Towards A Framework for the Evaluation of Machine Compositions

Marcus Pearce and Geraint Wiggins
Department of Computing, City University, Northampton Square, London EC1V OHB
{m.t.pearce, geraint}@city.ac.uk

Abstract

We outline a framework within which machine compositions may be evaluated objectively. In particular, the framework allows statements about those compositions to be refuted on the basis of empirical experimentation. We consider this to be fundamental if we wish to evaluate the degree to which our programs achieve their compositional aims. Furthermore, a review of the literature reveals that this is a largely ignored aspect of research into algorithmic composition. Our framework involves four components: specifying the compositional aims; inducing a critic from a set of example musical phrases; composing music that satisfies the critic; and evaluating specific claims about the compositions in experiments using human subjects. We describe a system which exemplifies these four stages and which demonstrates the practicality of the framework. Finally, the application of the framework to the evaluation of musical creativity is discussed and directions for future research are suggested.

1 Introduction

Our concern in this paper is the evaluation of the music composed by computer programs. The crux of the problem is that Artificial Intelligence and the Cognitive Sciences (including cognitive musicology) are scientific disciplines following a methodology which attempts to evaluate theories objectively through empirical experimentation. However, the evaluation of beauty or aesthetic value in works of art (including music) often comes down to individual subjective opinion. This, as noted by Spector and Alpern (1994), "presents a problem for AI scientists wishing to produce computational artists." How might we reconcile the objectivity that scientific methodology seems to require with the apparent subjectivity inherent in aesthetic evaluation of artworks?

In fact, the problem of evaluating the music generated by systems for algorithmic composition is one that is typically given little attention in the literature. It is, however, fundamental that such systems be evaluated objectively on the basis of the music they compose. How else can we decide whether or not the machine composer succeeds in fulfilling the specified compositional goals?

Such evaluation is also important (in a wider sense) if we are to develop "progressive research programmes" (Lakatos, 1970) in the field of cognitive musicology. As noted by Desain et al. (1998) "a computational model is [not] an aim unto itself but a means to compare and communicate theories between different research communities." We consider a common means of evaluation to be fundamental if we are to judge musical theories from other communities in our research programme.

Therefore, it is our opinion that any program for the algorithmic composition of music (and indeed the generation of other works of art) should be embedded in a theoretical model that allows its output to be evaluated in objective terms. There are (at least) two ways in which machine composers might be evaluated: first, in terms of the music they compose; and second, on the basis of the manner in which they compose music (which may or may not be important depending on the aims of the research). We set out here to outline a theoretical framework for the former means of evaluation and to discuss its implications.

This paper is structured as follows. First, we consider notions of the demarcation between scientific and non-scientific knowledge and how this relates to the problem of evaluating machine compositions. In Section 3 we review previous work concerning the evaluation of machine compositions finding that the little work done fails to provide means of *objectively* evaluating computer generated compositions. Our framework for evaluation is presented and discussed in Section 4 while in Section 5 we describe a system which embodies the framework. This work provided important directions for future research which are discussed in Section 6. Finally, in Section 7, the relevance of this work to the evaluation of the creativity of programs for the algorithmic composition of music is considered.

It is important before we start to distinguish two different uses of the word "evaluation". First, a compositional system may evaluate its own compositions during various phases of the compositional process. We label this evaluation module the "critic". The second sense concerns the evaluation of the machine compositions as a means of scientifically gauging the degree to which the system generates music that fulfills the specified compositional aims. We call this process "evaluation."

2 Science and Music

In an attempt to distinguish propositions of the scientific disciplines from the non-scientific, Karl Popper developed the approach of methodological falsificationism. Scientific statements must be embedded in a framework within which experiments may be designed that will allow them to be refuted:

> "statements, or systems of statements, convey information about the empirical world only if they are capable of clashing with experience; or, more precisely, only if they can be *systematically tested*, that is to say, if they can be subjected ... to tests which *might* result in their refutation." (Popper, 1959)

Therefore, what distinguishes scientific from non-scientific statements is not formality or precision nor weight of positive evidence but simply whether it is possible to carry out an experiment which may refute that statement[1]. Although not without its critics, Popper's epistemology of science has been one of the most influential of the last century.

Cross (1998) has considered "the relevance and utility of science for our understanding of music." At one extreme he considers the "immanentist" position which holds that "music has no physical reality or locus but is constituted and inferred from the human capacity to endow with meaning the contingent phenomena of the material world and of human interaction." He notes that this position both denies "*all* science *any* efficacy in respect of music" and that it "seems to pervade current musicological thinking and writing." From this standpoint the quest to find an objective means of evaluating machine compositions would clearly be a futile task.

Proponents of the immanentist view consider that science is irrelevant to music theory because of the latter's interpreted, cultural and intentional nature. Cross (1998) argues that this implies a (mis)conception of scientific method as positivist, of scientific knowledge as general (culture independent) and the objects of scientific research being exclusively material. In contrast, he argues that a conception of science based on falsificationism (rather than positivism) can "dispose of many of the objections of the immanentists."

In particular the *sophisticated methodological falsificationism* of Lakatos (1970), introduces the notion of *research programmes* as the basic unit of scientific achievement (in place of isolated hypotheses). Sufficient weight of change in the background knowledge of such a programme may contribute to its succession or radical change. Since these research programmes consist partly of local background knowledge and heuristics for change they are not unsuitable for "cultural exegesis" (Cross, 1998). Furthermore, the requirement that the scientific evidence be observable "is no hindrance to its application to the intentional sphere, while [this account's] provisional and dynamic nature is not dissonant with the idea that 'there are no genuine absolutes"' (Cross, 1998). Finally, this account seems to characterise well the progress of science (Lakatos, 1970) and "is an increasingly popular view of change in scientific theories" (Brown, 1989).

So where does this leave us? It is clear that the field of cognitive musicology is in the early stages of its development and research programmes are still only in their infant years. The notion of evaluation by falsification of theories in the *protective belt* (Lakatos, 1970) of these programmes is crucial so as to build up a theoretical *hardcore* as these theories continue to go unrefuted. Only in this manner may we begin to build predictive and *progressive* research programmes within the field of cognitive musicology. The development of a framework for the objective evaluation of our models of musical composition is a small, but necessary, step in this endeavour.

3 Background

Clearly the means of evaluating the compositions generated by a machine will depend on the aims of the designer. For example, some systems are designed to compose music based on critical feedback from the user and in these cases "the acceptability of the final melodic material is entirely up to the user" (Ralley, 1995). There would seem to be no way of objectively evaluating the music composed by the program.

More objective evaluation is possible when, for example, the system is designed to compose music according to critical criteria derived from music theory or in the style of a composer. An example of the former approach is reported by Phon-Amnuaisuk et al. (1999) who developed a Genetic Algorithm (GA) for harmonising traditional chorale melodies. The harmonisations were evaluated by a senior university music lecturer according to the criteria used for examining first year undergraduate students' harmony. The latter approach is exemplified by the work of Hild et al. (1992) who developed a system which would harmonise in the style of J.S.Bach. The harmonisations produced by their system were evaluated by "music professionals" possessing expert knowledge of the composer's work.

However, the situation becomes much more complicated in situations where the program has a more specific musical goal than simply to compose something that the user likes or where a formal musical theory or expert knowledge is not available for evaluation purposes. The following is a brief review of previous approaches to the problem of evaluating machine compositions.

The vast majority of research into algorithmic composition gives the topic of evaluation short thrift, typically concluding with a sentence such as: "Almost all of the generated individuals were pleasant to listen to" (Johanson and Poli, 1998). Such subjective evaluation by the

[1] See Gould (1985), chapter 6, for an elegant demonstration of this thesis.

author(s) of the system is clearly unsatisfactory not only due to the bias and subjectivity involved but also due to the lack of an objective criterion for success.

An alternative approach seems to be inspired by normal modes of presenting music: that is to organise concerts and use audience feedback as a measure of success (Biles, 1999; Hild et al., 1992). This provides a measurable criterion for success and removes the bias of the developer of the system from the evaluation. It also attempts to reduce the problem of subjectivity by collecting many judgements.

However, while a well received performance would seem a good criterion for the evaluation of new works (as in the case of Biles), in the case of Hild et al. (1992) whose system was designed to harmonise in the style of J.S.Bach it is unsatisfactory. First, it is not clear that all of the audience will be evaluating the patterns on the basis of the same criteria: factors such as musical taste and knowledge of the genre (as well as an awareness that the compositions are machine composed) will have significant impact on the individual judgements made.

Other attempts to evaluate machine compositions have used criteria drawn from information theory. Conklin and Witten (1995), for example, employed a framework in which a context model was used to infer the probabilities of musical events in a body of Bach chorales given a preceding context. Witten et al. (1994) demonstrated that their prediction model showed "striking" similarity with the expectancies of human listeners and their conjecture is that a highly predictive theory, as measured by its *entropy*, will also be a good generative theory. However, Conklin and Witten (1995) finally resort to subjective evaluation of an example chorale generated by the system saying that it "seems to be reasonable."

A final possibility is to use formalised rules for the evaluation of machine compositions. Ames (1992) surveys a number of means for quantitatively assessing the "merit" of machine compositions. These may be used "to assess to what extent a choice (an option, a provisional solution or a final result) conform to a set of criteria set forth by a composer or analyst" (Ames, 1992).

Spector and Alpern (1994) have taken up this approach in an attempt "to separate those components of an AI system to which aesthetic judgements should apply from those to which scientific judgement should apply." They have developed a GP system which takes as parameters a critic (criteria defining the "fitness" of a composition) and a culture (a prior body of works). They argue that a number of critical criteria from opposing parties may be plugged into the system for any particular set of musical works. If the system succeeds in satisfying all these critics then it can be said to have succeeded overall.

However, such critical criteria may not be used for the objective evaluation of machine compositions since they would be tainted by the subjectivity of the programmer who designed them. Essentially, this model simply replaces the human critic in an IGA with a human critic's personal choice of formalised critical evaluation criteria. Furthermore, Spector and Alpern (1994) note of their system that while "the response pleases the critic, it does not please us (the authors) very well." It was on the basis of subjective considerations such as these that Spector and Alpern (1995) extended their framework to use a trained multi-layer perceptron critic. Ultimately then this approach returns to the subjective evaluation we are trying to escape from.

It is clear from this review that previous approaches have either failed to evaluate the music composed by the system or failed to do so in objective terms.

4 The Proposed Framework

4.1 Overview

The proposed framework for the algorithmic composition of music and evaluation of those compositions builds and improves on these previous approaches in two general ways. First, it provides a means of objectively evaluating the degree to which the music composed by the system succeeds in attaining the compositional goals. Second, it places no limitations on the types of computational methods used for the composition of music.

There are four essential elements in the framework: specifying the compositional aims; inducing a critic from a corpus of data; composing music which satisfies the critic; and evaluating the music composed by the system.

4.2 Aims

First, the aims of the researcher in developing a compositional system should be clearly stated. While this seems obvious it is often overlooked with researchers being vague about the goals of their research. This fact begs a deeper analysis of what exactly there is to be specified!

A general distinction can be made between those systems which are designed to compose within a particular genre of music or in the style of a particular composer and those which designed to allow the generation of new styles (essentially an artistic pursuit). Ames (1992) calls these "empirical style modelling" and "active style synthesis" respectively and our framework is designed with the former activity in mind. Given this general aim there still exists wide variety in the specific aims of researchers. Are we modelling a musical genre or the style of a particular composer? Are we dealing with entire compositions or compositional subcomponents (e.g., harmonisation, rhythmic development and so on)? How strictly do we want our system to adhere to the style being modelled? And many other issues which must be specified in detail as compositional aims of the research.

4.3 Inducing the Critic

In the second phase, a critic is induced from a set of patterns representing the relevant musical genre using some machine learning technique. In theory, any suitable computational techniques may be used for this - the appropriate methodology is likely to depend on the musical domain. The use of a particular technique should however be clearly justified in terms of the compositional and academic goals of the research.

This method is preferred due to the difficulty of generating a comprehensive set of rules for musical genres lacking a well developed formal theory[2](especially the problems of capturing all the exceptions to rules). An underspecified rule base will not only fail to describe the genre adequately but will also suffer from bias introduced by the selection of rules by the knowledge engineer (Conklin and Witten, 1995). Finally, the failure to include the necessary rule exceptions may lead to a lack of diversity or rigidity in the music composed.

When using machine learning techniques, however, there also exist several sources of potential bias. These include the selection of training data, the representation language used and the level of abstraction employed (Widmer, 2000). Therefore, "any musicological assumptions that influenced these choices must be made explicit, as they also determine what conclusions may be legitimately be drawn from the results of the experiments" (Widmer, 2000).

4.4 Composition

The third phase of the framework involves the generation of musical compositions which satisfy the critic. Once again any appropriate computational methods may, in principle, be used for this process. The mechanism for composition may be the same as that used to induce the critic in the case of, for example, a grammar. However, as in the case of the critic the choice of computational mechanism should be justified in terms of the compositional and academic goals and any music-theoretic assumptions made explicit.

4.5 Evaluation

Finally, the generated music can be evaluated by asking human subjects to distinguish compositions taken from the data set from those generated by the system. If the system composed pieces are misclassified as human composed with a frequency that may not be distinguished (statistically) from random selection we can conclude that the machine compositions are indistinguishable from human composed pieces. As will be seen in Section 5.4 similar experiments can be devised to evaluate the degree to which a system fulfills other compositional aims.

It will be clear that this experimental procedure bears a certain resemblance to the famous "imitation game" of Turing (1950). It is, however, worth noting several differences:

1. While the Turing test is designed to test for the presence of machine-thinking (intelligence/consciousness) our test simply determines the (non-)membership of a machine composition in a set of human composed pieces of music.
2. While the interrogator in the Turing test may interact with the machine, in our test the subjects are simply passive listeners: there is no interaction with the machine.

Therefore, our *discrimination test* is only analogous to the Turing test in that in both cases a behavioural test (rather than one which analyses the structure of the processes underlying behaviour) is used to decide whether a behaviour may be included in a set: the set of intelligent behaviours on the one hand and the set of musical pieces in a particular style on the other. We argue in Section 4.6 that this provides a very powerful test[3].

4.6 Why is the Framework Useful?

This framework has several attractive features. First, the critic (which determines the value of a composition internally within the system) is extracted from examples of the compositional genre using accepted computational methods rather than relying on human expertise to generate sets of rules. We are, in general, notoriously unreliable in formalising our expert knowledge.

Second, the final machine compositions are evaluated objectively within a *closed* system which provides no place for subjective evaluation of aesthetic merit. The system is intended to model a style of music (represented by its corpus of training examples) and its compositions are evaluated by comparison with exactly that set of examples from which its critical knowledge was extracted.

A third attractive feature is the use of experiments (which are integral to the framework) that will potentially allow claims about the compositional capabilities of the system to be refuted. Questions such as: "Is this music good?" are being turned into statements such as "People cannot distinguish the machine composed music from human composed music" which may be refuted through empirical experimentation. In effect, we have a framework within which statements of the type: "I can say with certainty that [the generated musical phrases] rival the carefully prepared demo sequences distributed with most drum machines!" (Horowitz, 1994) may be refuted on objective grounds.

[2]However, Spector and Alpern (1994) find working in a domain governed by formalised valuation criteria unsatisfactory for three reasons. First, the existing formalisations are often "dead forms" and therefore not suitable for the production of creative works. Second, they note that adherence to rules may not be a good indicator of aesthetic value. Finally, work with rules in one genre may not generalise well to other areas where critical criteria are not so uniformly accepted.

[3]The use of a Turing test as a procedure for evaluating machine generated music has been criticised by Marsden (2000).

It is worth noting that, although simple, the discrimination test described above is very powerful. In fact, the success of a piece of machine composed music on this test would mean that there are absolutely *no* perceivable features present or absent in the music which allow experts to identify it as being composed by a machine rather than a human composer. These features may be taken to include such elusive notions as aesthetic quality or perceivable creativity.

Finally, the framework is general in three respects: first, examples from any style/type/genre of music can be supplied as parameters[4] to the system; second, experiments can be devised to evaluate a range of compositional aims; and finally, it places no restrictions on the types of computational techniques used for the critic and the compositional modules.

5 A Preliminary Study

This section describes a system based on a genetic algorithm which embodies the framework outlined in Section 4. The four stages in the development of this system are described in turn (see Pearce, 2000, for full details of this research).

5.1 Aims

The compositional aims were to develop a system that would generate drum patterns conforming to the following criteria:

1. They should be in the style of "drum and bass" (henceforth d&b).
2. They should be comparable with human generated patterns in this style.
3. The composed patterns should show a certain amount of variation both within and between runs of the system.

5.2 The Critic

The critic consisted of a multilayer perceptron (MLP) trained on a set of positive and negative examples of this style. A MLP was chosen over and above other machine learning techniques due its capacity for generalisation and tolerance of noise and contradictory data (Toivianen, 2000). The former property was considered desirable due to the potential to allow a degree of flexibility in the critic and therefore greater diversity in the generated drum patterns. The latter capacity seemed appropriate since it seemed unlikely that d&b patterns could be easily described by any consistent set of rules.

The use of a trained MLP as the critic in evolutionary compositional systems has proved problematic in previous research (see Todd and Werner, 1999, for a recent review of evolutionary approaches to algorithmic composition). An attempt was made here to improve upon these approaches in two main areas: the selection of the positive and negative training data and the number of instances used to train the network (Pearce, 2000).

The network learned to classify the training data with a final RMS error of 0.1476 and a classification rate of 93% on the test set, demonstrating that its classification performance generalised well to unseen data.

5.3 Composition

A generational GA with probabilistic binary tournament selection was used to evolve drum patterns using the trained MLP as a critic. The system employed single point crossover within instruments and three mutation operators: one which changed a gene to a randomly selected value; one which rotated each instrument about a randomly selected quaver timestep; and one which reversed the entire chromosome.

It became apparent that the MLP was providing imprecise evaluation of the chromosomes. For example, due to the random initialisation of the chromosome far too many notes appeared on demisemiquaver subdivisions. However, the MLP still gave these chromosomes high fitness. An informal analysis of the network weights suggested that those corresponding to these timesteps tended to be small and therefore exerted little influence on the classification of a drum pattern. It is suggested that this was due to a failure to cover this aspect of drum patterns in the negative training data. The network was also imprecise in other areas and this is likely also to have been a consequence of the negative training set failing to cover a large enough area of the space of negative features of the style.

Although a more sophisticated initialisation of the chromosomes and the addition of four rules to the critic improved the quality of the generated drum patterns, the development of appropriate techniques for inducing critics in compositional systems from example musical pieces is an area that warrants further investigation (see Section 6.1).

5.4 Evaluation

5.4.1 Introduction

Three evaluation experiments were performed using the system compositions corresponding to the compositional aims set out in Section 5.1. The first was our discrimination test (section 5.4.2); the second asked subjects to classify the patterns according to style (section 5.4.3); and the final experiment asked for judgements of the diversity present in groups of three system generated patterns taken from both between and within runs (section 5.4.4).

The experiments were carried out using 19 human subjects from the School of Artificial Intelligence at Edinburgh University. All experiments were conducted in one session with all 19 subjects present in order to maintain extraneous influences constant across subjects. The ques-

[4] It could perhaps be extended to cover the machine generation of other types of artwork such as paintings or stories.

tions pertaining to experiments one and two were answered with respect to the same set of drum patterns in an attempt to reduce the amount of listening the subjects would have to do. As noted by Biles (1999), subjects find active listening and criticism of music an extremely tiring task. The subjects were asked to state on a scale of between nought and five their knowledge and experience of the musical styles involved.

The patterns used in the experiments were generated using the same system parameters. All MIDI drum parts, both human and system generated, were one bar in length and recorded at a tempo of 150 BPM using the GS Roland 909 drum set. It was explained to the subjects that all patterns (both human and system generated) were quantised and recorded using electronic drum sounds.

All three experiments involve testing hypotheses about means and due to the small sample sizes involved the t-test was used. In the case of a one-sample t-test N was calculated as the number of subjects minus one, while in the case of the two sample t-test it was calculated as the number of subjects minus two[5].

A general discussion of these experimental results can be found in Section 6.2.

5.4.2 Experiment 1

In this test the subjects were asked to discriminate system generated patterns and human generated patterns from the training set. The system was considered to have succeeded if the subjects were unable to distinguish system from human generated patterns.

A set of drum patterns was constructed containing 10 system generated patterns taken from different runs of the GA and 10 human generated patterns randomly selected from the MLP training set. These 20 patterns were played in a randomised order to the subjects who were asked to state for each pattern heard whether they thought it was system or human generated. Subjects were also asked to state at the end of the experiment on what basis they were discriminating.

The proportions of system and human generated patterns correctly classified were calculated from the obtained results and the following hypotheses tested with a one sample t-test against the known mean of 0.5 (that expected if subjects were discriminating randomly).

- Null hypothesis one: the mean proportion of human generated patterns correctly classified is the same as that expected if the subjects were answering at random.
- Null hypothesis two: the mean proportion of system generated patterns correctly classified is the same as that expected if the subjects were answering at random.

The results of this experiment are shown in Table 1 [6]

[5]For further reading Cohen (1995) is an excellent text on experimental methods in AI.

	Mean	SD	DF	t	p
Human	0.516	0.224	18	0.311	0.241
System	0.679	0.181	18	4.241	0.999

Table 1: Results of Experiment 1

The results provided two statistical results using 95% confidence intervals. First, we could retain null hypothesis one and second, we could reject null hypothesis two in favour of the following hypothesis:

- Hypothesis two: the sample mean proportion of system generated patterns correctly classified is greater than that expected if the subjects were answering at random.

This result allows us to refute the claim that the system generated patterns are indistinguishable from human generated patterns in the same style.

5.4.3 Experiment 2

This experiment was designed to evaluate whether the generated patterns were in the intended style by asking subjects to specify a style for system and human generated patterns. If the proportion of system generated patterns correctly classified according to style was equal to or greater than the proportion of human generated patterns correctly classified then the system generated patterns could be considered to be in the correct style.

A set of drum patterns was constructed containing 10 system generated patterns taken from different runs of the GA, 10 human generated patterns randomly selected from the ANN training set and 10 human generated "techno" drum patterns. Techno was chosen since it is a distinct musical style from d&b but typically has a similar, fast tempo. These 30 patterns were played in a randomised order to the subjects who were asked to state for each pattern heard the style of the pattern from a choice of "drum&bass", "techno" and "other".

The mean proportions of human and system generated patterns correctly classified according to style were calculated from the experimental data and the following hypothesis was tested with a two sample t-test. In the case of system generated patterns "correctly classified" refers to classification in the intended style (d&b). The option "other" was counted as an incorrect classification in all cases.

- Null hypothesis: there is no difference in the mean proportions of human and system generated patterns correctly classified according to style.

The results of this experiment are shown in Table 2.

[6]In this description of our results the *degrees of freedom* are denoted by "DF", the *standard deviation* is denoted by "SD", "t" is the t statistic and "p" is the probability that the sample means come from two populations whose true means differ.

Human Mean	System Mean	DF	t	p
0.729	0.568	17	2.181	0.978

Table 2: Results of Experiment 2: against system mean

Within a confidence interval of 95%, we could reject the null hypothesis in favour of the following hypothesis:

- Hypothesis one: the mean proportion of correctly classified human generated patterns is significantly higher than the mean number of system generated patterns.

Given this result a further one-sample t-test was run against the known mean 0.33 (the expected result assuming the subjects were answering at random) using the null hypothesis:

- Null hypothesis: the mean proportion of correctly classified system patterns is equal to the mean expected if subjects were answering at random.

The result of this test is given in Table 3.

System Mean	Known Mean	DF	t	p
0.568	0.333	18	3.474	0.999

Table 3: Results of Experiment 2: against known mean

We could, therefore, within a confidence interval of 0.99, reject the null hypothesis in favour of the following hypothesis:

- Hypothesis one: the mean proportion of correctly classified system generated patterns is greater than the proportion expected if the subjects were answering randomly.

These statistical results allow us to refute the proposal that the system generated patterns are in the intended style (Table 2) although they also suggest that the set of system generated patterns does overlap with the set of patterns in the style of d&b (Table 3).

5.4.4 Experiment 3

This experiment was designed to evaluate the amount of musical variation in the patterns generated both within one run and between runs of the GA compared to the amount of variation in the training data. Perceived variation was chosen as more musically relevant than an analysis of the patterns themselves (using Hamming distance, for example). An intermediate degree of variation was desired since too much would take the patterns out of the intended style. The variation in the training data was chosen as a reasonable indication of a desirable amount.

A set of drum patterns was constructed containing 20 groups of three patterns. Five of these groups of three were constructed from patterns taken from within individual runs of the GA, another five from patterns taken from different runs of the GA and the final ten from patterns randomly selected from the training set. Subjects were played these 20 groups of patterns in a randomised order and asked to indicate on a scale of one to five how much variation they considered there to be within each group. The total amount of variation for the human, the within-run and the between-run groups was calculated for each subject and converted to a fraction between nought and one by dividing it by the maximum possible score. The mean of these values across subjects was then collected.

The mean variation of the within-run and between-run groups was compared to the mean variation of the human groups in a two sample t-test with the following null hypotheses:

- Null hypothesis one: there is no difference between the mean perceived variation of the within-run groups and the human groups.
- Null hypothesis two: there is no difference between the mean perceived variation of the between-run groups and the human groups.

Table 4 shows the results for machine generated patterns taken from within runs of the system while Table 5 shows the results for those taken from different runs.

Human Mean	System Mean	DF	t	p
0.601	0.502	17	3.055	0.996

Table 4: Results of Experiment 3: Within Run

Human Mean	System Mean	DF	t	p
0.601	0.502	17	3.055	0.996

Table 5: Results of Experiment 3: Between Run

These statistical results showed that within a 99% confidence interval we could reject both null hypotheses in favour of the following hypotheses:

- Hypothesis one: the mean perceived variation of the human groups of patterns is greater than that of the within-run groups of system generated patterns.
- Hypothesis two: the mean perceived variation of the human groups of patterns is greater than that of the between-run groups of system generated patterns.

These results indicate that the system generated patterns fail to reach the criterion level of perceived variation. We have refuted the assertion that there are equal amounts of variation in the system generated patterns and the human generated patterns.

6 Future Directions

This research has demonstrated the practicality of the proposed framework and also highlighted several areas that are worthy of further development.

6.1 Inducing the Critic

The failure of this study to achieve its aims was attributed largely to problems with using a MLP to learn to classify musical sequences even when steps were taken to ensure that there was a sufficient amount of training data and that positive training data came from an internally consistent source. The major obstacle seems to be finding a set of negative training instances that will sufficiently cover the space of musical phrases not in the target classification. This is a serious problem and one that must be dealt with if this method is to be used in the composition of music.

Since the proposed framework is general, however, other machine learning techniques can be applied to induce a critic (see Papadopoulos and Wiggins, 1999, for a recent review of techniques for algorithmic composition). For example, there is a body of research concerning the use of recurrent MLPs for the generation of music (e.g., Todd and Loy, 1991; Griffith and Todd, 1999). In this paradigm, the recurrent network is trained to predict the note on a particular timestep given a previous sequence of notes as a context. However, an inability to extract higher level features of music seems to be a problem that has dogged most attempts to compose with recurrent neural networks. Mozer (1994) comments that:

> "While the local contours made sense, the pieces were not musically coherent, lacking thematic structure and having minimal phrase structure and rhythmic organisation."

One exception is HARMONET (Hild et al., 1992). The aim of this study was to approximate the function mapping chorale melodies onto their harmonisation using a training set of 400 four-part chorales composed by J.S.Bach. They approached the problem by decomposing it into sub-tasks: generating a skeleton structure of the harmony based on local context; generating a chord structure consistent with the harmonic skeleton; and finally adding ornamental quavers to the chord skeleton. Neural networks were used for the first and third tasks and a symbolic constraint satisfaction approach was applied to the second sub-task. The resulting harmonisations were judged by an audience of "music professionals" to be on the level of an improvising organist. The authors conclude that:

> "By using a hybrid approach we allow the networks to concentrate on musical essentials instead of on structural constraints which may be hard if learned by a network but easy if expressed symbolically."

While the networks in these compositional systems essentially perform the functions of both critic and composer in the above framework, they are still amenable to the evaluatory system. Furthermore, recurrent MLPs require no set of negative training instances.

Another possibility is to use unsupervised learning techniques which also require only positive data. Burton and Vladimirova (1997) used an unsupervised ART network to develop clusters corresponding to drum patterns from different styles of music (rock, funk, disco, latin and fusion) from a set of training examples. The fitness of candidate patterns generated by a GA was given by their propinquity to the desired cluster. However, the ART network critic seemed to produce a certain homogeneity in the generated patterns (Burton, 1998).

Alternatively, symbolic machine learning techniques might be used to extract a critic from a set of musical data. Typically, this has involved the use of one of two AI techniques to extract a musical theory from a corpus of musical examples. First, Markov models have been used to extract context based note transition probabilities from a corpus of data (e.g., Conklin and Witten, 1995). However, these approaches once again suffer from the problems of an inability to extract higher level structure in music. A second approach has been to extract grammars through statistical analysis of a set of musical pieces (e.g., Cope, 1991; Ponsford et al., 1999). Among the main drawbacks of these approaches are dealing with ambiguity and the potential to generate large numbers of strings of questionable quality (Papadopoulos and Wiggins, 1999).

The appropriate methods to use will depend crucially on the musical domain being modeled. However, we believe that an approach that applies different AI techniques to those critical and compositional subtasks to which they are best suited (as in HARMONET) is likely to prove most fruitful.

6.2 Experimental Design

The experiments performed to evaluate the drum patterns generated by the system proved inadequate in several respects. It is interesting to note that in Experiment 1 the subject's classification performance on the human generated patterns was no better than random. This suggests two things: that the subject's familiarity with the domain was low; and a bias towards classifying the patterns as system generated.

The first suggestion is supported by the the low average experience and knowledge of d&b professed by the subjects (two out of five) and also by the low mean proportion of human generated patterns correctly classified according to style in Experiment 2. The subject's self-professed lack of knowledge of the relevant musical genres made their judgements hard to evaluate. Ideally such experiments should be made with subjects who are highly familiar with the genre of music being composed by the system[7].

[7] Although the subjects must not be familiar with the human composed pieces used in the test

The second problem concerns the bias towards classifying drum patterns as system composed. Some reasons for this bias were suggested by an informal collection of the criteria used by the subjects to distinguish system and human generated patterns. It seemed that they were, in general, looking out for negative features[8] of the patterns which would classify them as system generated. A sense that they were being asked to "catch the system out" may have lead them to overclassify the patterns as system generated. Those subjects who were looking for features of human generated patterns searched for "smoothness", "coherency", "large scale structure", "subtleties" and such features as whether it qualified as part of a song or similarity to rhythms they had heard in songs. Given that the drum patterns were short, lacking musical context and in an unfamiliar style for most subjects, the use of these criteria may have lead to the bias towards classifying patterns as system generated.

Urwin (1997), in a similar experiment, asked subjects to assume that a pattern was human generated if they were unsure (and obtained 85% misclassification of the system generated patterns). However, this is likely to have produced a bias in the opposite direction. There are two obvious means of countering these kinds of biases. The first would be to use a control experiment in which subjects are given a set of human compositions. The proportion misclassified as machine generated could then be taken as a baseline to be factored into the statistical analysis of the actual experiments. A second solution would be to inform the subjects that the set of musical phrases contained equal proportions of machine and human generated compositions. An extension of this idea would be to present the subjects with a set of compositions only one of which is be machine generated. The task would then become to decide which composition has been composed by the machine.

A further possibility would be to set up the test in a manner more akin to the Turing test. A computer interface could be designed which presented two buttons, one of which would play compositions randomly selected from the training set while the other would play compositions randomly selected from the set of machine compositions. The subjects would have to decide which button corresponded to the system generated compositions. Statistics such as the number of times each button was pressed and so on could be collected for each subject.

Finally, a few points made by the subjects concerning the experiments are worth noting. First, it was suggested that the short duration of the patterns (just one bar) may have forced subjects to quick and unreliable decisions while the lack of musical context for the drum patterns made the evaluation difficult. Second, the merging of experiments one and two may have lead to unreliable decisions since subjects had to answer two different questions (relating to whether the pattern was system or human generated and what style it was in) about the same pattern. Once again, this may have forced hurried and unreliable responses from the subjects.

Therefore, some suggestions for better designed experiments would be to use separate experiments for each individual test, to use more knowledgeable subjects and to use longer patterns. Finally, the problem of the bias towards classifying patterns as system generated should be addressed.

6.3 What do the Results Mean?

The discrimination test by itself simply tells us whether the system generated patterns are perceptually distinguishable from human generated patterns in the same style. This tells us nothing about which subcomponents of the system and its behaviour are in need of further development. However, this information is very important if our research programmes are to be progressive as described in Section 2.

The other experiments described here were designed to be able to refute other specific claims about the drum patterns composed by the system. Experiment 2 would allow us to refute the claim that the patterns were in the intended style. However, since membership of a stylistic group is probably not a discrete concept, a better experiment might have asked for judgements of the *degree* to which the patterns were considered d&b patterns.

Experiment 3 would allow us to refute the claim that there existed as much perceptual diversity in the system generated patterns as in the human generated training set. Another experiment which asked subjects to distinguish system generated patterns from human generated examples of the style which were not included in the data set could also be used to test the claim that the knowledge possessed by the system was generalised to the style under consideration rather than reflecting only the training corpus.

It can be seen that experiments could be designed to test claims about many other aspects of the system generated patterns. For example, the output of creative systems may be evaluated not only in terms of set membership but also using qualitative measures. Therefore, an experiment asking for an aesthetic evaluation of a set of patterns containing machine and human composed music might be helpful in determining not only whether the system generated pieces are comparable to human composed pieces and in the correct style but also how "good" they are considered to be within the style. It would be interesting to see how much consensus there would be between subjects on such aesthetic matters.

So what do the results of these experiments mean? It should be noted that these experiments are not intended as replacements for the comments of musicians and musi-

[8] Examples of these features were lack of originality, randomness (or how chaotic the patterns seemed), predictability and mechanicality lead to classification as system generated. It is interesting to note that both extreme conformity to the prototype of a style and extreme randomness in a pattern classified it as system generated in the eyes (or ears) of the subjects.

cologists which may be extremely insightful and useful in terms of improving our computational models of composition. However, these experiments do allow us to make scientific (refutable) claims about the music generated by our compositional systems. Nevertheless, many questions remain. Are we justified in assuming that if a group of knowledgeable subjects misclassified 50% of the system generated patterns as human generated then they can be taken to be answering at random? Can the claim of indistinguishability be refuted by a single correct classification of a pattern as system generated?

7 Evaluating Musical Creativity

No mention has yet been made of musical creativity – does our framework have any relevance to the evaluation of the creativity of machine composers? The framework is designed for the evaluation of machine compositions within a specified style. It might therefore be objected that the really creative musical acts involve the founding of a new style or genre. However, as noted by Garnham (1994) most creative achievement in the arts does not follow this form: "the origins of the symphony are lost in history and its major triumphs are the work of composers who did not invent the basic symphonic form." Most creative work is carried out within styles or genres.

Creativity can be defined in two ways: what Boden (1990) calls the Psychological and Historical (P- and H-) forms of creativity. The former refers to the generation of a creative product that is novel for the individual while the latter indicates that something never before conceived of by mankind has been generated. Since H-creativity can be seen as a subset of P-creativity depending also on historical accident and social fashion (Boden, 1990), our concern here is with P-creativity.

How might we go about evaluating the P-creativity of our compositional system? There would seem to be two aspects of the system to be subjected to evaluation. First, the music composed and second, the internal workings of the system itself.

Regarding the former (which has been the major focus of this paper), we have argued that the system generated compositions will only succeed on our discrimination test if there are absolutely no perceivable features which can be used to distinguish the set of machine compositions from the set of human compositions. If it is possible to perceive creativity in music (or to *infer* the P-creativity of the composer) then this would be among these features.

In fact, the perceived creativity of a work of art or piece of music is likely to be closely related to its perceived aesthetic value and it is possible that this was considered by the subjects in their attempts to discriminate human and system generated patterns. This conjecture is supported to some degree by the comments of the subjects in the experiments described above: both extreme conformity to the prototype of a style and extreme randomness in a pattern as indicative that it had been machine generated. This suggests that *guided* exploration of the space of possible drum patterns was considered indicative of human composition. This, in turn, accords with the notion that creative products must be both original (p-novel) and "appropriate" (Boden, 1990).

The other experiments in the research described in Section 5 may also be pertinent here. The second experiment ensures that the patterns are in the correct style and therefore "appropriate". Finally, the third experiment was looking at the ability of the system to continually and thoroughly explore its the space of drum patterns in a non-repetitive manner. Similarly, we would expect creative individuals to consistently and continually generate creative products.

Other experiments could be devised along similar lines to probe other aspects of creative composition. For example, by obtaining judgements of the perceptual distance between pairs of training examples and training example/system generated pairs it would be possible to evaluate how far the program explores away from the experienced musical examples.

It might be suggested that evaluation of machine compositions themselves can tell us only so much about the creativity of a compositional system: we would want to know about the internal workings of the system (its compositional processes) before we called it creative. As noted by Boden (1990) this appears to be an important criterion by which people are reluctant to attribute creativity to machines. Furthermore, Cohen (1999) refuses to attribute Aaron (his program for the generation of artworks) with creativity although it generates pieces it has never painted before and has a unique and characteristic style. This is largely because he doesn't believe it is creating the paintings in the right way.

While it would seem important to complement behavioural evaluation of our creative systems with what we might call "cognitive" evaluation [9] our tests can show some light on the internal mechanics of the system. Hofstadter (1994) has argued that the premise that "*covert mechanisms* can be deeply probed and eventually revealed merely by means of watching *overt behaviour*... lies at the very heart of modern science." In particular, he argues that the Turing test offers a multitude of probes which may be used in long-term interaction with a cognitive model to infer the mechanisms underlying its behaviour.

To give an example, a system which stored samples from various songs and simply pasted them together to produce new compositions might pass the discrimination test initially. However, it would seem likely that over repeated experiments the underlying mechanisms of "composition" would be inferred by the subjects. This example emphasises two important features of the experiments: first, the criteria used by the subjects for evaluation are useful as pointers to the types of behaviour they identify as exposing non-human mechanisms in the compositional

[9] Although important this is a topic for another paper.

system[10]; second, the subjects should ideally be allowed to take the test repeatedly.

8 Summary and Conclusions

This paper has provided a tentative first step towards the development of a general framework for the evaluation of machine generated music by computer programs based on AI techniques. The evaluation of algorithmic compositions is an important issue since without it we have no means of telling whether the systems we develop succeed in their compositional aims and, if not, why not. This in turn is important if we are to develop progressive research programmes within the field of cognitive musicology. The issue of evaluation is also one that is frequently given less attention than it deserves in the literature on algorithmic composition.

The framework involves four stages: specification of compositional aims; induction of a critic from examples of the relevant musical genre; composition of music that satisfies the critic; and evaluation of the machine compositions using human subjects.

The framework has several attractive features, one being that it places no restrictions on the compositional aims, the music generated and the AI techniques used in the research. However, our presentation in Section 5 of a compositional system which embodies the framework demonstrates that is not so general as to be meaningless. In effect, it allows us to make refutable, and therefore scientific, claims about the degree to which a system fulfills its compositional aims. The refutation of these claims may allow us to identify areas in which are compositional models are lacking. Finally, the framework may be extended to evaluate the musical creativity of machine composers.

In addition we have highlighted several issues worthy of future work including the following:

- A reliable and appropriate means of inducing a critic from a body of music.
- Various issues concerning the experimental protocol used for evaluation including: the use of expert subjects; countering classification biases; and the presentation of the composed music in a natural context for evaluation by the subjects.
- The evaluation of musical creativity both in terms of the compositions produced and how the composition allow subjects to infer the underlying mechanisms of the system.
- Finally, the framework should be applied to systems with a wider range of compositional aims and the generation of different styles and types of music. It would also be interesting consider the application of the framework to the evaluation of other creative systems for the generation of, for example, visual art, stories and jokes.

[10] as noted in section 6.2

We expect to address these issues in future research.

Acknowledgements

We would like to thank Alan Smaill for useful and careful guidance during the course of parts of the research presented here. Thanks is also due to Dave Meredith for his stimulating discussion of the issues raised in this paper and to two anonymous reviewers for their useful comments on earlier drafts. This research has been supported by EPSRC via studentship numbers 99407250 and 00303840.

References

C. Ames. Quantifying musical merit. *Interface*, 21:53–93, 1992.

J.A. Biles. Life with GenJam: interacting with a musical GA. In *Proceedings of the 1999 IEEE Systems, Man and Cybernetics Conference*, Tokyo, 1999.

M.A. Boden. *The Creative Mind: Myths and Mechanisms*. Weidenfield and Nicholson, London, 1990.

R.T. Brown. Creativity: what are we to measure? In J.A. Glover, R.R. Ronning, and C.R. Reynolds, editors, *Handbook of Creativity*, pages 3–32. Plenum Press, New York, 1989.

A.R. Burton. *A hybrid neuro-genetic pattern evolution system applied to musical composition*. PhD thesis, University of Surrey, 1998.

A.R. Burton and T. Vladimirova. A genetic algorithm utilising neural network fitness evaluation for musical composition. In *Proceedings of the 1997 International Conference on Artificial Neural Networks and Genetic Algorithms*, pages 220–224, 1997.

H. Cohen. Colouring without seeing: a problem in machine creativity. *AISB Quarterly*, 102:26–35, 1999.

P. Cohen. *Empirical Methods for AI*. MIT Press, Cambridge, MA, 1995.

D. Conklin and I.H. Witten. Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24:51–73, 1995.

D. Cope. *Computers and Musical Style*. Oxford University Press, Oxford, UK, 1991.

I. Cross. Music and science: three views. *Revue Belge de Musicologie*, 52:207–214, 1998.

P. Desain, H. Honing, H. van Thienen, and L. Windsor. Computational modelling of music cognition: problem or solution. *Music Perception*, 16(1):151–166, 1998.

A. Garnham. Art for arts's sake. *Behavioural and Brain Sciences*, 17(3):543–544, 1994.

S. J. Gould. *The Flamingo's Smile*. W. W. Norton, New York and London, 1985.

N. Griffith and P. M. Todd. *Musical Networks: Parallel Distributed Perception and Performance*. MIT Press/Bradford Books, Cambridge, MA, 1999.

H. Hild, J. Feulner, and D. Menzel. HARMONET: a neural net for harmonizing chorals in the style of J.S. Bach. In R.P. Lippmann, J.E. Moody, and D.S. Touretzky, editors, *Advances in Neural Information Processing 4*. Morgan Kaufmann, San Francisco, CA, 1992.

D. Hofstadter. Creativity, brain mechanisms and the turing test. In D. Hofstadter, editor, *Fluid Concepts and Creative Analogies*, pages 467–491. Harper Collins, NY, 1994.

D. Horowitz. Generating rhythms with genetic algorithms. In *Proceedings of the 1994 International Computer Music Conference*, Aarhus, Denmark, 1994.

B. Johanson and R. Poli. GP-music: an interactive genetic programming system for music generation with automated fitness raters. In *Proceedings of the third International conference on Genetic Programming*, Cambridge, MA, 1998.

I. Lakatos. Falsification and the methodology of scientific research programmes. In I. Lakatos and A. Musgrave, editors, *Criticism and the Growth of Knowledge*, pages 91–196. CUP, Cambridge, UK, 1970.

A. Marsden. Music, intelligence and artificiality. In E. R. Miranda, editor, *Readings in Music and Artificial Intelligence*, pages 15–28. Harwood Academic Publishers, 2000.

M. C. Mozer. Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multi-scale processing. *Connection Science*, 6(2-3):247–280, 1994.

G. Papadopoulos and G. Wiggins. AI methods for algorithmic composition: a survey, a critical view and future prospects. In *Proceedings of the AISB'99 Symposium on Musical Creativity*, Edinburgh, UK, 1999.

M. T. Pearce. Generating rhythmic patterns: a combined neural and evolutionary approach. Master's thesis, Department of Artificial Intelligence, University of Edinburgh, Scotland, 2000. URL `http://www.soi.city.ac.uk/~ek735/msc/msc.html`.

S. Phon-Amnuaisuk, A. Tuson, and G. Wiggins. Evolving musical harmonisation. In *ICANNGA '99*, Slovenia, 1999.

D. Ponsford, G. Wiggins, and C. Mellish. Statistical learning of harmonic movement. *Computer Music Journal*, 28(2), 1999.

K. Popper. *The Logic of Scientific Discovery*. Hutchinson and Co., London, 1959.

D. Ralley. Genetic algorithms as a tool for melodic development. In *Proceedings of the International Computer Music Conference*, pages 501–502, 1995.

L. Spector and A. Alpern. Criticism, culture and the automatic generation of artworks. In *Proceedings of the Twelfth National Conference on Artificial Intelligence, AAAI-94*, pages 3–8, 1994.

L. Spector and A. Alpern. Induction and recapitulation of deep musical structure. In *Proceedings of the IFCAI-95 Workshop on Artificial Intelligence and Music*, pages 41–48, 1995.

P. M. Todd and D. G. Loy, editors. *Music and Connectionism*. MIT Press, Cambridge, MA, 1991.

P. M. Todd and G. Werner. Frankensteinian approaches to evolutionary music composition. In N. Griffith and P. M. Todd, editors, *Musical Networks: Parallel Distributed Perception and Performance*, pages 313–339. MIT Press/Bradford Books, Cambridge, MA, 1999.

P. Toivianen. Symbolic ai versus connecionism in music research. In E. R. Miranda, editor, *Readings in Music and Artificial Intelligence*. Harwood Academic Publishers, 2000.

A. M. Turing. Computing machinery and intelligence. *Mind*, 59:433–460, 1950.

R. Urwin. Connectionist methods for musical rhythm composition. Undergraduate thesis, Department of Artificial Intelligence, University of Edinburgh, Scotland, 1997.

G. Widmer. On the potential of machine learning for music research. In E. R. Miranda, editor, *Readings in Music and Artificial Intelligence*. Harwood Academic Publishers, 2000.

I. H. Witten, L. C. Manzara, and D. Conklin. Comparing human and computational models of music prediction. *Computer Music Journal*, 18(1):70–80, 1994.