Musical Metacreation: Papers from the 2013 AIIDE Workshop (WS-13-22)

Tracking Creative Musical Structure: The Hunt for the Intrinsically Motivated Generative Agent

Benjamin D. Smith
Indiana University Purdue University Indianapolis
535 W. Michigan St.
Indianapolis, Indiana 46202

Abstract

Neural networks have been employed to learn, generalize, and generate musical pieces with a constrained notion of creativity. Yet, these computational models typically suffer from an inability to characterize and reproduce long-term dependencies indicative of musical structure. Hierarchical and deep learning models propose to remedy this deficiency, but remain to be adequately proven. We describe and examine a novel dynamic bayesian network model with the goal of learning and reproducing longer-term formal musical structures. Incorporating a computational model of intrinsic motivation and novelty, this hierarchical probabilistic model is able to generate pastiches based on exemplars.

Introduction

What is the primary task of a generative music system positioned as "creative?" A pragmatic approach is to answer with a system model which, given a specific context, can produce a creative continuation (in the form of a new musical statement). If the space of potential continuations can be known, or computed tractably, then this model effectively seeks to continually answer "is this particular option more or less creative than every other?" However, this pragmatic approach, and the evaluation of any results, hinges on a working formulation of the notion of "creativity."

Due to the inherent subjectivity and differences of experience, culture, goals, and individual contexts wrapped up with human creativity this proves a difficult task. Taking a pluralistic approach, Boden (2004) describes "creativity" as a relative concept dependent on the current context, knowledge, and understandings of the individuals involved in the reification or evaluation of the process. This opens the concept of creativity to encompass instances of any size, from a child's drawings to Nobel Prize winners (suggesting the assignment of different metrics and values based on scale and impact). From this description a highly constrained notion of creativity can be derived, wherein a generative agent (human or computational) seeks to satisfy a narrowly scoped model of novelty and innovation. The model described herein claims "creativity" based on its ability to locate novel combinations of musical material suggested by pattern matching on given training exemplars. In a sense, the system takes the "child's drawing" approach, incorporating previous experiences and knowledge in a small scale exploration.

The compositional process employed, of continually evaluating with narrow foresight while generating a musical work, is often implemented through neural networks and recurrent neural networks (Eck and Schmidhuber 2002) due to their ability to efficiently learn and reproduce note-to-note dependencies. However, the results typically suffer from problems of local myopia (failing to produce creative structures beyond note-to-note sequences) and the "vanishing gradients" problem (Hochreiter et al. 2001) (wherein the error flow in gradient learning methods vanishes or explodes, resulting in an inability to robustly store past information about inputs). The result can easily be seen as aimless and incoherent music without definitive formal structure, style, and phrasing (Mozer 1999; Smith and Garnett 2012b).

Work examining hierarchical neural networks (Smith and Garnett 2012a) shows promise but has yet to be fully evaluated. These models attempt to entrain and produce longer time-scale forms through a stacked structure of neural networks wherein each layer guides the output of the adjoining layers. However, the evaluation in (Smith and Garnett 2012a) employs a brute-force method rather than a true probabilistic resolution, producing limited results. Taking inspiration from this work we propose a dynamic bayesian network (DBN) (Murphy 2002) to generate music that exhibits both micro and macro levels of form. The initial results, included below, show promise towards creating works that mimic the training corpus with novel variations.

System Goals

The focus of the proposed system is to identify and produce musical sequences based on computational models of creativity and novelty. Currently this takes the form of "improvisations in the style of" pastiches. The results should be two-fold: the generation of novel, valuable musical works, as well as serving as a framework for the evaluation of equations claiming creative results. In order for the DBN framework to prove effective a calculation must be performed that proposes the "creative" value of each transition (i.e. new note) in a piece (either as a training exemplar or a generated composition). This takes the form of a calculation, or equation, that claims the characteristic of creative valuation.

The initial orientation of the system is to generate a sequence of pitches that a human expert would agree exhibits a degree of creativity or novelty. This is the minimal case, intentionally stripping out factors such as dynamics, articulations, rhythmic variation, polyphony, etc. (which, not coincidentally, is characteristic of certain works from western classical music of the Baroque and early Classical era.) Additionally the system is constrained to four octaves of the western, equal-tempered, 12-tone scale for purposes of tractability (this is also a common range for acoustic melodic instruments in conventional practice). Ideally these constraints will allow a focused comparison and evaluation of the system's performance.

The "novelty" seeking, intrinsically motivated reinforcement learning models proposed by Schmidhuber (2009) serve as a starting point for the system's operation. Rather than explicitly follow the reinforcement learning model, which takes a trial-and-error approach and must make "mistakes" in order to discover better solutions, our proposal is to plug the "novelty" rewarding algorithm (from Schmidhuber) into the tracking functionality of a DBN to predict the best solutions at every time point, converging on the most rewarding musical path over time. Again, the purpose of the system is to both discover what these algorithms can produce towards a notion of creative music, as well as evaluate the limits of said algorithms.

The notion of "intrinsic motivation" describes a process wherein an agent continually seeks out new inputs that strike a balance between a sense of *boredom* and *confusion.* In a general sense this is a matter of finding new sensory stimulation (musical sequences, in this case) that is closely related to previously processed events, but does not merely exhibit the repetition of known patterns (which is deemed "boring"). The alternative is stimulation that is so extremely different that no patterning can be identified, given the agent's knowledge base, resulting in "confusion." An intrinsically motivated agent continually processes new inputs, gradually seeking out new stimulation to expand its pattern database and learn its environment.

We draw on the metaphor of a musical *path* or *journey* through a piece, which exists in a musical *landscape* defined by relationships between works, styles, and exemplar compositions. Thus the system should predict a maximally, intrinsically satisfying path (i.e. novel or inventive sequence of notes) through a landscape (database of melodic patterns encoded with dependancies). This task is further complicated by the impact that the walking of that path (i.e. the playing of the piece) has on the landscape itself. A simple example of this is the importance that the exposition of a Sonata has on the rest of the movement and subsequent movements. In effect the walk through the exposition carves out new musical space which bears special significance for everything that follows.

The model presented below starts with given training exemplars (i.e. a piece of music composed and selected by a human) and is then tasked to generate new material as a regular sequence of notes. The selection of these notes is weighted by the "novelty" of the resulting patterns and the chosen results are fed back into the system, which adapts and reinforces the choices made.

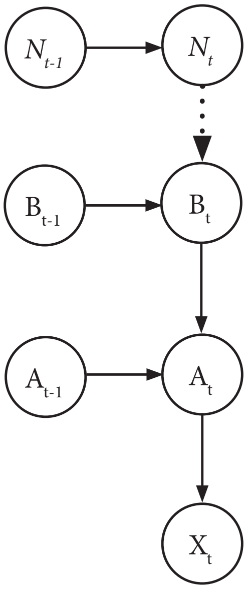

Figure 1: Single time slice of the DAG. X is the sequence of notes, $A...N$ are abstracted pattern sequences.

Model

The presupposition of this working model is that one or more *creative* sequences of musical events exist, within the constraints presented for a given piece (including the individual, subjective aesthetic preferences of any chosen listener.) Additionally, that some sequences are more creative than others and an initial algorithm, based on the notion of "novelty," proposes to obtain this valuation. With such sequences theoretically available a DBN can be employed to track this sequence (albeit with some degree of approximation or error). Figure 1 shows the proposed network model.

The DBN updates at the note rate (one time step is one 16th note), attempting to identify the maximally interesting next choice with each prediction step. The observed sequence of notes X is predicted through an arbitrarily deep stack of abstract dependencies $(A, B, ...N)$. While a nearly infinite stack of nodes is theoretically desirable towards encoding longer-term dependencies, in implementation the stack is fixed by the operator. The DAG in figure 1 admits the following factorization:

$$\begin{aligned} P(N_{1:T}, B_{1:T}, A_{1:T}, X_{1:T}) = P(N_1)P(B_1|N_1) \qquad (1) \\ \times P(A_1|B_1)P(X_1|A_1) \times \prod_{t=2}^{T} P(N_t|N_{t-1}) \\ \times P(B_t|B_{t-1}, N_t)P(A_t|A_{t-1}, B_t)P(X_t|A_t) \end{aligned}$$

Certain characteristics of melodic pitch sequences have become widely accepted in western theory and these allow the refactoring of A as shown in figure 2. Specifically,

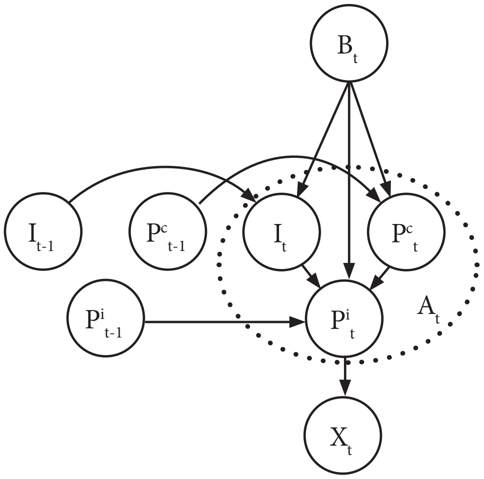

Figure 2: Timeslice of the DAG showing the internal expansion of A with dependencies. P^i is pitch, P^c is pitch class, I is interval.

three first order characteristics of musical notes are tracked interdependently: the absolute pitch of the note (P^i), the pitch class of the note (P^c, based on the notion of octave-equivalency in the western, equal-tempered scale), and interval (I, correlated with the signed difference between X_t and X_{t-1}). This results in the DAG in figure 2 and following factorization of a single time slice:

$$P(A_t|A_{t-1}, B_t) = P(P^c_t|P^c_{t-1}, B_t)P(I_t|I_{t-1}, B_t) \times P(P^i_t|P^i_{t-1}, P^c_t, I_t, B_t)P(X_t|P^i_t) \quad (2)$$

In this formulation X is the sequence of notes we seek to track, P^i is the sequence of pitches of those notes, P^c the pitch classes, and I the intervals between pitches (note: currently X and P^i are equivalent, however this is designed with the potential to expand X in the DAG to encompass rhythm, dynamics, etc.). The upper layers, $B...N$, each track the state-transitions of the next lower layer. A parallel concept is the notion of form wherein musical motif 1 (a short sequence of notes) may be followed by motif 2 (a different short sequence), then 3, and then back to motif 1. In the DBN, each layer trains such sequences at levels of abstraction increasing from the note level at the bottom.

Data Representation

The probability of any lateral transition ($P(B_t|B_{t-1})$, for example) is a continually changing function based on the knowledge base of our system-agent and the developments observed in the currently generated piece (i.e. $B_0...B_{t-1}$). As previously discussed, the agent is learning during the generation, adjusting its notion of what qualifies as "novel" based on what it has heard previously and/or generated. This is implemented from Schmidhuber (2009), who presents many formulations of a continually developing intrinsic reward function based on knowledge acquisition. Based on Smith (2012a) we calculate the transition probability (such as $P(B_t|B_{t-1})$), R as:

$$R_t = \frac{s(r_t + a)}{\Delta n_t} \quad (3)$$

which has the property of rewarding longer sequences (s, of states or notes) that can be classified (clumped) more efficiently (where n is the size, in data memory, of the neural net at this layer). The change in data memory size Δn_t gives an indication of novelty in that previously seen input patterns will not require the neural nets to grow, while radically new patterns will. The learning residual component (r_t, the change in weights of the nodes of the neural nets encoding a given layer) similarly reflects the new-ness of the input, and gives more weight to new patterns, discounting purely repetitive sequences (higher residuals result in a higher weight). The constant $0 < a \leq 1$ offsets r, preventing $R_t = 0$ which would result in rapid reductions of the probability space, discarding potentially rewarding musical paths before they can be explored.

Placing equation 3 in context gives:

$$P(B_t|B_{t-1}) = \frac{s_{B_t}(r_{B_t} + a)}{\Delta n_{B_t}} \quad (4)$$

The dataset that informs the probabilities for each node in the DBN are retained by an adaptive neural network model (i.e. an adaptive classifier). We retain this data through fuzzy Adaptive Resonance Theory (ART) (Carpenter, Grossberg, and Rosen 1991) networks. The ART model is an adaptive neural network that classifies inputs into automatically derived categories (or clumps). Inputs typically take the form of fixed dimension, normalized feature vectors (of real numbers, in the fuzzy ART variant). Running in a sequential fashion, ARTs create new category designations as needed when new inputs are presented. There is no theoretical limit on the number of categories an ART network can assign, instead the clumping is based on a "vigilance" parameter which guides the ART in segmenting the feature space (low vigilance allows larger clumps to be formed, while high vigilance enforces stricter discrimination between inputs).

The "novelty" equation works on three elements that are known at each transition: the length of the note sequence (s), the learning residual (r, the amount of change in the ART node weights resulting from the presentation of the next token), and the size of the ART (n, number of bytes required to store the dynamically growing neural net).

All nodes in the DAG (above A) represent a dynamically growing state space involving a token sequence denoting discrete state changes over time. In this way a state transition can be expressed as "how novel is X new state given our previous state sequence?" Because each ART is expanding to classify new state-change patterns the state-token space is continually growing, adapting and allocating new neural nodes to encode the new patterns.

For the purposes of tractable evaluation the PDF for transitions of P_i is modeled as a uniform distribution of ± 12 semi-tones from the previous state. One time step is equated to one note event, decoupled from duration (which is not considered by the model). An event, X_t, can be any of the

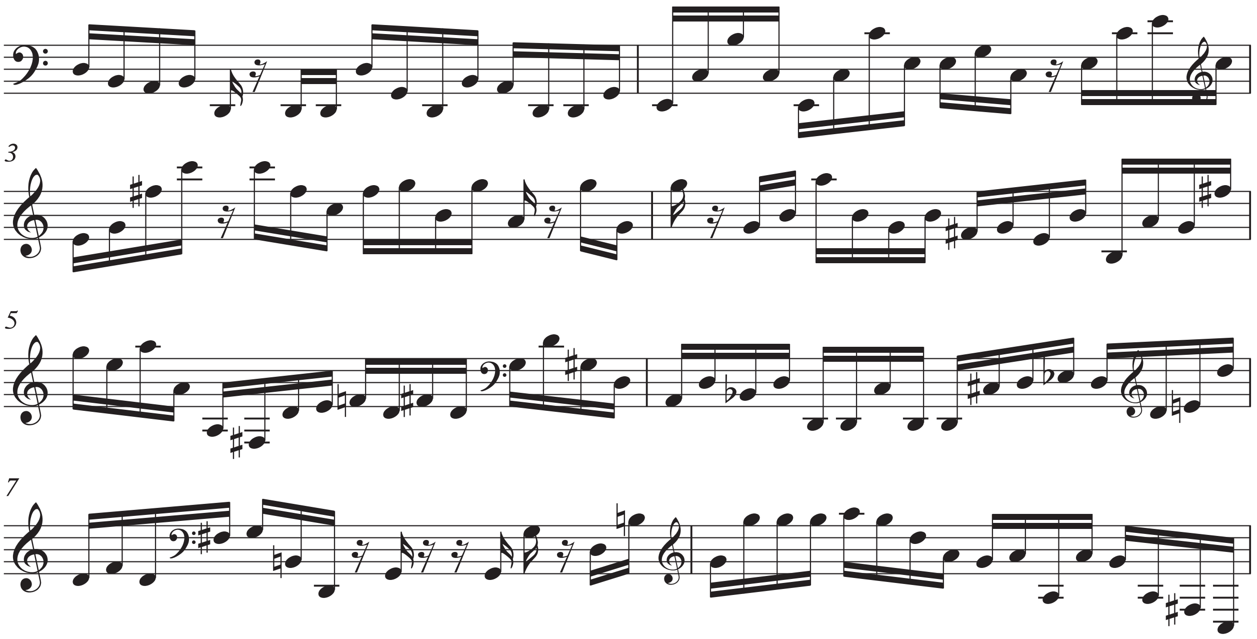

Figure 3: Example musical segment generated by the DBN showing harmonic movement typical of the training exemplar.

48 chromatic pitches (in 4 octaves) or a rest. Thus our generated data will be an ordered sequence of integer note events ($X_0...X_t$).

In order for the ART to classify on short patterns the token sequence for each node is transformed into a fixed-length vector employing spatial encoding (Gjerdingen 1990; Smith and Garnett 2011; 2012b). Modeled on theories of human attention and short-term memory function, spatial encoding converts a token sequence (of pitch-classes, for example) through a simple neural-net to a vector representing the recency of each token-feature. The spatial encoding net employs a single node layer with one node for each token in the feature set, which is fully activated when that token is presented, simultaneously damping all the other nodes. The implications of this encoding model are that sequences such as ABCD and ACBD (or even ACD or AD) may be efficiently compared and a relative distance measure calculated.

Evaluation

Figure 3 shows the efficacy of the system in generating musical passages, here trained on Bach's *Prelude from the Suite No. 1 for unaccompanied Violoncello* (as are all the examples included herein). This example, which is drawn from a much longer "piece," displays references to the training material, employing the broken chord figuration used throughout the source and opening chord sequence of the original Bach (in this example: G-major in the first bar moving to C-major, then a D7 in bar 3 resolving to a G-major in the middle of the bar). The remainder of this example displays chordal relationships that have been learned from the training piece, employing harmonic motion that is typical of the Baroque era. Additionally, the chromatic ascent with alternating Ds seen in bars 5–7 is a close reference to the training work. From this, as a typical exemplar of the system's output, we conclude that the work is capable of learning and reproducing formal dependencies of a longer than note-to-note term duration.

The starting point for qualifying and quantifying the limits of the proposed system is to examine the generative capabilities in terms of range of musical output. In addition to an overall evaluation of the system in terms of creative output, the specific behavior of the neural network models is effected through several parameters that can have an impact on the generated material. In the following section we present and compare musical sections resulting from different ranges of parameters towards understanding their effects in qualitative and aesthetic fashions.

A Monte Carlo approach is employed to actualize the DBN design through the use of a typical particle filter. One significant deviation from the standard particle filter model is the absence of an observation/update step in the algorithm. This results from the system tracking what we propose as a creative musical sequence that is purely derived from the algorithms and DAG structure, without any externally appreciable "observations." The particle filter predicts transitions and the reward function (eq. 3) described previously serves to assign weights to each particle for resampling.

One final design consideration is the translation of token sequences into feature vectors for categorization by the ART neural networks. The continuous spatial encoding model exposes two parameters that have impact on the system's generative possibilities: the attention weight assigned to new tokens and the decay rate of the encoder (i.e. how quickly it

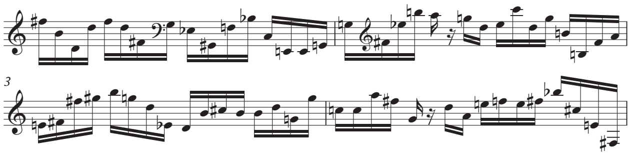

Figure 4: Example generation with a vigilance of $v = 0.5$.

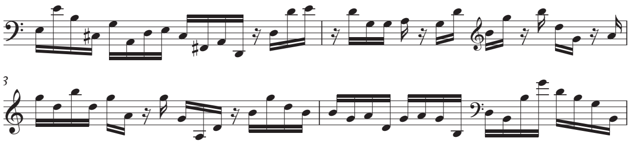

Figure 5: Vigilance of $v = 0.9$.

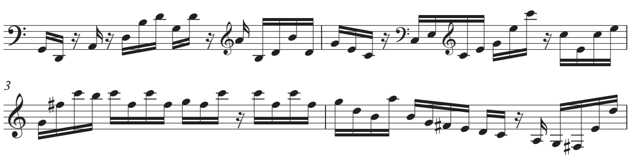

Figure 6: Vigilance of $v = 0.99$.

forgets, determining how many tokens in the sequence are represented at a given time step).

The parameters under consideration include:

- the ART network's vigilance parameter v,
- the ART network's learning rate l,
- the spatial encoders' activation amount,
- the spatial encoders' decay rate,
- the number of nodes stacked in the DAG ($B...N$).

The probabilistic nature of the DBN leads to different musical results with every execution. This makes direct, quantitative comparisons between different parameter settings difficult. However, the examples shown here are chosen as representative of the material produced with the given settings under examination.

The vigilance parameter controls the size of each category identified by the ART. Larger values cause the ART to be more discriminating, resulting in many more categories identified for a given token sequence. Lower values result in many larger, more general categories. Unless otherwise stated the generated examples employ a DBN configured: 8 nodes stacked above A, a vigilance of $v = 0.975$, a learning rate of $l = 1$, spatial encoder activation: 0.25, decay: 0.85714286. Note that while the ART algorithm specifies a range for vigilance $v = [0, 1]$, practical limits are imposed by the particle filter. One one end low vigilances flatten the PDF to such an extent that all predictions are accepted (in our implementation $v = 0.8$ was found as a practical lower bound). On the other end extreme high vigilance settings result in over-fitting of the training data denying any original material generation.

The musical example in figure 4, with a lower vigilance, fails to show any musical direction or structure. While there is constant variation and intervallic leaps the harmonic movement of the previous figure is lacking. Figure 5, with an intermediate vigilance setting, exhibits more structure, hinting at an A-major to D-major to G-major harmonic sequence in the first two bars. The highest level of vigilance, fig. 6, shows even more focus and direction, mimicking the training piece more closely. Not only does the chord progression (G to C to D) from the source show strongly but the timing (one chord per bar) is reproduced as well. These three examples (fig. 4–6) fit the model's implications that higher vigilance settings should allow more discrimination in the DBN (coming closer and closer to reproducing the source exactly).

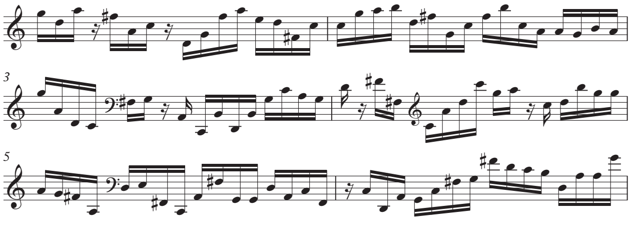

Figure 7: Generation with a learning rate of $l = 0.01$.

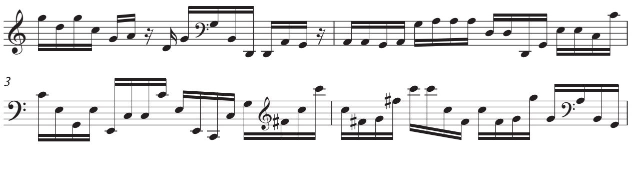

Figure 8: With spatial encoding activation of 1 and decay of 0.85714286.

The ART network's learning rate is aptly named as it controls how quickly the neural network nodes change weights to incorporate new patterns. This directly impacts how much learning residual results from the presentation of new patterns, but can mimic an attention span (in a sense), allowing the system to repeat patterns more often (lower learning rate), or encourage more rapid exploration of new material (higher learning rates). Figures 3–6 are the result of higher learning rates and fig. 7 is supplied to show the contrast of a lower setting.

The low learning rate employed in fig. 7 results in a passage that moves back and forth between hints of a D-major harmony and resolving to G-major. In one sense the DBN can be seen as more patient, or focused on smaller details, content to repeat the same segments many times. However, the motion lacks the clarity and structural movement of material generated with higher learning rates (especially figs. 3 and 6).

The impact of the short-term memory configuration is exposed through the parameters of the spatial encoding model. The two operative variables work in conjunction to control the length of pattern sequence that the ART will encode. Figures 3–7 were generated with an activation of 0.25 and a decay of 0.85714286 (indicating that a single presentation of a token will remain in the encoder for at least seven time steps, and the encoder will distinguish between multiple pre-

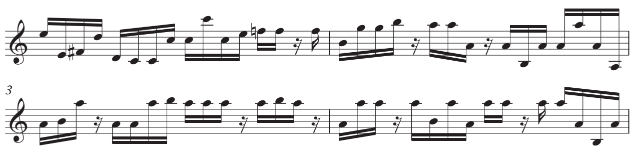

Figure 9: With spatial encoding activation of 1 and decay of 0.5.

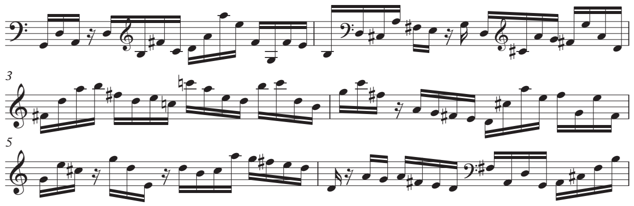

Figure 10: With 1 upper layer.

sentations of a token before over saturation). Figure 8 shows an excerpt generated with an activation of 1 and a decay of 0.85714286 (retaining a single presentation for seven steps). Figure 9 results from an activation of 1 and a decay of 0.5 (retaining a single presentation for 2 steps).

As with several previous examples, fig. 8 repeats the same opening chord progression (G in bar 1 through an ambiguous bar to C, the D7 in bar 4 with a final resolution to G). Examining many examples (not reproduced here) indicates the conclusion that the activation amount has an impact on direct repetitions of pitches (fig. 8 shows more directly repeated notes than previous examples), but otherwise has limited effect on the material.

The excerpt in fig. 9 only hints at the training material, instead becoming centrally fixated on the pitch class A. After 1000 notes this excerpt was still repeating pitch classes A and B, with varying rhythmic and octave placement as shown here. Examples generated with these spatial encoder settings exhibit the same characteristics, yet the fixation settles on different pitches. From this we can conclude that the shorter retention in the feature spatial encoding denies the structural abstraction that the DBN is aimed at (presumably as a result of the ARTs inability to adequately classify the resulting abbreviated note sequence chunks).

The final consideration is the number of node layers to employ (nodes $B...N$ in the DBN). Unfortunately, musical material generated with anywhere from 1 to 512 upper layers do not present immediately transparent differences. Figure 10 shows material representative of the results with only 1 upper layer (i.e. just the DAG shown in fig. 2). This example shows local-level figuration that can be seen as reminiscent of the Bach and includes hints of harmonic movement between A-major and D-major (or D7). However, as with the low learning rate example (fig. 7), the movement appears aimless after a few bars indicating a lack of longer-time scale direction. This appears to fit with the hypothesis that more layers results in the generation of longer term dependencies, but is not conclusive.

Future Directions and Conclusions

The design of the DBN employed suggests additions and extensions to expand the musical output capacities of the system. One direction is to incorporate contextual constraints that allow the system to track a creative musical path through externally dictated musical styles. For example, rules governing harmonic movement could be imposed in the probability distributions employed by the predictive algorithm to encourage results that more overtly obey the rules of harmonic-based musical styles. These rules could eventually be learned through the examination of exemplars drawn from a corpus of target works (such as through a hidden-markov model construction).

A similar extension is the incorporation of new nodes in the DAG that address harmonic implications, rhythm, tempo, meter and other conventions of western classical music. Many of these may be distinct DBNs nested within the larger model, providing the ability to track many distinctive elements of different musical styles. Once in place each sub-DBN may function with imposed PDF constraints learned from a stylistically targeted corpus.

The opportunity for an observation step in the particle filter operation hints at a touching point for an interactive model. Functioning in conjunction with a human musician (or other agent) the generative agent could accept musical input and attempt to converge on the same musical sequence. The interactive loop is completed through the musician accepting the DBN-based agent's output as suggestion for what to play next. With a human improvisor in the loop the agent's output can either be taken as a score to closely follow or as suggestion, should the performer desire external stimulation. The metaphor of path and landscape implicates the performer as driver who takes input from a navigator (the generative agent) as to which musical direction might be most rewarding. With a rich enough training set the DBN may be able to accurately imitate a given improviser's style.

The initial evaluations contained herein show the success of the implementation in realizing the DBN and the efficacy of this novel system design towards generating a type of pastiche based on a training corpus. While the examples provided in this context are necessarily limited they indicate a varied patterning that is a hallmark of novel, inventive improvisations.

References

Boden, M. 2004. *The creative mind: Myths and mechanisms*. Psychology Press.

Carpenter, G.; Grossberg, S.; and Rosen, D. 1991. Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system. *Neural networks* 4(6):759–771.

Eck, D., and Schmidhuber, J. 2002. A first look at music composition using lstm recurrent neural networks. *Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale*.

Gjerdingen, R. 1990. Categorization of musical patterns by self-organizing neuronlike networks. *Music Perception* 339–369.

Hochreiter, S.; Bengio, Y.; Frasconi, P.; and Schmidhuber, J. 2001. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies.

Mozer, M. 1999. Neural network music composition by prediction: Exploring the benefits of psychoacoustic constraints and multiscale processing. *Musical Networks: Parallel Distributed Perception and Performance* 227.

Murphy, K. 2002. *Dynamic bayesian networks: representation, inference and learning*. Ph.D. Dissertation, University of California.

Schmidhuber, J. 2009. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Anticipatory Behavior in Adaptive Learning Systems* 48–76.

Smith, B., and Garnett, G. 2011. The self-supervising machine. In *Proc. New Interfaces for Musical Expression*.

Smith, B., and Garnett, G. 2012a. Improvising musical structure with hierarchical neural nets. In *Proc. AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*.

Smith, B., and Garnett, G. 2012b. Reinforcement learning and the creative, automated music improviser. In *Proc. EvoMUSART*.