and Deep Architectures: New Directions for Music Informatics. *Journal of Intelligent Information Systems* 41 (2013), 461–481.
A. Huq, J.P. Bello, and R. Rowe. 2010. Automated Music Emotion Recognition: A Systematic Evaluation. *Journal of New Music Research* 39, 3 (2010), 227–244.
D. Huron. 2006. *Sweet Anticipation: Music and the Psychology of Expectation*. MIT Press, Cambridge, MA.
P. Juslin. 2013. What Does Music Express? Basic Emotions and Beyond. *Frontiers in Psychology* 4, article 596 (2013).
Y. Kim, E. Schmidt, R. Migneco, B. Morton, P. Richardson, J. Scott, J. Speck, and D. Turnbull. 2010. Music Emotion Recognition: A State of the Art Review. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*. Utrecht, The Netherlands.
P. Knees and M. Schedl. 2013. A Survey of Music Similarity and Recommendation from Music Context Data. *ACM Transactions on Multimedia Computing, Communication and Applications* 10, 1 (2013), 2:1–2:21.
F. Krebs, S. Böck, and G. Widmer. 2015. An Efficient State-Space Model for Joint Tempo and Meter Tracking. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*. Malaga, Spain.
M. Leman. 2008. *Embodied Music Cognition and Mediation Technology*. MIT Press, Cambridge, MA.
J. London. 2000. Musical Expression and Musical Meaning in Context. In *6th International Conference on Music Perception and Cognition (ICMPC 2000)*. Keele, UK. http://www.people.carleton.edu/~jlondon/musical_expression_and_mus.htm.
J. Madsen, B.S. Jensen, and J. Larsen. 2014. Modeling Temporal Structure in Music for Emotion Prediction Using Pairwise Comparisons. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*. Taipei, Taiwan.

L.B. Meyer. 1956. *Emotion and Meaning in Music*. Chicago University Press, Chicago, IL.
A. Moles. 1966. *Information Theory and Aesthetic Perception*. University of Illinois Press, Urbana, IL.
M. Müller. 2015. *Fundamentals of Music Processing. Audio, Analysis, Algorithms, Applications*. Springer Verlag, Berlin.
E. Narmour. 1992. *The Analysis and Cognition of Melodic Complexity: The Implication-Realization Model*. University of Chicago Press, Chicago, IL.
J. Nika and M. Chemillier. 2012. Improtek: Integrating Harmonic Controls into Improvisation in the Filiation of OMax. In *Proceedings of International Computer Music Conference (ICMC 2012)*. Ljubljana, Slovenia.
F. Pachet. 2003. The Continuator: Musical Interaction with Style. *Journal of New Music Research* 32, 3 (2003), 333–341.
C. Palmer. 1997. Music Performance. *Annual Review of Psychology* 48, 1 (1997), 115–138.
A. Papadopoulos, F. Pachet, P. Roy, and J. Sakellariou. 2015. Exact Sampling for Regular and Markov Constraints with Belief Propagation. In *Proceedings of the 21st International Conference on Principles and Practice of Constraint Programming (CP 2015)*. Cork, Ireland.
A. Patel. 2008. *Music, Language and the Brain*. Oxford University Press, Oxford, UK.
J. Paulus, M. Müller, and A. Klapuri. 2010. State of the Art Report: Audio-based Music Structure Analysis. In *11th International Society for Music Information Retrieval Conference (ISMIR 2010)*. Utrecht, The Netherlands.
M. Pearce, M. Herrojo Ruiz, S. Kapasi, G. Wiggins, and J. Bhattacharya. 2010a. Unsupervised Statistical Learning Underpins Computational, Behavioural, and Neural Manifestations of Musical Expectation. *NeuroImage* 50, 1 (2010), 302–313.
M. Pearce, D. Müllensiefen, and G. Wiggins. 2010b. The Role of Expectation and Probabilistic Learning in Auditory Boundary Perception: A Model Comparison. *Perception* 39, 10 (2010), 1365–1391.
M. Pearce and G. Wiggins. 2012. Auditory Expectation: The Information Dynamics of Music Perception and Cognition. *Topics in Cognitive Science* 4 (2012), 625–652.
A. Porter, D. Bogdanov, R. Kaye, R. Tsukanov, and X. Serra. 2015. AcousticBrainz: A Community Platform for Gathering Music Information Obtained from Audio. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*. Malaga, Spain.
C. Raphael. 2010. Music Plus One and Machine Learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*. Haifa, Israel.
P. Roy and F. Pachet. 2013. Enforcing Meter in Finite-Length Markov Sequences. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI 2013)*. Bellevue, WA.
J.A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.
P. Russell. 1982. Relationships Between Judgements of the Complexity, Pleasingness and Interestingness of Music. *Current Psychological Research* 2 (1982), 195–202.
M. Schedl, A. Flexer, and J. Urbano. 2013. The Neglected User in Music Information Retrieval Research. *Journal of Intelligent Information Systems* 41, 3 (2013), 523–539.
J. Schlüter and R. Sonnleitner. 2012. Unsupervised Feature Learning for Speech and Music Detection in Radio Broadcasts. In *Proceedings of the 15th Intenational Conference on Digital Audio Effects (DAFx 2012)*. York, UK.
X. Serra. 2011. A Multicultural Approach in Music Information Research. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*

(ISMIR 2011). Miami, FL.
X. Serra, M. Magas, E. Benetos, M. Chudy, S. Dixon, A. Flexer, E. Gómez, F. Gouyon, P. Herrera, S. Jordà, O. Paytuvi, G. Peeters, J. Schlüter, H. Vinet, and G. Widmer. 2013. *Roadmap for Music Information ReSearch*. Creative Commons BY-NC-ND 3.0 license ISBN: 978-2-9540351-1-6. Available at http://mires.eecs.qmul.ac.uk.
S. Sigtia, N. Boulanger-Lewandowski, and S. Dixon. 2015. Audio Chord Recognition with a Hybrid Recurrent Neural Network. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*. Malaga, Spain.
Y. Song, S. Dixon, and M. Pearce. 2012. A Survey of Music Recommendation Systems and Future Perspectives. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR 2012)*. London, UK.
B.L. Sturm. 2014. A Simple Method to Determine if a Music Information Retrieval System is a Horse. *IEEE Transactions on Multimedia* 16, 6 (2014), 1636–1644.
D. Temperley. 2007. *Music and Probability*. MIT Press, Cambridge, MA.
R. E. Thayer. 1989. *The Biopsychology of Mood and Arousal*. Oxford University Press, New York, NY.
K. Ullrich, J. Schlüter, and T. Grill. 2014. Boundary Detection in Music Structure Analysis using Convolutional Neural Networks. In *Proceedings of the 15th International Society for Music Information Retrieval Conference (ISMIR 2014)*. Taipei, Taiwan.
Y. Vaizman, R. Granot, and G. Lanckriet. 2011. Modeling Dynamic Patterns for Emotional Content in Music. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*. Miami, FL.
O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. *Show and Tell: A Neural Image Caption Generator*. arxiv preprint arxiv:1411.4555 (2015).
M. Wertheimer. 1938. Laws of Organization in Perceptual Forms (Reprint). In *A Source Book of Gestalt Psychology*, W. D. Ellis (Ed.). Kegan Paul, Trench, Trübner & Company, London, pp. 71–88.
T. Weyde, S. Cottrell, J. Dykes, E. Benetos, D. Wolff, A. Kachkaev, S. Dixon, S. Hargreaves, M. Barthet, N. Gold, S. Abdallah, D. Tidhar, and M. Plumbley. 2015. The Digital Music Lab: A Big Data Infrastructure for Digital Musicology. In *16th International Society for Music Information Retrieval Conference (ISMIR 2015), Demos and Late Breaking News Session*. Malaga, Spain.
G. Widmer, S. Flossmann, and M. Grachten. 2009. YQX Plays Chopin. *AI Magazine* 30, 3 (2009), 35–48.
G. Widmer and W. Goebl. 2004. Computational Models of Expressive Music Performance: The State of the Art. *Journal of New Music Research* 33, 3 (2004), 203–216.
G. Wiggins, D. Müllensiefen, and M. Pearce. 2010. On the Non-Existence of Music: Why Music Theory is a Figment of the Imagination. *Musicae Scientiae* Discussion Forum 5 (2010), 231–255.
Y.-H. Yang and H. Chen. 2012. Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology* 3, 3 (2012), 40:1–30.
M. Zentner, D. Grandjean, and K. Scherer. 2008. Emotions Evoked by the Sound of Music. Characterization, Classification, and Measurement. *Emotion* 8, 4 (2008), 494–521.

Received October 2015; revised January 2016; accepted February 2016