Evaluating Cognitive Models of Musical Composition

Marcus T. Pearce and Geraint A. Wiggins
Centre for Cognition, Computation and Culture
Goldsmiths, University of London
New Cross, London SE14 5SG, UK
{m.pearce,g.wiggins}@gold.ac.uk

Abstract

We present a method for the evaluation of creative systems. We deploy a learning-based perceptual model of musical melodic listening in the generation of tonal melodies and evaluate its output quantitatively and objectively, using human judges. Then we show how the system can be enhanced by the application of mathematical methods over data supplied by the judges. The outcome to some extent addresses the criticisms of the experts. We suggest that this is a first step on the road to autonomously learning, introspective, creative systems.

1 Introduction

We examine, at the computational level, the demands of the melodic composition task, focusing on constraints placed on the representational primitives and the expressive power of the composition system. We use three multiple-feature Markov models trained on a corpus of chorale melodies to generate novel pitch structures for seven existing chorale melodies. We propose null hypotheses that each model is consistently capable of generating chorale melodies that are rated as equally successful examples of the style as the original chorale melodies in our dataset. To examine the hypotheses, experienced judges rated the generated melodies together with the original chorale melodies, using a variant of the Consensual Assessment Technique (Amabile, 1996) for investigating psychological components of human creativity. The results warrant rejection of the null hypothesis for all three of the systems. Even so, further analysis identifies some objective features of the chorale melodies that exhibit significant relationships with the ratings of stylistic success, suggesting how the computational models fail to meet intrinsic stylistic constraints of the genre. Adding new features to address these concerns significantly improves our systems' prediction performance.

We present our experiment and the evaluation method, which, we suggest, forms a basis for systems capable of introspection based on feedback on their output.

2 Background

2.1 Music Generation from Statistical Models

Conklin (2003) examines four methods of generating high-probability music according to a statistical model. The simplest is sequential random sampling: an event is sampled from the estimated event distribution at each sequential position up to a given length. Events are generated in a random walk, so there is a danger of straying into local minima in the space of possible compositions. Even so, most statistical generation of music uses this method.

The Hidden Markov Model (HMM) addresses these problems; it generates observed events from hidden states (Rabiner, 1989). An HMM is trained by adjusting the probabilities conditioning the initial hidden state, the transitions between hidden states and the emission of observed events from hidden states, so as to maximise the probability of a training set of observed sequences. A trained HMM can be used to estimate the probability of an observed sequence of events and to find the most probable sequence of hidden states given an observed sequence of events. This can be achieved efficiently for a first-order HMM using the Viterbi algorithm; a similar algorithm exists for first-order (visible) Markov models. However, Viterbi's time complexity is exponential in the context length of the underlying Markov model (Conklin, 2003).

Tractable methods for sampling from complex statistical models (such as those presented here) which address the limitations of random sampling do exist, however (Conklin, 2003). The *Metropolis-Hastings algorithm* is a Markov Chain Monte Carlo (MCMC) sampling method (MacKay, 1998). The following description applies it within our generation framework. Given a trained multiple-feature model m for some basic feature τ_b, in order to sample from the target distribution $p_m(s \in [\tau_b]^*)$, the algorithm constructs a Markov chain in the space of possible feature sequences $[\tau_b]^*$ as follows:

1. number of iterations $N \leftarrow$ a large value; iteration number $k \leftarrow 0$; initial state $s_0 \leftarrow$ some feature sequence $t_1^j \in [\tau_b]^*$ of length j;
2. select event index $1 \leq i \leq j$ at random or based on

some ordering of the indices;

3. let s'_k be the sequence obtained by replacing event t_i at index i of s_k with a new event t'_i sampled from a distribution q which may depend on the current state s_k – in the present context, an obvious choice for q would be $\{p_m(t|t_1^{i-1})\}_{t\in[\tau_b]}$;
4. accept the proposed sequence with probability

$$\min\left[1,\ \frac{p_m(s'_k\ )\ \cdot\ q(t_i)}{p_m(s_k)\ \cdot\ q(t'_i)}\right];$$

5. if accepted, $s_{k+1} \leftarrow s'_k$, else $s_{k+1} \leftarrow s_k$;
6. if $k < N$, k++ and iterate from 2, else return s_k.

If N is large enough, the resulting event sequence s_{N-1} is guaranteed to be an unbiased sample from the target distribution $p_m([\tau_b]^*)$. However, there is no method of assessing the convergence of MCMCs nor of estimating the number of iterations required to obtain an unbiased sample (MacKay, 1998). Because these sampling algorithms explore the state space using a random walk, they can still be trapped in local minima.

Event-wise substitution is unlikely to provide a satisfactory model of phrase- or motif-level structure. Our model has a short-term component, to model intra-opus structure, but generation still relies on single-event substitutions. Pattern-discovery algorithms may be used to reveal phrase level structure, which may subsequently be preserved during stochastic sampling (Conklin, 2003).

2.2 Evaluating Computer Models of Composition

Analysis by synthesis evaluates computational models of composition by generating pieces and evaluating them with respect to the objectives of the implemented model. The method has a long history; Ames and Domino (1992) argue that a primary advantage of computational analysis of musical style is the ability to evaluate new pieces generated from an implemented theory. However, evaluation of the generated music raises methodological issues which have typically compromised the potential benefits thus afforded (Pearce et al., 2002). Often, compositions are evaluated with a single subjective comment, *e.g.,*: "[the compositions] are realistic enough that an unknowing listener cannot discern their artificial origin" (Ames and Domino, 1992, p. 186). This lack of precision makes it hard to compare theories intersubjectively.

Other research has used expert stylistic analyses to evaluate computer compositions. This is possible when a computational model is developed to account for some reasonably well-defined stylistic competence or according to critical criteria derived from music theory or music psychology. For example, Ponsford et al. (1999) gave an informal stylistic appraisal of the harmonic progressions generated by their n-gram models.

However, even when stylistic analyses are undertaken by groups of experts, the results obtained are typically still qualitative. For fully intersubjective analysis by synthesis, the evaluation of the generated compositions must be empirical. One could use an adaptation of the Turing test, where subjects are presented with pairs of compositions (one computer-generated, the other human-composed) and asked which they believe to be the computer-generated one (Marsden, 2000). Musical Turing tests yield empirical, quantitative results which may be appraised intersubjectively. They have demonstrated the inability of subjects to distinguish reliably between computer- and human-composed music. But the method can be biased by preconceptions about computer music, allows ill-informed judgements, and fails to examine the criteria being used to judge the compositions.

2.3 Evaluating Human Composition

Amabile (1996) proposes a conceptual definition of creativity in terms of processes resulting in novel, appropriate solutions to heuristic, open-ended or ill-defined tasks. However, while agreeing that creativity can only be assessed through subjective assessments of products, she criticises the use of *a priori* theoretical definitions of creativity in rating schemes and failure to distinguish creativity from other constructs. While a conceptual definition is important for guiding empirical research, a clear operational definition is necessary for the development of useful empirical methods of assessment. Accordingly, she presents a consensual definition of creativity in which a product is deemed creative to the extent that observers who are familiar with the relevant domain independently agree that it is creative. To the extent that this construct is internally consistent (independent judges agree in their ratings of creativity), one can empirically examine the objective or subjective features of creative products which contribute to their perceived creativity.

Amabile (1996) used this operational definition to develop the *consensual assessment technique* (CAT), an empirical method for evaluating creativity. Its requirements are that the task be open-ended enough to permit considerable flexibility and novelty in the response, which must be an observable product which can be rated by judges. Regarding the procedure, the judges must:

1. be experienced in the relevant domain;
2. make independent assessments;
3. assess other aspects of the products such as technical accomplishment, aesthetic appeal or originality;
4. make relative judgements of each product in relation to the rest of the stimuli;
5. be presented with stimuli and provide ratings in orders randomised differently for each judge.

Most importantly, in analysing the collected data, the inter-judge reliability of the subjective rating scales must be determined. If—and only if—reliability is high, we may correlate creativity ratings with other objective or subjective features of creative products.

Numerous studies of verbal, artistic and problem solving creativity have demonstrated the ability of the CAT to obtain reliable subjective assessments of creativity in a range of domains (Amabile, 1996, ch. 3, gives a review).

The CAT overcomes the limitations of the Turing test in evaluating computational models of musical composition. First, it requires the use of judges expert in the task

System	Features	H
A	Pitch	2.337
B	Int1stInPiece, ScaleDegree ⊗DurRatio, Thread1stInPhrase	2.163
C	Interval⊗Duration, ScaleDegree ⊗Int1stInPiece, Pitch⊗Duration, ScaleDegree⊗1stInBar, ThreadTactus, ScaleDegree⊗Duration, Interval⊗DurRatio, Int1stInPiece, Thread1stInPhrase	1.953

Table 1: The component features of Systems A, B and C and their average information content computed by 10-fold cross-validation over the dataset.

domain. Second, since it has been developed for research on human creativity, no mention is made of the computational origins of the stimuli; this avoids bias due to preconceptions. Third, and most importantly, the methodology allows more detailed examination of the objective and subjective dimensions of the creative products. Crucially, the objective attributes of the products may include features of the generative models (corresponding with cognitive or stylistic hypotheses) which produced them. Thus, we can empirically compare different musicological theories of a given style or hypotheses about the cognitive processes involved in composing in that style.

3 The Experiment

3.1 Introduction

Following Johnson-Laird (1991), we analyse the computational constraints of the melody composition task in two ways: first, examining whether our learned finite context grammars can compose stylistically-successful melodies or whether more expressive grammars are needed; and second, determining which representational structures are needed for the composition of successful melodies.

Our experiment is designed to test the hypothesis that our statistical models are capable of generating melodies which are deemed stylistically successful in the context of a specified tradition. Three multiple-feature Markov models (Pearce, 2005) trained on a dataset of chorale melodies were used to generate melodies which were then empirically evaluated: System A is a single-feature system; System B is a multiple-feature system developed through forward, stepwise feature selection to provide the closest fit to the human expectancy judgements obtained by Manzara et al. (1992); and System C is a multiple-feature system developed through forward, stepwise feature selection to yield the best prediction performance over the chorale dataset. The Systems were parameterised optimally and differ only in the features they use (Table 1).

Our work differs in several ways from extant statistical modelling for music generation, in particular, in that no symbolic constraints were imposed on the generation process—it was based entirely on the learned models. This focuses the analysis more sharply on the inherent capacities of statistical finite context grammars, since our goal was to examine the synthetic capabilities of purely statistical, data-driven models of melodic structure.

Our strategy improves on previous work in several ways. The variable order selection policy of PPM* (Cleary and Teahan, 1997) is used to address concerns that low, fixed order models tend to generate features uncharacteristic of the target style (Ponsford et al., 1999). Other model parameters are optimised to improve prediction performance over a range of different melodic styles. Systems B and C operate over rich representational spaces supplied by the multiple-feature framework; their features were selected on the basis of objective and empirical criteria (*cf.* Conklin and Witten, 1995). Our Systems use a novel model combination strategy, which improves prediction performance over the chorale dataset (Pearce, 2005). While most previous approaches used sequential random sampling to generate music from statistical models, in the present research melodies were generated using Metropolis sampling. We expect that this method will be capable of generating melodies which are more representative of the inherent capacities of the Systems. We do not propose Metropolis sampling as a cognitive model of melodic composition, but use it merely as a means of generating melodies which reflect the internal state of knowledge and capacities of the trained models.

Finally, to evaluate the systems as computational models of melodic composition, we developed a method based on the CAT. The method, described fully by Pearce (2005), obtains ratings by expert judges of the stylistic success of computer generated compositions and existing compositions in the target genre. The empirical nature of this method makes it preferable to the exclusively qualitative analyses typically adopted and we expect it to yield more revealing results than the Turing test methodology.

3.2 Hypotheses

We use three different Systems to examine which representational structures are needed for competent melody generation. Our null hypotheses are that each System can generate melodies rated as equally stylistically successful in the target style as existing, human-composed melodies. We expect the null hypothesis for the simplistic System A to be refuted.

For System B, Baroni's (1999) proposal that composition and listening involve equivalent grammatical structures is relevant. If the representational structures underlying perception and composition of music are similar, we would expect grammars which model perceptual processes well to generate satisfactory compositions. Since System B represents a satisfactory model of the perception of pitch structure in the chorale genre, we may expect to retain the null hypothesis for this system.

Pearce and Wiggins (2006) demonstrate a relationship between prediction performance and fit to human expectancy data (Manzara et al., 1992), suggesting that human perceptual systems base their predictions on uncertainty-reducing representational features. In terms of model selection for music generation, highly predictive theories of a musical style, as measured by information content, should generate original and acceptable works in the style (Conklin and Witten, 1995). Systems A, B and C in turn exhibit decreasing uncertainty in predicting unseen melodies from the dataset (Table 1). Therefore, we may expect to retain the null hypothesis for System C.

3.3 Method

3.3.1 Judges

Our judges were 16 music researchers or students at City University, London, Goldsmiths, University of London, and the Royal College of Music. Five were male and eleven female, and their age range was 20–46 years (mean 25.9, SD 6.5). They had been formally musically trained for 2–40 years (mean 13.8, SD 9.4). Seven judges reported high familiarity with the chorale genre and nine were moderately familiar. All judges received a nominal payment, and worked for approximately an hour.

3.3.2 Apparatus and Stimulus Materials

Our dataset is a subset of the chorale melodies placed in the soprano voice and harmonised in four parts by J. S. Bach. These melodies are characterised by stepwise patterns of conjunct intervallic motion and simple, uniform rhythmic and metric structure. Phrase structure is explicitly notated. Most phrases begin on the tonic, mediant or dominant and end on the tonic or dominant; the final phrase almost always ends with a cadence to the tonic.

Our stimuli were as follows. Seven existing *base* melodies were randomly selected from the set of chorales in the midrange of the distribution of average information content (cross-entropy) values computed by System A. All 7 were in common time; 6 were in major keys and 1 was minor; they were 8–14 bars (mean 11.14) and 33–57 events (mean 43.43) long. The base melodies were removed from the training dataset.

7 novel melodies were generated by each System, *via* 5000 iterations of Metropolis sampling using the 7 base chorales as initial states. Only pitch was sampled: time and key signatures and rhythmic and phrase structure were left unchanged. Figure 1 shows one base chorale melody and the three melodies generated using it; Pearce (2005) gives further examples.

Each melody was stored as a quantised MIDI file. A pattern of velocity accents was added to emphasise the metrical structure and a one-beat rest was inserted after each fermata to disambiguate the phrase structure. The stimuli were recorded to CD-quality audio files on a PC using the piano tone of a Roland XP10 synthesiser connected via MIDI to a Terratec EWS88 MT soundcard, at a uniform 90 beats per minute. They were presented over Technics RP-F290 stereo headphones fed from a laptop PC running a software media player. The judges recorded their responses in writing in a response booklet.

3.3.3 Procedure

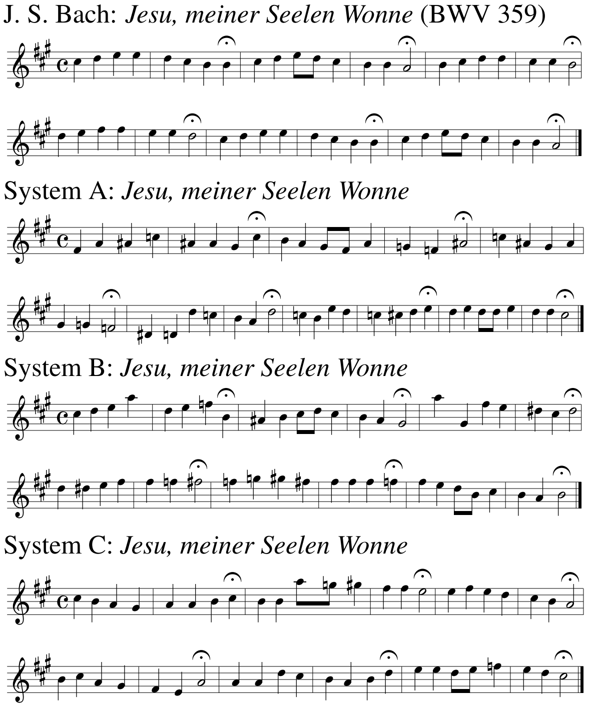
J. S. Bach: *Jesu, meiner Seelen Wonne* (BWV 359)
System A: *Jesu, meiner Seelen Wonne*
System B: *Jesu, meiner Seelen Wonne*
System C: *Jesu, meiner Seelen Wonne*

Figure 1: An example of one base chorale melody and the three melodies generated using it.

Our judges supplied their responses individually and received instructions verbally and in writing. We told them they would hear a series of chorale melodies in the style of Lutheran hymns and asked them to listen to each entire melody before answering two questions about it by placing circles on discrete scales in the response booklet. The first question[1] was, “How successful is the composition as a chorale melody?” Judges were advised that their answers should reflect such factors as conformity to important stylistic features, tonal organisation, melodic shape and interval structure; and melodic form. Answers to this question were given on a seven-point numerical scale, 1–7, with anchors marked low (1), medium (4) and high (7). To promote an analytic approach to the task, judges were asked to briefly justify their responses to the first question. The second question was, “Do you recognise the melody?” Judges were advised to answer “yes” only if they could specifically identify the composition as one they were familiar with.

We explained to the judges that after both questions had been answered for a melody, they could listen to the next one by pressing a single key on the PC. We asked them to bear in mind that their task was to rate the composition of each melody rather than the performance and urged them to use the full range of the scales, reserving 1 and 7 for extreme cases. There were no constraints on the time taken to answer the questions.

The experiment began with a practice session during which judges heard two melodies from the same genre (but not one of those in the test set). These practice trials were intended to set a judgemental standard for the subsequent test session. This departs from the CAT, which encourages judges to rate each stimulus in relation to the others by experiencing all stimuli before making their ratings. However, here, we intended the judges to use their expertise to rate the stimuli against an absolute standard: the body of existing chorale melodies. Judges responded

[1] This is a variant on the original CAT, whose primary judgement was about creativity. We justify this on the grounds that stylistic success is a directly comparable kind of property.

as described above for both of the items in the practice block. The experimenter remained in the room for the duration of the practice session after which the judges were given an opportunity to ask any further questions; he then left the room before the start of the test session.

In the test session, the 28 melodies were presented to the judges, who responded to the questions. The melodies were presented in random order subject to the constraints that no melody generated by the same system nor based on the same chorale were presented sequentially. A reverse counterbalanced design was used, with eight of the judges listening to the melodies in one such order and the other eight listening to them in the reverse order.

After the test session, the judges filled out a questionnaire detailing their age, sex, number of years of music training (instrument and theory) and familiarity with the chorales harmonised by J. S. Bach (high/medium/low).

3.4 Results

3.4.1 Inter-judge Consistency

We report analyses of the 28 melodies from our test session: we discarded the data from the practice block. First, we examine the consistency of the judges' ratings.

All but two of the 120 pairwise correlations between judges were significant at $p < 0.05$ with a mean coefficient of $r(26) = 0.65$ ($p < 0.01$). Since there was no apparent reason to reject the judges involved in the two non-significant correlations, we did not do so. This high consistency warrants averaging the ratings for each stimulus across individual judges in subsequent analyses.

3.4.2 Presentation Order and Prior Familiarity

Two factors which might influence the judges' ratings are the order of presentation of the stimuli and prior familiarity. The correlation between the mean success ratings for judges in the two groups was $r(26) = 0.91, p < 0.01$ indicating a high degree of consistency across the two orders of presentation, and warranting the averaging of responses across the two groups; and, although the mean success ratings tended to be slightly higher when judges recognised the stimulus, a paired t test revealed no significant difference: $t(6) = 2.07, p = 0.08$.

3.4.3 Influence of Generative System and Base Chorale

Now we examine the primary question: the influence of generative system on the ratings of stylistic success. The mean success ratings for each stimulus are shown in Table 2. The mean ratings suggest that the original chorale melodies were rated higher than the computer-generated melodies while the ratings for the latter show an influence of base chorale but not of generative system. Melody C249 is an exception, attracting high average ratings of success. Our preferred analysis would have been a multivariate ANOVA using within-subjects factors for generative system with 4 levels (Original, System A, B, C) and base chorale with 7 levels (249, 238, 365, 264, 44, 153 and 147) with the null hypotheses of no main or interaction effects of generative system or base chorale. However, Levene's test revealed significant non-homogeneity of variance with respect to the factor for generative system $F(3) = 6.58, p < 0.01$, so ANOVA was not applicable. Therefore, we used Friedman's rank sum tests, as a non-parametric alternative; this does not allow examination of interactions between the two factors.

Base	System A	System B	System C	Original	Mean
249	2.56	2.44	5.00	6.44	4.11
238	3.31	2.94	3.19	5.31	3.69
365	2.69	1.69	2.50	6.25	3.28
264	1.75	2.00	2.38	6.00	3.03
44	4.25	4.38	4.00	6.12	4.69
141	3.38	2.12	3.19	5.50	3.55
147	2.38	1.88	1.94	6.50	3.17
Mean	2.90	2.49	3.17	6.02	3.65

Table 2: The mean success ratings for each stimulus and means aggregated by generative system and base chorale.

Statistic	System A	System B	System C	Original
Median	2.86	2.57	3.07	5.93
Q1	2.68	2.25	2.68	5.86
Q3	3.29	2.75	3.61	6.29
IQR	0.61	0.50	0.93	0.43

Table 3: The median, quartiles and inter-quartile range of the mean success ratings for each generative system.

We examined the influence of generative system in an unreplicated complete blocked design using the mean success ratings aggregated for each subject and generative system across the individual base chorales. Summary statistics for this data are shown in Table 3. The Friedman test revealed a significant within-subject effect of generative system on the mean success ratings: $\chi^2(3) = 33.4, p < 0.01$. We compared the factor levels pairwise using Wilcoxon rank sum tests with Holm's Bonferroni correction for multiple comparisons: the ratings for the original chorale melodies differ significantly from the ratings of melodies generated by all three computational systems ($p < 0.01$). Furthermore, the mean success ratings for the melodies generated by System B were found to be significantly different from those of the melodies generated by Systems A and C ($p < 0.03$). These results suggest that none of the systems is capable of consistently generating chorale melodies which are rated as equally stylistically successful as those in the dataset and that System B performed especially poorly.

4 Learning from Qualitative Feedback

4.1 Objective Features of the Chorales

Next, we aim to explain how the Systems lack compositionally, by examining which objective musical features of the stimuli the judges used in making their ratings of stylistic success. This could explain how the systems are lacking compositionally. To achieve this, we analysed the stimuli qualitatively and developed a set of corresponding objective descriptors, which we then applied in a series of multiple regression analyses using the rating scheme, averaged across stimuli, as a dependent variable. We now present the descriptive variables, their quantitative coding and the analysis results.

The chorales generated by our systems are mostly not very stylistically characteristic of the dataset, especially in higher-level form. From the judges' qualitative comments, we identified stylistic constraints describing the stimuli and distinguishing the original melodies. We grouped them into five categories—pitch range; melodic structure; tonal structure; phrase structure; and rhythmic structure—each covered by a predictor variable.

Pitch Range The dataset melodies span a pitch range of about an octave above and below C_5, favouring the centre of this range. The generated melodies are constrained to this range, but some tend towards extreme tessitura. We developed a predictor variable *pitch centre* to capture this difference, reflecting the absolute distance, in semitones, of the mean pitch of a melody from the mean pitch of the dataset (von Hippel, 2000). Another issue is the overall pitch range of the generated chorales. The dataset melodies span an average range of 11.8 semitones. By contrast, several of the generated melodies span pitch ranges of 16 or 17 semitones, with a mean pitch range of 13.9 semitones; others have a rather narrow pitch range. We captured these qualitative considerations in a quantitative predictor variable *pitch range*, representing the absolute distance, in semitones, of the pitch range of a melody from the mean pitch range of the dataset.

Melodic Structure There are several ways in which the generated melodies do not consistently reproduce salient melodic features of the original chorales. The most obvious is a failure to maintain a stepwise pattern of movement. While some generated melodies are relatively coherent, others contain stylistically uncharacteristic leaps of an octave or more. Of 9042 intervals in the dataset melodies, only 57 exceed a perfect fifth and none exceeds an octave. To capture these deviations, we created a quantitative predictor variable called *interval size*, representing the number of intervals greater than a perfect octave in a melody. The generated chorales also contain uncharacteristic discords such as tritones or sevenths. Only 8 of the 9042 intervals in the dataset are tritones or sevenths (or their enharmonic equivalents). To capture these deviations, we created a quantitative predictor variable *interval dissonance*, representing the number of dissonant intervals greater than a perfect fourth in a melody.

Tonal Structure Since System A operates exclusively over representations of pitch, it is not surprising that most of its melodies fail to establish a key note and exhibit little tonal structure. However, we might expect Systems B and C to do better. While the comments of the judges suggest otherwsie, they may have arrived at a tonal interpretation at odds with the intended key of the base chorale. To independently estimate the perceived tonality of the test melodies, Krumhansl's (1990) key-finding algorithm, using the revised key profiles of Temperley (1999) was applied to each of the stimuli. The algorithm assigns the correct keys to all seven original chorale melodies. While the suggested keys of the melodies generated by System A confirm that it does not consider tonal constraints, the melodies generated by Systems B and C retain the key of their base chorale in two and five cases respectively. Furthermore, especially in the case of System C, deviations from the base chorale key tend to be to related keys (either in the circle of fifths or through relative and parallel major/minor relationships). This suggests some success on the part of the more sophisticated systems in retaining the tonal characteristics of the base chorales.

Nonetheless, the generated melodies are often unacceptably chromatic, which obscures the tonality. Therefore, we developed a quantitative predictor called *chromaticism*, representing the number of chromatic tones in the algorithm's suggested key.

Phrase Structure The generated chorales typically fail to reproduce the implied harmonic rhythm of the originals and its characteristically strong relationship to phrase structure. In particular, while some of the generated melodies close on the tonic, many fail to imply stylistically satisfactory harmonic closure. To capture such effects, we created a variable called *harmonic closure*, which is 0 if a melody closes on the tonic of the key assigned by the algorithm and 1 otherwise. Secondly, the generated melodies frequently fail to respect thematic repetition and development of melodic material embedded in the phrase structure of the chorales. However, these kinds of repetition and development of melodic material are not represented in the present model. Instead, as a simple indicator of complexity in phrase structure, we created a variable *phrase length*, which is 0 if all phrases are of equal length and 1 otherwise.

Rhythmic Structure Although the chorale melodies in the dataset tend to be rhythmically simple, the judges' comments revealed that they were taking account of rhythmic structure. Therefore, we adapted three further quantitative predictors modelling rhythmic features from Eerola and North's (2000) expectancy-based model of melodic complexity. *Rhythmic density* is the mean number of events per tactus beat. *Rhythmic variability* is the degree of change in note duration (*i.e.,* the standard deviation of the log of the event durations) in a melody. *Syncopation* estimates the degree of syncopation by assigning notes a strength in a metric hierarchy and averaging the strengths of all the notes in a melody; pulses are coded such that lower values are assigned to tones on metrically stronger beats. All three quantities increase the difficulty of perceiving or producing melodies (Eerola and North, 2000).

The mean success ratings for each stimulus were regressed on the predictor variables in a multiple regression analysis. The following pairwise correlations between the predictors were significant at $p < 0.05$: interval size, positively with interval dissonance ($r = 0.6$) and chromaticism ($r = 0.39$); harmonic closure, positively with chromaticism ($r = 0.49$); rhythmic variation, positively with syncopation ($r = 0.61$) and phrase length ($r = 0.73$); and rhythmic density, positively with syncopation ($r = 0.62$) and negatively with phrase length ($r = -0.54$). Because of this collinearity, in each analysis, redundant predictors were removed through backwards stepwise elimination using the Akaike Information Criterion: $AIC = n \log(RSS/n) + 2p + c$, for a regression model with p predictors and n observations, where c is a constant and RSS is the residual sum of squares of the model (Venables and Ripley, 2002). Since larger models

Predictor	β	Std. Error	t	p
Pitch Range	-0.29	0.08	-3.57	< 0.01
Pitch Centre	-0.21	0.10	-2.01	< 0.1
Interval Dissonance	-0.70	0.28	-2.54	< 0.05
Chromaticism	-0.27	0.03	-8.09	< 0.01
Phrase Length	-0.53	0.28	-1.91	< 0.1

Overall model: $R = 0.92$, $R^2_{adj} = 0.81$,
$F(5, 22) = 25.04, p < 0.01$

Table 4: Multiple regression results for the mean success ratings of each test melody.

Stage	Feature Added	H
1	`Interval⊗Duration`	2.214
2	`ScaleDegree⊗Mode`	2.006
3	`ScaleDegree ⊗Int1stInPiece`	1.961
4	`Pitch⊗Duration`	1.943
5	`Thread1stInPhrase`	1.933
6	`ScaleDegree ⊗LastInPhrase`	1.925
7	`Interval⊗DurRatio`	1.919
8	`Interval⊗InScale`	1.917
9	`ScaleDegree⊗Duration`	1.912
10	`Int1stInPhrase`	1.911

Table 5: Results of feature selection for reduced information content over the dataset using an extended feature set.

provide better fits, this criterion balances model size, represented by p, with the fit of the model to the dependent variable, RSS.

More positive values of the predictors indicate greater deviation from the standards of the dataset (for pitch range and centre) or increased melodic complexity (for the remaining predictors), so we expect each predictor to show a negative relationship with the success ratings. The results of the multiple regression analysis with the mean success ratings as the dependent variable are shown in Table 4. The overall model accounts for approximately 85% of the variance in the mean success ratings. Apart from rhythmic structure, at least one predictor from each category made at least a marginally significant contribution to the fit of the model. Coefficients of all the selected predictors are negative as predicted. Overall, the model indicates that the judged success of a stimulus decreases as its pitch range and centre depart from the mean range and centre of the dataset, with increasing numbers of dissonant intervals and chromatic tones and if it has unequal phrase lengths.

4.2 Improving the Computational Systems

The constraints identified above mainly concern pitch range, intervallic structure and tonal structure. It seems likely that the confusion of relative minor and major modes is due to the failure of any of the Systems to represent mode. To examine this hypothesis, a linked feature `ScaleDegree⊗Mode` was added to the feature space. Furthermore, we hypothesise that the skewed distribution of pitch classes at phrase beginnings and endings can be better modelled by two linked features `ScaleDegree⊗1stInPhrase` and `ScaleDegree⊗LastInPhrase`. On the hypothesis that intervallic structure is constrained by tonal structure, we included another linked feature `Interval⊗InScale`.

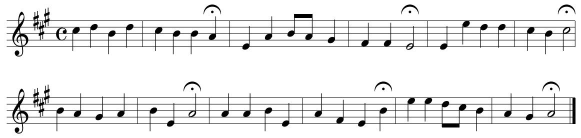

Figure 2: Melody generated by System D, based on the same chorale as Figure 1.

To examine whether the Systems can be improved to respect such constraints, we added the four selected features to the feature selection set used for System C. We ran the same feature selection algorithm over this extended feature space to select feature subsets which improve prediction performance; the results are shown in Table 5. In general, the resulting multiple-feature System, D, shows a great deal of overlap with System C. Just three of the nine features present in System C were not selected for inclusion in System D: `ScaleDegree⊗1stInBar`; `ThreadTactus`; and `Int1stInPiece`. This is probably because three of the four new features selected for inclusion in System D, were strongly related: `ScaleDegree⊗Mode`; `ScaleDegree⊗LastInPhrase`; and `Interval⊗InScale`. The first two of these, in particular, were selected early in the selection process; the existing feature `Int1stInPhrase` was added in the final stage. Ultimately, System D exhibits a lower average information content ($H = 1.91$) than System C ($H = 1.95$) in predicting unseen compositions in the dataset. The significance of this difference was confirmed by paired t tests over all 185 chorale melodies: $t(184) = 6.00, p < 0.01$, and averaged for each 10-fold partition of the dataset: $t(9) = 12.00, p < 0.01$.

4.3 A Melody Generated by System D

We now present preliminary results on System D's capacity to generate stylistically successful chorale melodies. System D uses the features in Table 5; it exhibits significantly lower entropy than System C in predicting unseen melodies. We used it to generate several melodies, as described above, with the same base melodies.

Figure 2 shows System D's most successful melody, based on Chorale 365. Its tonal and melodic structure are much more coherent than System C's melodies. Our multiple regression model, developed above to account for the judges' ratings of stylistic success, predicts that this melody would receive a rating of 6.4 on a seven-point scale of success as a chorale melody. While this result is positive, other melodies were less successful; System D must be analysed using our method to examine its ability to *consistently* compose stylistically successful melodies.

5 Discussion and Conclusions

Our statistical finite context grammars did not meet the computational demands of chorale melody composition, regardless of the representational primitives used. Since we attempted to address the limitations of previous context-modelling approaches to generating music, we might conclude that more powerful grammars are needed for this task. However, other approaches are possible. Further analysis of the capacities of finite context modelling systems may prove fruitful: future research should use the methodology developed here to analyse System D, and identify and correct its weaknesses. The MCMC generation algorithm may be responsible for failure, rather than the limitation of the models to finite context representations of melodic structure: more structured generation strategies, such as pattern-based sampling techniques, may be able to conserve phrase-level regularity and repetition in ways that our Systems were not.

Our evaluation method also warrants discussion. The adapted CAT yielded insightful results for ratings of stylistic success even though the judges were encouraged to rate the stimuli according to an absolute standard (*cf.* Amabile, 1996). However, the results suggest possible improvements: first, avoid any possibility of method artefacts by randomising the presentation order of both test and practice items for each judge and also the order in which rating scales are presented; second, the judges' comments sometimes reflected aesthetic judgements, so they should also give ratings of aesthetic appeal, to delineate subjective dimensions of the product domain in the assessment (Amabile, 1996); and third, though influence of prior familiarity with the test items was ambiguous, bias resulting from recognition should be avoided.

Our results suggest that the task of composing a stylistically successful chorale melody presents significant challenges as a first step in modelling cognitive processes in composition. Nonetheless, our evaluation method proved fruitful in examining the generated melodies in the context of existing pieces in the style. It facilitated empirical examination of specific hypotheses about the models through detailed comparison of the generated and original melodies on several dimensions. It also permitted examination of objective features of the melodies which influenced the ratings and subsequent identification of weaknesses in the Systems and directions for improving them. This practically demonstrates the utility of analysis by synthesis for evaluating cognitive models of composition—if it is combined with an empirical methodology for evaluation such as that developed here.

References

Amabile, T. M. (1996). *Creativity in Context*. Westview Press, Boulder, Colorado.

Ames, C. and Domino, M. (1992). Cybernetic Composer: An overview. In Balaban, M., Ebcioğlu, K., and Laske, O., editors, *Understanding Music with AI: Perspectives on Music Cognition*, pages 186–205. MIT Press, Cambridge, MA.

Baroni, M. (1999). Musical grammar and the cognitive processes of composition. *Musicæ Scientiæ*, 3(1):3–19.

Cleary, J. G. and Teahan, W. J. (1997). Unbounded length contexts for PPM. *The Computer Journal*, 40(2/3):67–75.

Conklin, D. (2003). Music generation from statistical models. In *Proceedings of the AISB 2003 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, pages 30–35, Brighton, UK. SSAISB.

Conklin, D. and Witten, I. H. (1995). Multiple viewpoint systems for music prediction. *Journal of New Music Research*, 24(1):51–73.

Eerola, T. and North, A. C. (2000). Expectancy-based model of melodic complexity. In Woods, C., Luck, G., Brochard, R., Seddon, F., and Sloboda, J. A., editors, *Proceedings of the Sixth International Conference on Music Perception and Cognition*, Keele, UK. Keele University.

Johnson-Laird, P. N. (1991). Jazz improvisation: A theory at the computational level. In Howell, P., West, R., and Cross, I., editors, *Representing Musical Structure*, pages 291–325. Academic Press, London.

Krumhansl, C. L. (1990). *Cognitive Foundations of Musical Pitch*. Oxford University Press, Oxford.

MacKay, D. J. C. (1998). Introduction to Monte Carlo methods. In Jordan, M. I., editor, *Learning in Graphical Models*, NATO Science Series, pages 175–204. Kluwer Academic Press, Dordrecht, The Netherlands.

Manzara, L. C., Witten, I. H., and James, M. (1992). On the entropy of music: An experiment with Bach chorale melodies. *Leonardo*, 2(1):81–88.

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

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition*. PhD thesis, Department of Computing, City University, London, UK.

Pearce, M. T., Meredith, D., and Wiggins, G. A. (2002). Motivations and methodologies for automation of the compositional process. *Musicae Scientiae*, 6(2):119–147.

Pearce, M. T. and Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, 23(5):377–406.

Ponsford, D., Wiggins, G. A., and Mellish, C. (1999). Statistical learning of harmonic movement. *Journal of New Music Research*, 28(2):150–177.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285.

Temperley, D. (1999). What's key for key? The Krumhansl-Schmuckler key-finding algorithm reconsidered. *Music Perception*, 17(1):65–100.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York.

von Hippel, P. T. (2000). Redefining pitch proximity: Tessitura and mobility as constraints on melodic intervals. *Music Perception*, 17(3):315–127.