Critical issues in evaluating freely improvising interactive music systems

Adam Linson, Chris Dobbyn and Robin Laney
Faculty of Mathematics, Computing and Technology
Department of Computing
The Open University
Milton Keynes UK
{a.linson, c.h.dobbyn, r.c.laney}@open.ac.uk

Abstract

As freely improvised music continues to be performed, it also continues to be implemented in interactive computer systems. For the scientific study of such systems to be possible, it is important to ensure the fitness for purpose of available evaluation methods. This paper will review several approaches to evaluating interactive computer music systems. It will also examine the uncritically-accepted assumption that quantitative evaluation invariably yields significant data, irrespective of context. Ultimately, it will be argued that, for some interactive computer systems, such as those designed for freely improvised music, qualitative evaluation by experts is the most appropriate evaluation method.

Introduction

Freely improvising computer systems, modelled on an established musical practice that has been called "non-idiomatic" (Bailey 1980/1993), have been around since at least the 1990s (see Rowe 1993; Lewis 1999). There has been a significant amount of academic writing on the topic, including a chapter in *Machine Musicianship* (Rowe 2001), and an entire book, *Hyperimprovisation* (Dean 2003), dedicated to the topic of its subtitle, "*computer-interactive sound improvisation*". As freely improvised music continues to be performed, it also continues to be implemented in interactive computer systems (see, for example, Blackwell and Young 2004; Hsu 2005; Collins 2006). For the scientific study of such systems to be possible, it is important to ensure the fitness for purpose of available evaluation methods.

A significant amount of research is conducted on dominant forms of instrumental and computer music, which has led to a number of evaluation methods and technologies. For example, music with well-defined style-based rules that constrain melodic, harmonic, and/or rhythmic constructs lends itself to generation and analysis techniques based on traditional musical notation. However, less widely studied forms of music such as freely improvised music have different evaluation criteria, and thus pose unique problems to widely adopted approaches to musicological and computational analysis. In particular, for music such as free improvisation, formalisable musical rules and symbolic notation fail to account for the fundamental aspects of the musical practice.

Defining the practice of freely improvised music is not trivial. As MacDonald, et al. (2011) point out, "while there is no generally accepted single definition of improvisation, most accounts highlight the spontaneously generated nature of the musical material and the real-time negotiation of unfolding musical interactions". This characterisation is explicitly extended to cover contemporary improvisation practices including free improvisation. In Clarke's examination of creativity in performance (2005a), he refers to an empirical study on freely improvised music showing that the "interweaving of social and structural factors" serves a central role in such music. (For those unfamiliar with freely improvised music, the artists and recordings mentioned, for example, in Bailey 1980/1993 and Smith and Dean 1997 may provide a useful starting point.)

In considering research into computer music systems that have been developed to perform freely improvised music, it is important to find an appropriate method of evaluation that is well-suited to the context. When computer music systems for free improvisation are assessed according to inappropriate criteria, it can have a potentially stifling effect on the development of new approaches to such systems, as well as potentially devaluing existing effective systems. This paper will review several approaches to evaluating interactive computer music systems. It will also examine the uncritically-accepted assumption that quantitative evaluation invariably yields significant data, irrespective of context. Ultimately, it will be argued that, for some interactive computer systems, such as those designed for freely improvised music, qualitative evaluation by experts is the most appropriate evaluation method.

Evaluation methods in computer music research

Computer music researchers generally acknowledge the need to determine an evaluation method appropriate to their specific research. It is not always apparent, however, to what extent these methods apply to other research in the wider field of computer music. Stowell, et al. (2009) consider a number of quantitative and qualitative approaches to evaluating "live human-computer music-making" (Stowell, et al. 2009), although they do not consider generative systems. Collins (2008), on the other hand, considers approaches to evaluating generative systems, finding promise in approaches that take into account the relationship of software to musical output. These authors find musical improvisation significant enough to merit acknowledgement, although they do not engage with the evaluation issues unique to "player-paradigm" interactive improvising systems, that is, systems with "a musical voice which may be related to, but is still in an audible way distinct from, the performance of a human partner" (Rowe 1996).

Notably, Stowell, et al. (2009) favour studies of expert performers for the evaluation of interactive digital musical instruments that are under performer control, although they do not mention a logical extension of this view, namely, that the same approach can be extended to interactive systems that are not under performer control. Similarly, Pearse and Wiggins (2001) find experts to be capable evaluators of music with enumerable rules (such as period harmonisations), but they do not address the evaluation of music without enumerable rules, such as freely improvised music. Among these researchers, there is a clear recognition that expert human analysis has something to offer, although pragmatic concerns lead to the consideration of alternatives to using human experts, especially computational approaches. But while computational approaches to evaluation can be expected to yield appropriate results in some research contexts, in others, computer-based evaluation techniques may be in principle incapable of discovering evidence that is relevant to the investigation. In support of this claim, Collins (2008) acknowledges that computational analysis, in failing to address emergent features of complex musical output, may have a destructive effect on the (musical) object of study.

A quantitative approach to evaluating improvised music

Pressing (1987), in a comprehensive study of quantitative analysis and improvisation, concludes that while "idiomatic" (Bailey 1980/1993) improvisation such as jazz lends itself well to both macro- and microstructural quantitative analysis, in freely improvised music "the musical meaning is not well described" by the same quantitative analytical approach. To clarify Pressing's terminology, "macroanalysis uses the full panoply of devices from traditional music theory" (primarily those generally found in musicological analysis of composed works), and microanalysis addresses parameters more likely to be found in perception studies of expressivity, such as "interonset and duration distributions", dynamic contours, and "legatoness". Pressing devised a specialised model to account for some of the general structural features of improvised music, which he validated in quantitative empirical studies. In further studies, he found that while his model functioned effectively for analysing improvised jazz music, it could not be effectively extended to freely improvised music without an arbitrary (and thereby subjective) partitioning of "polyphonically overlapping phrase structures". When comparing a jazz improvisation and a free improvisation—both subjectively regarded by Pressing as aesthetically successful—he found that the jazz improvisation contained extensive quantitative evidence of "micro-micro" *and* "micro-macro" correlations (which thus appears to validate his subjective assessment); in the free improvisation, both types of quantitative correlation were "nearly completely absent" (Pressing 1987). His findings suggest that even when quantitative analysis succeeds in apparently similar musical contexts, it is not trivial to extend such analysis to evaluating freely improvised music, whether human- or computer-generated.

Music and qualitative analysis

In some performances, musical features that are apparently insignificant become significant in the course of an analysis. This poses a difficulty for approaches to analysing data that screen for features whose relevance has been determined in advance. A computer-based quantitative analysis cannot overcome this problem, despite other strengths in detecting specific correlations, statistical significance, or other quantitative constructs such as self-similarity. Thus, computational quantitative analysis is limited in what it can discover; it is often confined to providing answers about whether or not a given data set complies to a given rule set.

While computers are a powerful tool to rapidly sift through vast amounts of data, computational analyses are notoriously bad at picking out long-term dependencies or large-scale structures from a body of time-series data, in contrast to human experts. Consider, for example, an analysis by composer, musician, and historian Gunther Schuller (1958) of three Sonny Rollins saxophone solos, all from within a single performance of Rollins' piece, "Blue 7" on the album "Saxophone Colossus" (Prestige LP 7079). Part of Schuller's argument for the merits of Rollins' solos includes the assessment that they are not merely following the fixed harmonic chord progression, nor are they merely variations on a melodic theme, nor do

they merely fulfil both of these (quantitatively measurable) criteria, both of which are typically used to determine that a jazz solo is (formally) allowable. Rather, Schuller points out the musical significance of a number of creative decisions made during the solos. His argument, based on his background expertise in the field, is fundamentally qualitative, and is nonetheless extensively backed up with (in some cases, quantitatively measurable) material evidence (i.e., musicological specificity about particular pitches, phrases, rhythms, etc.). The structural features he identifies in his analysis stand in sharp contrast to those that could be discovered by rule-based quantitative approaches: he identifies semantic information in the particular configuration of musical elements, thus extending his analysis beyond the quantitative measurement of compliance to musical rules. (Another example of this distinction can be found in Clarke's analysis of Jimi Hendrix's "Star Spangled Banner"; Clarke 2005b, Chapter 2.) Other quantitative approaches, such as conducting a survey of listeners' opinions, require large enough sample sizes to find statistical significance, and thus may be applicable when studying the capacities of a given listener population. By determining what is relevant to an analysis, a given analytic approach not only investigates but also characterises the object of study.

Schuller's (1958) expert analysis contains a wide range of assessments that illustrate the strengths of qualitative analysis over quantitative analysis, computational or otherwise. Take, for example, his assertion that a musical phrase introduced by Rollins "at first, seems gratuitous", whereas later in the piece, it "becomes apparent that [the phrase] was not at all gratuitous or a mere chance result, but part of an overall plan". Or, for example, the notion that the final restatement of an initial theme "is drained of all excess notes" and that the "rests [in the original statement of the theme] are filled out by long held notes," serving both to end the piece and "sum up all that came before". His analysis even briefly isolates the Max Roach drum solo, pointing out that two musical ideas, a triplet figure and a snare roll, are built up through permutations and alternations into a complex solo; then, eleven bars after the drum solo has ended, the drummer interestingly and meaningfully re-uses these two elements "in an accompanimental capacity". These examples can be viewed as arguments for the significance of specific musical decisions—what was chosen, or, in some cases, not chosen—among an allowable range of options. For instance, several possible notes may fit a given chord, but there may be a significance to the particular note that is played, such as a long-term dependency that is outside the scope of a computational analysis. Furthermore, a particular note may be chosen over another because of a social connotation, in principle irreducible to a quantitative framework.

In general, assessments of musical significance are relative to listener knowledge and expectation, as well as being strongly affected by listening context (for an extended discussion of this point, see Clarke 2005b). Furthermore, differences in listeners' accounts may extend beyond traditional musicology, and new concepts may be introduced that were not built into the initial evaluation framework. This is not possible when a computer has been limited in advance to a particular analytic framework. Also, in contrast to a quantitative approach, differing assessments of the same material need not contradict each other. In Clarke's Hendrix example, three listeners assess the significance of a particular arpeggiation: Clarke hears a destructive melodic rupture verging on dissolution, another hears the bugle of a military funeral, and yet another hears a pattern of fingerboard traversal. For Clarke's example, an imagined computational analysis would run the risk of shifting the framework of significance to only what can be discovered computationally, potentially excluding *a priori* the three listener assessments. It is difficult to imagine what a computational or other quantitative approach could contribute in this case, beyond support (confirm that it is an arpeggiation; identify the statistical likelihood for the presence and location of the arpeggiation within the melody; confirm its similarity to a given bugle call; investigate melodic possibilities constrained by fingerboard layout). And even if in principle computational analysis could discover *any* item of significance, the necessity of making prior decisions as to what counts as significant is a profound limitation.

Clarke's engagement with musical meaning finds support in the empirical listener perception studies conducted by Deliege, et al. (1997). These studies identify two primary types of perceived musical cues: those that can be confirmed by consulting the musical notation—"'objective' cues (themes, registral usages, etc.)"—and, in contrast, "'subjective' cues, which have psycho-dynamic functions (impressions, for example, of development, or of commencement) which may be experienced differently from one listener to another and are not necessarily identifiable in the score" (Deliege et al. 1997). This account of cues highlights specific, narrowly-defined observations (such as development and commencement), as opposed to the broader semantic framework of Clarke. But both accounts point to the fact that different listeners experience the same musical material in different ways, underscoring the fact that human listeners may be sensitive to information that could otherwise be obscured by more constrained assessments of the same material.

It is not currently possible to computationally model the entirety of human listening possibilities. Thus, when a particular research question is framed to empirically validate a computational model of human listening, the boundaries of listening are constrained, for example, to investigate melodic or harmonic expectations. But for research questions that seek, for example, to uncover the inherent polysemy of a given guitar solo, the diversity of embodied cultural expertise captured by multiple

qualitative accounts is no less scientific, and likely more relevant to the question at hand, than a quantitative study.

The role of experts

Expertise is not necessarily confined to an unworkably small set of specialists. With respect to Clarke's example, the ability to recognise a particular bugle call or guitar fingering can be considered forms of expertise that are shared by many. In practice, these recognitions eluded and thus enhanced his own musicologically astute account of melodic dissolution. Returning to the topic of improvisation, Smith and Dean, in their extensive investigation of improvisation in the arts, suggest that with an improvised work, "the possibility of finite interpretation is not to be expected, or even desirable," and "the ideas of improvisors themselves are very interesting sources for the analysis and understanding of improvisation" (Smith and Dean 1997). The substance of their study is found in the differing perspectives of practising improvisors who are regarded as experts. As Clarke (2005a) states, "the boundaries between the mundane, the creative, and the unacceptably idiosyncratic are constantly shifting, and [...] their position and evaluative significance is a function of judgements made within a shifting cultural and historical context". If we define experts as those with significant experience operating within the given cultural and historical context of a musical practice, it follows that such individuals are better equipped to make effective evaluations about the practice being studied. Especially in light of the aforementioned centrality of the "interweaving of social and structural factors" in freely improvised music, an experienced improvisor is well-suited to serve as an expert qualitative evaluator, capable of attunement to both subtle and complex emergent criteria.

Although some aspects of freely improvised music are amenable to various quantitative criteria (such as those that borrow from compositional analysis, especially melodic and harmonic information), the unique aspects of the music being studied do not necessarily reside in such criteria (see Lehmann and Kopiez 2010). To identify shared features across classical compositions by a single composer, a quantitative analysis would likely suffice, because the melodic and harmonic information comprise a significant degree of what constitutes the compositions. On the other hand, with freely improvised music, Smith and Dean (1997) find that "a multiplicity of semiotic frames can be continually merging and disrupting during a 'free' [...] improvisation," which they find to be an essential characteristic of such music. This represents at least one finding that is more effectively discovered by qualitative human expertise. Furthermore, in their elaborate taxonomy of improvisation, Smith and Dean refer to what they term "stipulated" improvisation, which describes a type of improvisation that derives structure and characteristic style from stipulated aesthetic parameters that are internalised by a community of performers. According to their account, the "stipulated" approach does not fully exploit improvisation because it does not permit the "breaking, remoulding and rebreaking of such 'parameters'", as does freely improvised music, which fundamentally allows for the possibility of "reformulating the parameters on each occasion" (Smith and Dean 1997). Thus, for some complex objects of study, expert qualitative analysis should be recognised as fulfilling an essential role that, at times, can be empirically supported by quantitative means, but never entirely replaced by these means.

Research context and conclusion

Quantitative approaches certainly have independently useful scientific functions (such as examining physical mechanics or features of perception). Yet expert qualitative analysis has the potential to offer a set of results that may, in fact, be more relevant to the particular research being conducted. Unfortunately, qualitative study is often assumed to diminish scientific rigour, despite the well-known criticisms of quantitative studies concerning test bias, determination of statistical significance, and assumptions implicit in classifications and standardised procedures (Hammersley 2009).

Generally speaking, for empirical study, the research question ought to be the determinant of experiment design and evaluation. Among the varieties of computer music research, there are some computer music systems that are not interactive, such as systems designed to output rule-based compositions. In many of these cases, quantitative computational analysis may be the most practical approach to evaluating whether or not a given computer system is successful in achieving its aims, such as rule compliance. When listener surveys are used to evaluate system success, it may be appropriate to use discrimination tests of fixed musical material (for more on discrimination tests, see Ariza 2009). For computer systems that generate widely divergent musical material, studies that focus on the underlying software may offer results more relevant to some research questions (Collins 2008). Alternatively, for studies of interactive computer music systems, the human-computer interaction, rather than the music, may be at the centre of the research. For these studies, the relation between performer intention and system responsiveness is one area of investigation that benefits from both quantitative and qualitative study, such as looking into actual and perceived timing issues (Stowell, et al. 2009). However, when considering interactive computer systems that are not under direct performer control, there is no well-established evaluation method that is widely recognised in the literature.

For some studies of such player-paradigm systems, the focus may be on idiomatic music, for which the evaluation

approaches mentioned for generative composition systems are found to be applicable (Pachet 2002). But for studies of *interaction experience* with player-paradigm systems, it is essential to use expert qualitative analysis to avoid the danger of "measurement that fails to ensure that the assumptions built into measurement procedures correspond to the structure of the phenomena being investigated" (Hammersley 2009). It is a common aim of many studies of computer systems to iteratively improve a system based on assessments of its strengths and weaknesses. In the case of player-paradigm systems, expert qualitative evaluation can be used to identify even broadly defined—or potentially undefinable—weaknesses such as whether or not (and why) a human musical interaction with a system is, for example, "boring". Qualitative expert analysis in this context, though not widely acknowledged, is not entirely disregarded. For example, in Collins' brief account of a "free improvisation simulation" (2006), expert interview data is the primary source of evaluation.

It has been argued here that using qualitative data from experts is one way to approach the problem of evaluating a freely improvising computer music system. This approach is especially relevant for determining whether or not a player-paradigm system itself performs at the level of a human expert. Accounts of interaction experiences such as interview data can be collected, correlated, and analysed, with the aim of applying the data to improve the system. In practice, as part of a longer research program, qualitative data can function in the same manner as quantitative data: after identifying a system's strengths and weaknesses, a second iteration of the system can be built, and a follow-up study can determine what aims have been achieved. In this way, despite the predominance of quantitative evaluation in computer music, qualitative expert analysis can be a viable means of investigating phenomena, and qualitative studies can ultimately serve in making novel contributions to the research field.

References

Ariza, C. 2009. "The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems". *Computer Music Journal* 33:2, 48–70.

Bailey, D. 1980/1993. *Improvisation: its nature and practice in music*. Da Capo Press.

Blackwell, T. and M. Young. 2004. "Swarm granulator". *App. of Evolutionary Computing*: 399-408. Springer.

Clarke, E. 2005a. "Creativity in performance". *Musicae Scientiae* 9:1, 157-182.

Clarke, E. 2005b. *Ways of listening: An ecological approach to the perception of musical meaning*. Oxford University Press.

Collins, N. 2006. "Towards Autonomous Agents for Live Computer Music: Real-time Machine Listening and Interactive Music Systems". PhD Thesis. Centre for Science and Music, University of Cambridge.

Collins, N. 2008. "The analysis of generative music programs". *Organised Sound* 13, No. 3: 237-248.

Dean, R. T. 2003. *Hyperimprovisation: computer-interactive sound improvisation*. A-R Editions.

Deliège, I., M. Mélen, D. Stammers, and I. Cross. 1997. "Musical Schemata in Real-Time Listening to a Piece of Music". *Music Perception* 14(2): 117–60.

Hammersley, M. 2009. "Is social measurement possible, and is it necessary?" In: *The SAGE handbook of measurement*, Sage Publications Ltd.

Hsu, W. 2005. "Using timbre in a computer-based improvisation system". *Proc. of the ICMC*, Barcelona.

Lehmann, A. and R. Kopiez. 2010. "The difficulty of discerning between composed and improvised music". *Musicae Scientiae* 14: 113-129.

Lewis, G. 1999. "Interacting with latter-day musical automata". *Contemporary Music Review* 18:3, 99-112.

MacDonald, R., G. Wilson, and D. Miell. 2011. "Improvisation as a creative process within contemporary music". In: *Musical Imaginations: Multidisciplinary perspectives on creativity, performance and perception*, D. Hargeaves, D. Miell, and R. Macdonald, Eds. Oxford University Press.

Pachet, F. 2002. "The Continuator: Musical Interaction with Style" in *Proc. of the International Computer Music Conference* (ICMA), Gothenberg, Sweden.

Pearce, M. T., and G. A. Wiggins. 2001. "Towards a Framework for the Evaluation of Machine Compositions". In: *Proc. of the AISB'01 Sym. on Artificial Intelligence and Creativity in the Arts and Sciences*. Brighton: 22–32.

Pressing, J. 1987 "The Micro- and Macrostructural Design of Improvised Music". *Music Perception* 5(2): 133-172.

Rowe, R. 1993. *Interactive music systems: machine listening and composing*. MIT Press.

Rowe, R. 1996. "Incrementally Improving Interactive Music Systems." *Contemporary Music Review* 13(2): 47-62.

Rowe, R. 2001. *Machine Musicianship*. MIT Press.

Schuller, G. 1958. "Sonny Rollins and the challenge of thematic improvisation". *Jazz Review* 1, No. 11: 6-9.

Smith, H. and R. T. Dean. 1997. *Improvisation, Hypermedia and the Arts since 1945*. Routledge.

Stowell, D., A. Robertson, N. Bryan-Kinns, and M. D. Plumbley. 2009. "Evaluation of live human-computer music-making: Quantitative and qualitative approaches". *Int. J. of Human-Computer Studies* 67, no. 11: 960-975.

Linson, A., Dobbyn, C., Laney, R.: Critical issues in evaluating freely improvising interactive music systems. In: Maher, M., Hammond, K., Pease, A., PÃ©rez, R., Ventura, D., Wiggins, G. (eds.) Proceedings of the Third International Conference on Computational Creativity, pp. 145â€“149 (2012b)Google Scholar. 13. Monson, I.: Doubleness and jazz improvisation: Irony, parody, and ethnomusicology.Â Linson A., Dobbyn C., Laney R. (2013) A Parsimonious Cognitive Architecture for Human-Computer Interactive Musical Free Improvisation. In: Chella A., Pirrone R., Sorbello R., JÃ³hannsdÃ³ttir K. (eds) Biologically Inspired Cognitive Architectures 2012. Advances in Intelligent Systems and Computing, vol 196. Springer, Berlin, Heidelberg. Musical improvisation is usually defined as the composition of music while simultaneously singing or playing an instrument. In other words, the art of improvisation can be understood as composing music "on the fly". There have been previous experiments by Charles Limb, using functional magnetic resonance imaging, that show the brain activity during musical improvisation.[7] Limb was able to show an increased activity in the medial prefrontal cortex, which is an area associated with an increase in self-expression. Further, there was decreased activity in the lateral prefrontal cortex, In the case of improvising music systems, user study and evaluation of a systemâ€™s ability to improvise may be useful in the ethnomusicological study of musical interaction in contemporary improvised music.Â Placing interactive system evaluation in the context of the ongoing debate on the efficacy of the Turing test, Ariza makes the important observation that such studies must prioritize the perspective of the human subject who interacts with such systems, emphasizing that this perspective differs greatly from the removed vantage point of the audience, academic, or domain-expert (Ariza 2009).Â Results of this study will not be discussed in full here, but some methodological issues of evaluating such systems are clarified, issues