Evaluating Musical Metacreation in a Live Performance Context

Arne Eigenfeldt
Contemporary Arts
Simon Fraser University
Vancouver, BC CANADA
arne_e@sfu.ca

Adam Burnett
Cognitive Science
Simon Fraser University
Burnaby, BC CANADA
ajb14@sfu.ca

Philippe Pasquier
Interactive Arts and Technology
Simon Fraser University
Surrey, BC CANADA
pasquier@sfu.ca

Abstract

We present an evaluation study of several musical metacreations. An audience that attended a public concert of music performed by string quartet, percussion, and Disklavier was asked to participate in a study to determine its success: 46 complete surveys were returned. Ten compositions, by two composer/programmers, were created by five different software systems. For purposes of validation, two of these works were human-composed, while a third was computer-assisted: the audience was not informed which compositions were human-composed. We briefly discuss the different systems, and present the artistic intent of each work, the methodology used in gathering audience responses, and the interpreted results of our analyses.

Introduction

The Musical Metacreation project[1] is an ongoing research collaboration between scientists and composer/musicians at Simon Fraser University that explores the theory and practice of metacreation – the notion of developing software that demonstrates creative behaviour (Whitelaw 2004). The objectives include not only developing software, but producing and presenting artistic works that use the software, and validating their musical success.

The research team includes a composer of acoustic and electroacoustic music who has created music composition and performance systems for over twenty years, an artificial intelligence researcher whose specialty includes multi-agent systems and cognitive modeling (and who is himself a creative artist in the field of computer music, sound design, audio and media arts), and several research assistants who are composers and/or scientists.

The fields of musical metacreation revolves around two central tasks:

- The composition task: the aim of this task is to produce music in the form of a symbolic representation, often a musical score. If the system takes existing compositions as input, it will be said to be corpus-based.
- The interpretation task: given some symbolic musical notation, this task consists of generating an acoustic signal.

Sometimes, these two tasks collide. For example, in electroacoustic music (in which we include electronica), an acoustic signal is directly generated as the output of the composition task. In the case of improvised music, composition and interpretation can be seen to happen simultaneously. The systems described in this paper, along with their evaluation, are all addressing the composition task.

The creative systems produced by our research team have already been described in conference proceedings and journals, while the music produced has been presented in public concerts and festivals. On the surface, therefore, we could state that our work has already been validated; however, there are deeper issues involved that we discuss in this paper.

In considering how a metacreative system might be validated, there are at least five potential viewpoints that can be considered:

1. The designer: the designer of the system accepts the output as artistically valid;
2. The audience: the work is presented publicly, and the audience accepts the work;
3. The academic experts: the system is described in a technical peer-reviewed paper and accepted for conference or journal publication;
4. The domain experts: the system receives critical attention through the media or non-academic artists via demonstration;
5. Controlled experiments: the system is validated through scientifically accepted empirical methods, using statistical analysis of the results in order to accept or reject the hypothesis made about the system.

In the first instance, any artwork created by a human, and publicly presented, conceivably requires the artist to consider it complete and successful and representative of the artist's aesthetic vision. Similarly, metacreative works have, so far and to our knowledge, reflected the artistic sentiment of their designers. According to this viewpoint, the system evaluation is made directly by the designer. In our case, our metacreative systems have produced works that we find artistically interesting.

The second step reflects an artist's desire to share their work with the public. Whether the audience accepts, appreciates, or enjoys the work is, unfortunately, often difficult to ascertain, as many audiences will politely applaud any work. One could include more quantitative measures, such as audience counts, album sales, or online downloads.

The third case involves peer-review, albeit for a description of the system in technical terms. A different criteria is in place, one dependent less upon the artistic output, and more upon the technical contribution of the system in its novelty and usefulness. Often, the evaluation is also an evaluation of the originality and soundness of the process encoded in the system in regard to the computational creativity literature (Colton, 2008).

[1] http://www.metacreation.net/

Both metacreation software and their output can be discussed in the media. Journalists and critics are different from the regular audience, in that their opinion will be further diffused to the audience: this may influence the audience judgment and the work can gain or lose notoriety as a consequence.

Lastly, empirical quantitative or qualitative validation studies can be undertaken that involve methods long supported by the research community for generating knowledge within the hard and soft sciences. While the computational creativity literature has started investigating these (Pearce and Wiggins, 2001; Pease et al., 2001, Ritchie, 2007, Jordanous, 2011), a great deal remains to be done.

While most previous work regarding the evaluation of musical metacreation (and computationally creative software in general for that matter) have been focusing on dimensions 1, 3 and 5, this paper presents an experimental study realized in the context of the public presentation of artworks in a concert setting (mixing dimensions 2 and 5). Also, there are very few instances of evaluation studies that consider more than one metacreative system at a time; our study is a comparative study of five different systems for computer-generated or computer-assisted composition.

The remainder of this paper discusses our evaluation study and the results we received, but also the questions that were raised. We first describe the different software systems involved, as well as the artistic intent of the compositions produced. We then present the methodology used in gathering audience responses to the compositions, as well as the results garnered from these responses. Finally, we posit our conclusions, as well as potential future work in this area.

Description

The public presentation of the metacreative software systems described in this paper took place as a public concert in December, 2011. Audience included members of the general public, as well some students of the first and third authors. Ten compositions, by two composers, were performed by a professional string quartet, percussionist, and Disklavier (a mechanized piano equipped to interpret MIDI input). The music was produced by five different software systems designed, and coded individually by the two composers. For comparison purposes, two of the pieces were composed without software; in other words, composed completely by human; a third was computer-assisted. The audience was informed beforehand that at least two of the works were human-composed, but were not informed as to which pieces these were; however, the program notes made it rather obvious that *fundatio* and *experiri* were, at most, computer-assisted. See Table 1 for a list of compositions.

The Systems and Compositions

In Equilibrio was generated by a real-time multi-agent system, described in (Eigenfeldt, 2009b). The system is concerned with agent interaction and negotiation towards a integrated melodic, harmonic, and rhythmic framework; its final output are MIDI events. The generated MIDI data was sent to a Yamaha Disklavier; no effort was made to disguise the fact that the performance was by a mechanical musical instrument. Along with the Disklavier and some high-level performance control by the composer, this system was responsible for both the "live" composition and its interpretation.

One of the Above consists of three movements for solo percussion. The music is notated by a system described in (Eigenfeldt and Pasquier, 2012). This system uses multiple evolutionary algorithms, including genetic algorithms, to control how a population of musical motives is presented in time, and how it is combined with other populations of motives. Intended for solo percussionist, the composition is a concentrated investigation in development of rhythmic motives. Each movement of the composition was presented separately, and treated as a unique composition within the evaluation. One additional movement, composed with the same intentions as the other three in this series, is human-composed (for reasons discussed in the Evaluation section).

Dead Slow / Look Left is a notated composition for string quartet and percussion, by a system that employs the harmonic generation algorithm described in (Eigenfeldt and Pasquier, 2010). The composition consists of a continuous overlapping harmonic progression generated using a harmonic analysis of 87 compositions by Pat Metheny, and a third-order Markov model based upon this analysis. In this corpus-based system, durations, dynamics, playing style, range, and harmonic spread were determined using patterns generated by a genetic algorithm. These continuous harmonies were interrupted by contrapuntal sections that interpret tendency masks (Truax 1991), which define such parameters as sequence length, number of instruments, subdivisions, playing style, number of playing styles, dynamics, and the number of gestures in a section.

Other, Previously was generated by a system described generally in (Eigenfeldt, 2009a), while the composition is described more fully in (Eigenfeldt, 2012b). A corpus of MIDI files – in this case 16 measures of the traditional Javanese ensemble composition *Ladrang Wilugeng* – was analysed, and generative rules regarding rhythmic construction was derived from the corpus. These rules were used by a genetic operator to create a population of ever-evolving melodies and rhythms that the system reassembled in a multi-agent environment over a rotating harmonic field. The real-time output was transcribed in a music notation program, and performed by string quartet. The end result is a piece of notated music that reflects many of the tendencies of the original corpus material, without direct quotation. The composer's role was limited to dynamic markings, orchestration, and assembling sections.

Gradual was generated by an extension of the system used to generate *One of the Above*, with an additional module to control pitch aspects integrated into the system. The final output was a notated work for marimba, violin, and Disklavier. While the system achieved the composition on its own, the interpretation was mixed: humans were playing the marimba and violin while the system was in charge of operating the Disklavier.

	Composition	Instrumentation	Experience Level		
			Expert	Novice	Combined
1	*In Equilibrio* [*c*]	Disklavier	3.17 (0.99)	2.71 (1.23)	2.90 (1.14)
2	*One of the Above #1* [*h*]	Solo percussion	4.00 (1.00)	3.36 (1.19)	3.67 (1.13)
3	*Dead Slow /Look Left* [*c*]	String quartet and percussion	4.16 (0.90)	3.08 (1.15)	3.51 (1.16)
4	*One of the Above #2* [*c*]	Solo percussion	3.68 (0.67)	3.16 (1.07)	3.42 (0.93)
5	*fundatio* [*h*]	String quartet	4.29 (0.80)	4.24 (0.83)	4.24 (0.81)
6	*experiri* [*c-a*]	String quartet	4.47 (0.61)	4.36 (0.86)	4.40 (0.76)
7	*One of the Above #3* [*c*]	Solo percussion	3.39 (0.76)	3.12 (1.20)	3.22 (1.04)
8	*Other, Previously* [*c*]	String quartet	4.31 (0.75)	4.50 (0.59)	4.40 (0.66)
9	*One of the Above #4* [*c*]	Solo percussion	3.63 (1.16)	2.71 (1.00)	3.10 (1.16)
10	*Gradual* [*c*]	Violin, marimba, Disklavier	4.05 (0.85)	3.88 (0.95)	3.93 (0.89)

Table 1. Individual composition engagement score means (out of 5). Standard deviations appear in parentheses. [c] = computer- composed. [h] = human-composed. [c-a] = computer-assisted.

fundatio and *experiri* were created by composer and software designer James Maxwell, with the help of his generative composition software that rests on a cognitive model of music learning and production. This software, *ManuScore*, is partially described in (Maxwell et al. 2009, 2011). *ManuScore* is a notation-based, interactive music composition environment. It is not a purely generative system, but rather a system which allows the composer to load a corpus, and proceed with that compositional process while enjoying recommendations from the system of possible continuations as suggested by the model.

fundatio was written using the commercial music notation software, Sibelius, following the compositional process used by the composer for many years, while *experiri* was written using *ManuScore*. Although this latter work remains clearly human-composed, the formal development of the music, and much of the melodic material used, were both directly influenced by the software.

Performances of the compositions can be viewed here:
In Equilibrio: http://youtu.be/x5fIdHbqEhY
Other; Previously: http://youtu.be/gaQfyhOiRio
One of the Above #2: http://youtu.be/gAIjQOiMG54;
One of the Above #3: http://youtu.be/bUYr7T7DKGs;
One of the Above #4: http://youtu.be/cQNQKinbJ-s.
Gradual: http://youtu.be/HZ2_Pr35KyU.
experiri: http://youtu.be/Gr5E7UVUoE8
fundatio: http://youtu.be/rNXt8b-kLMQ

Evaluation Study

The public concert was meant to serve two purposes: firstly, to present the artworks of the metacreative systems to the public, and secondly, to explore the idea of conducting evaluation in concert settings.

The opportunity for serious validation prompted the first composer to write an additional work separate from the metacreative systems, with the same musical goal. The purpose was not to fool the audience in making them guess which piece was not composed by machine, but rather to add human-composed material to the comparative study. While we hope that audiences will, one day, accept machine generated music without bias, Moffat and Kelly (2006) suggest this is not yet occurring. In our case, given three works for solo percussionist, composed in a particularly modernist style, it would be difficult to ascertain whether an audience's appreciation – or lack thereof – was due to the musical style, the restricted timbral palette, the lack of melodic and harmonic material, or any failings of the metacreative system. The human-composed piece allowed the composer to demonstrate the above-mentioned aspects, yet composed by the system designer. If the audience's rating of the human-composed piece was statistically similar to the metacreative works, it would demonstrate that the audience's preferences were based upon style, rather than musical creativity and/or quality.

Methods

Participants were 46 audience members from the general public (rather than only students) who attended a paid concert put on by Simon Fraser University. A program distributed to each audience member explicitly indicated that "machine-composed and machine-assisted musical compositions" would be performed. Each audience member also received an evaluation card on which they were encouraged to provide feedback. Audience members were asked to indicate, on a Likert-scale from 1 to 5, their level of familiarity with contemporary music, followed by ten similar 5 point Likert-scales regarding how "engaging" they found each piece to be. Audience members were also asked to indicate which three pieces they felt were the most directly human-composed. Audience members where also given space to write in their own comments. See Table 1.

Hypotheses

We hypothesized that the machine-generated and computer-assisted works were sufficiently similar in quality and style to the human-composed pieces that audience members would show no preference for the timbrally similar human-composed pieces (null hypothesis). This preference would be indicated by audience members' indication of how "engaging" they found each piece.

Analysis

In order to avoid alpha inflation that arises from multiple comparisons, statistical tests were made using post-hoc Bonferroni corrected alpha levels of .005 (0.5/10). For part of the analysis, the 46 audience members were divided into novice and expert groups depending on the score they indicated for the "familiarity with contemporary music" question. The "novice" group consisted of audience members that gave a score 1, 2, or 3 out of 5 on the familiarity scale (N = 25). The "expert" group consisted of the remaining audience members who gave a 4 or 5 (N = 19). Two audience members failed to provide a familiarity score, so their data was excluded from group comparisons.

Audience did not seem to discriminate between all the percussion pieces. Comparing the average engagement scores for the human-composed solo percussion piece *One of the Above #1* (M =3.59, SD = 1.15) with the average scores for the machine-composed *One of the Above #2* through #4 (M = 3.28, SD = 1.02) was not significant, t(44) = 1.43; p = .16 ns, leaving us unable to suggest that participants were able to discriminate between the human and machine-composed percussion pieces.

Audience did not "recognize" which piece was not computer-made. Assuming participants would find human-composed pieces more engaging, participants' engagement rating of the individual pieces were interpreted as an indication of whether participants could implicitly distinguish human-composed from machine-composed pieces. Tests comparing expert listeners' engagement scores for the human-composed *One of the Above #1* (M = 4.00, SD = 1.00) against the machine-composed alternatives (M = 3.57, SD = 0.88) were not significant (t(18)=1.68; p = 0.11 ns). Similarly, novice listeners' scores for *One of the Above #1* (M = 3.33, SD = 1.20) compared to the alternatives (M = 3.01, SD = 1.08) demonstrated no significant preference for the human-composed piece, t(23)=0.96; p = 0.34 ns.

Comparisons between the expert listener engagement ratings for the two string quartet pieces, the human-composed *fundatio* (M = 4.29, SD = 0.81) and the machine-assisted *experiri* (M = 4.47, SD = .61) were non-significant, t(18) = 1.00; p = .33 ns. Novice ratings for *fundatio* (M = 4.24, SD = 0.83) and *experiri* (M = 4.36, SD = 0.86) were similarly non-significant, t(24) = .72; p = .48 ns. This also failed to show that audience was discerning between the computer-assisted composition made using *ManuScore* and the human-made composition by the same composer.

Together, these results do not support the hypothesis that audience members were able to implicitly pick out which pieces were human-composed.

There was no difference between experts and novice choices. To determine whether audience members' ability to explicitly pick out the human-composed piece could depend on one's familiarity with contemporary music, a chi-square test compared novice and expert listeners' three "most directly human-composed" choices. The results of this test were non-significant, X2 (9, N = 113) = 14.17; p = .51 ns. This result fails to support the hypothesis that expert and novice listeners differ in their ability to explicitly discriminate human-composed pieces from machine-composed pieces.

Discussion

In addition to the above results, several further remarks can be made.

Overall, the evaluation results were pretty successful, showing both a rather high level of engagement from the audience, as well as good range with ranking means varying from 2.7/5 to 4.5/5. The audience did not discern computer composed from human-composed material, which seems to give credit to the five systems presented above. More precisely, this might just mean that the system were successful in portraying the goal, aesthetic and style of the two composers who developed them.

One further general observation that can be made is that while an evaluation in a concert setting allows us to capture the audience reaction to musical output in its "natural" presentation environment, it also introduces many variables that get us out of the usual controlled environment setting. The experimental protocol is also more difficult to follow.

On the other hand, controlled experiments are not the traditional setting in which a musical artwork is presented and this does introduce a number of biases in this type of evaluation. While these are well known, and solutions exist to circumvent them, our goal was to conduct an evaluation study in a live concert setting. We were concerned if conducting an evaluation in a concert setting would risk upsetting the audience's appreciation of the artwork. To our surprise, it did not seem to be the case, and the feedback forms were really welcomed. The whole process triggered a

longer than expected question and answer session at the end of the show. It is to be noted that very few audience members left before the end of the Q&A session.

Conclusions and Future Work

Finally, the whole process shed some light on the difficulty of evaluating computational creativity (and creativity in general). Artificial intelligence addresses the problem of emulating intelligence by having the computer achieve tasks that would require intelligence if achieved by humans. These tasks are usually formalized as well-formed problems. Rational problem solving is then evaluated by comparison to some optimal solution. If the optimal solution is theoretical and not attainable, optimization and approximation techniques can be used to get closer to the optimal, or at least improve the quality of the solution according to some metrics. Computational creativity is faced with the dilemma that, while creative behavior is intelligent behavior, such notions of optimality are not defined. It is often unclear which metrics need to be used to track progresses in the area. As demonstrated by this paper, it is at least an issue for the evaluation of composition systems.

Musical success is subjective in nature. This is why we resort to a comparative study capturing the relative level of success, rather than absolute ones. In the absence of formal metrics, we used human subjects to evaluate musical metacreation. However, creativity is a process (Boden, 2033). When evaluating a musical composition system, one particularly challenging aspect is that the system is capable of generating numerous pieces, with possibly varying levels of success: designing methodologies to measure that variability is an inherent challenge of the area. This is especially true when one has to use human subjects, since getting average relative evaluations of the average system production makes the experimental design particularly challenging.

To our knowledge, this paper is the first one to report on an evaluation experiment of machine-generated material conducted in real-world public situation. Beside the findings exposed above, the research instrument discussed here is a contribution in itself. As the systems presented are musical metacreations, validation and evaluation of such a system's output is itself a relatively novel and challenging research area. Our future work will continue to investigate and try to evaluate the methodologies to do so. Meanwhile, besides the finding exposed above, the paper raises a number of concerns and questions that will likely need further consideration in future work.

Acknowledgements

This research was funded by a grant from the Canada Council for the Arts, and the Natural Sciences and Engineering Research Council of Canada.

References

Boden, M. 2003. The Creative Mind - Myths and Mechanisms (2. ed.). Routledge I-XIII, 1-344

Colton, S. 2008. Creativity versus the perception of creativity in computational systems. In Creative Intelligent Systems: Papers from the AAAI Spring Symposium.

Eigenfeldt, A. 2009a. The Evolution of Evolutionary Software: Intelligent Rhythm Generation in Kinetic Engine. Applications of Evolutionary Computing, Berlin.

Eigenfeldt, A. 2009b. Multi-Agency and Realtime Composition: In Equilibrio. eContact 11.4 Toronto Electroacoustic Symposium 2009 http://cec.concordia.ca/econtact/11_4/

Eigenfeldt, A., Pasquier, P. 2010. Realtime Generation of Harmonic Progressions Using Constrained Markov Selection. Proceedings of the First International Conference on Computational Creativity, Lisbon.

Eigenfeldt, A., Pasquier, P. 2012a. Populations of Populations - Composing with Multiple Evolutionary Algorithms, P. Machado, J. Romero, and A. Carballal (Eds.): EvoMUSART 2012, LNCS 7247, 72–83. Springer, Heidelberg.

Eigenfeldt, A. 2012b. Corpus-based Recombinant Composition using a Genetic Algorithm. Soft Computing - A Fusion of Foundations, Methodologies and Applications. Springer Special issue on Evolutionary Music, forthcoming.

Jordanous, A. 2011. Evaluating Evaluation: Assessing Progress in Computational Creativity. Proceedings of the Second International Conference on Computational Creativity, Mexico City.

Maxwell, J., Pasquier, P. and Eigenfeldt, A. 2009. Hierarchical Sequential Memory for Music: A Cognitive Model. Proceedings of the International Society of Music Information Retrieval Conference, Kobe.

Maxwell, J., Pasquier, P. and Eigenfeldt, A. 2011. The Closure-based Cueing Model: Cognitively-Inspired Learning and Generation of Musical Sequences, Proceedings of the 8th Sound and Music Computing Conference, Padova.

Moffat, D., and Kelly, M. 2006. An investigation into people's bias against computational creativity in music composition. Third Joint Workshop on Computational Creativity, Riva del Garda.

Pearce, M. and Wiggins, G. 2001. Towards a framework for the evaluation of machine compositions. In Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in Arts and Science. Brighton: SSAISB. 22–32.

Pease, A., Winterstein, D., and Colton, S. 2001. Evaluating machine creativity. In Workshop on Creative Systems, 4th International Conference on Case Based Reasoning, Vancouver, 56–61.

Ritchie, G. 2007. Some Empirical Criteria for Attributing Creativity to a Computer Program. Minds & Machines. 17, 67–99.

Truax, B. 1991. Capturing musical knowledge in software systems. Interface. 20:3-4, 217–233.

Whitelaw, M. 2004. Metacreation. Art and Artificial Life. Cambridge, MA: MIT Press.