Reinforcement Learning and the Creative, Automated Music Improviser

Benjamin D. Smith and Guy E. Garnett

University of Illinois at Urbana-Champaign, United States

Abstract. Automated creativity, giving a machine the ability to originate meaningful new concepts and ideas, is a significant challenge. Machine learning models make advances in this direction but are typically limited to reproducing already known material. Self-motivated reinforcement learning models present new possibilities in computational creativity, conceptually mimicking human learning to enable automated discovery of interesting or surprising patterns. This work describes a musical intrinsically motivated reinforcement learning model, built on adaptive resonance theory algorithms, towards the goal of producing humanly valuable creative music. The capabilities of the prototype system are examined through a series of short, promising compositions, revealing an extreme sensitivity to feature selection and parameter settings, and the need for further development of hierarchical models.

Keywords: Computational creativity, machine learning, music, composition, reinforcement learning, adaptive resonance theory.

1 Introduction

Can beauty be measured? Can innovation and inspiration be quantified? What elements inform creativity and can they be identified? Questions such as these remain largely unanswered across every academic discipline, yet solutions could have wide reaching, transformative implications. Reliably identifying innovation could lead to more productive funding allocations, commissions, grant awards, and ultimately a more productive citizenry. Measures of aesthetic beauty, while conceptually problematic, could provide additional methods for analyzing art, music, and dance. Understandings of creativity could lead to new models of artificial intelligence, enabling ever more responsive, interactive, and natural computer systems. Algorithmic computer music, seeking to locate the ultimate pure music, could eventually realize a vision of self-organizing creative art.

Machine learning algorithms present one promising line of inquiry towards the duplication and/or extension of human intelligence in computers. Such applications, often formulated around models of biological brain function, seek to understand observed phenomena through pattern abstraction and reduction, giving the machine access to underlying conceptual models and relationships. However, these systems typically lack the capability to truly extemporize or

P. Machado, J. Romero, and A. Carballal (Eds.): EvoMUSART 2012, LNCS 7247, pp. 223–234, 2012.

improvise and have no motivation to pursue the mapping of new patterns or concepts.

Reinforcement Learning, a specialized machine learning model that enables automated exploration and discovery, presents possibilities in computational creativity. An enhanced, intrinsically motivated reinforcement learner (RL) [9] can actively create and/or discover novel or surprising patterns, arguably explaining some essential aspects of intelligence and aesthetics through the process. The RL is able to evaluate its discoveries within a given context, encouraging the agent to identify new patterns and new algorithms that enable a better mapping (i.e. understanding) of the agent's environment.

We describe herein the design and analysis of a novel intrinsically motivated RL prototype created to improvise and analyze music, examining the computational potential for generating compelling, creative music through this novel application. Employing unsupervised machine learning algorithms, this RL is based on theories of human perception and cognition, intentionally modeling human creativity and inspiration at a fundamental level. This work is presented as a process of creation and evaluation in the form of a composition lesson with the artificial, RL composer.

2 Motivation

2.1 Reinforcement Learning

The basic premise of the intrinsically motivated RL model is characterized by Schmidhuber [9], based on learning processes observed in human development. While humans readily learn from sensory stimuli and their environment, avoiding heat, injury, hunger, and thirst, they take more than a passive role in this process. Babies actively conduct experiments of the nature "what sensory feedback do I get if I move my eyes or my fingers... just like that?" In this way the individual is always seeking new effects that exhibit "some yet unexplained but *easily learnable* regularity" (as also described by the Wundt curve). Stimuli observed previously is quickly deemed boring, while entirely new input is regarded as incomprehensible noise. Through this gradual mapping of behaviors and patterns the learner gradually acquires more and more complex behaviors, eventually leading to the extreme examples of academic thought, scientific innovation, and aesthetic inspiration.

A simple algorithmic mechanism is proposed by Schmidhuber to explain this learning phenomena, which uses RL to maximize the "internal joy" of the discovery of "novel patterns." Patterns can be understood as regularities in a dataset that can be abstracted in some fashion and effectively reduced, in complexity or size, as a result (i.e. data compression). When an agent discovers a regularity or a new model that allows phenomena to be compressed, the pattern is deemed temporarily *interesting* or *surprising.* Thus a measure of this learning, i.e. at any given moment how much new data is being compressed (or *understood*), can be calculated and used to drive a process of discovery whereby the agent takes an active role in seeking out new phenomena to analyze. Attempting to maximize the efficiency of the compression model (termed *reward*), both in terms of performance (speed and processing load) and product (how much the dataset is reduced), motivates the agent to seek out surprising situations.

The crucial ingredients of an intrinsically motivated RL implementation are:

1. An adaptive world model, essentially a predictor or compressor of the continually growing history of actions/events/sensory inputs, reflecting what is currently known about how the world works,
2. A learning algorithm that continually improves the model (detecting novel, initially surprising spatiotemporal patterns that subsequently become known patterns),
3. Intrinsic rewards measuring the model's improvements (first derivative of the learning progress) due to the learning algorithm (thus measuring the *degree* of subjective surprise or fun),
4. A separate reward optimizer or reinforcement learner, which translates those rewards into action sequences or behaviors expected to optimize future reward.

The general model that best satisfies these requirements is not calculable in finite time. However, simplifications and non-general implementations can both prove revelatory about RL theory and provide compelling examinations of the human behaviors being modeled. The later is a significant objective of this work.

Going back to Newell et al. [7], this model exhibits several theoretical aspects of creativity and problem solving. By definition the learning involves novelty and intrinsic, or personal value (Boden's P-creativity, [1]), in (3) and (4). Further, the exploration requires modification of previously learned material, through (1), and ideally requires continual reaffirmation or rejection of concepts, in (2). However, insuring global significance (H-creativity) requires context that the model does not explicitly contain, and it also demands a relevant selection of features to define the agent's environment. Ideally this model will produce music that is deemed creative in human terms [5], which may be evaluated but yet necessitate research beyond the current scope.

Building a musical RL requires 1) a model of music, a predictor/compressor containing all the music heard by the agent and containing everything that is "known" about music; 2) an algorithm that learns how music works (improves the model in 1); 3) a reward measure of the model's improvements in (2); and 4) an agent that creates more music anticipating maximal future reward. In other words, (1) is an analysis of all the music presented to the agent and (2) is the set of working theories that explain these analyses. Acting on these theories takes place through (3) and (4).

For example, in functional-harmony tonal theory a Roman Numeral chord analysis of a chorale provides a compressed representation of the piece, reflecting some aspects of patterns and regularities observed in the music. In fact the concept of *style*, of which chord analysis may be considered an informing example, is effectively a form of compression, allowing the description and categorization of large collections of music according to some similarities exhibited in exemplar cases.

To implement (1) and (2) above we employ Adaptive Resonance Theory (ART) [2], an unsupervised machine learning model mimicking elements of human cognition and perception. (3) is understood as a measure of the *relative entropy* between the ART's prior and posterior states, and (4) is implemented as a comprehensive predictor that anticipates the reward measure for every potential stimuli. These algorithms are efficient enough to run in real-time on pro-sumer grade hardware, enabling live testing and performance with the final application.

In order to verify the applicability of this model and design we restrict *music* to a containable space. Our model processes monophonic pitch within four octaves, C2 to B5, treating elements of dynamic, rhythm, and timbre as uniform. Further, common practice concepts such as meter and harmony are not input explicitly. While these constraints may deny most human musical expression, still many cases fit within this paradigm (folk musics, some solo string works by J.S. Bach, and many examples of minimalism, as an example).

2.2 Adaptive Resonance Theory

The ART is a self-organizing neural network model developed by Carpenter et al. [2], which enables efficient, online classification and categorization of data vectors. The theory is based on understandings of human cognition and the ability to distinguish different categories of stimuli from a continuous data stream. ART implementations have been used in music analysis previously to analyze tonal music [4], auto discovering patterns of harmony, consonance, and dissonance [8]. Interactive applications are also being developed employing ARTs to enable improvisatory performance with multi-media systems [10].

The basic functionality examines a series of feature vectors, extracted from a stream of data (i.e. music), and locates distinctive categories within these features. One ready analogy is to that of theoretical concepts such as *chord, major* and *minor, motive,* and *phrase.* Given a sequence of musical features the ART can locate similar concepts, based on mathematical distance (or "resonance") calculations.

When presented with a new input vector the ART algorithm first obtains a resonance measure through the comparison of each known category with the new input.

$$T_j(\mathbf{I}) = \frac{|\mathbf{I} \wedge \mathbf{w}_j|}{\gamma + |\mathbf{w}_j|} \tag{1}$$

For a given input $\mathbf{I}$ the resonance measure is calculated with choice function T, comparing the input with the adaptive weights $\mathbf{w}$ of each category j. A choice parameter γ affects the matching of inputs to the closest subset category, and is typically set close to 0 to achieve this. The fuzzy AND operator $\wedge$ is defined by

$$(\mathbf{x} \wedge \mathbf{y}) = min(x_i, y_i) \tag{2}$$

and the norm $| \bullet |$ is the L1 norm

$$|\mathbf{x}| = \sum_{i=1} |x_i| \tag{3}$$

If by incorporating the new input the strongest resonating node remains within a preset limit (or "vigilance") it is selected and allowed to learn based on the input. On the other hand if by incorporating the new input the category size (in feature space) would increase beyond this limit p, then this node is rejected for this iteration and the next most resonant node is considered,

$$\frac{|\mathbf{I} \wedge \mathbf{w_j}|}{|\mathbf{I}|} < p \tag{4}$$

The rejection of all existing category nodes results in the creation and training of a new category node. Each feature vector is complement coded before it is examined, through the concatenation of I:

$$(\mathbf{I}) = (I_i, 1 - I_i) \tag{5}$$

Thus the fuzzy AND operator effectively encodes the limits of each category, comparing the minimum and maximum values for each element separately. The details of the ART algorithm are described at length by Carpenter et al [1].

Another parameter with operative ramifications is the "learning rate" β of the ART network. This parameter allows the network to both train new inputs immediately (setting $\beta = 1$ for category creation) and still adapt slowly, retaining the identity of older categories. Setting the learning rate high causes categories to expand and fully incorporate new inputs while setting it low causes the categories to adjust slowly, settling into an average area of the feature space:

$$\mathbf{w}' = \beta(\mathbf{w} \wedge \mathbf{I}) + (1 - \beta)\mathbf{w} \tag{6}$$

2.3 Feature Encoding

The choice of input features has a fundamental impact on the machine's formulation of what music is and how it works. As a result of the novelty seeking proclivities of the system any new, comprehensible pattern in the feature vector will be rewarded, satisfying definitive requirements of personal creative value [1]. However, if these patterns do not map to human perception the result will be rejected and deemed boring by human observers. Thus the choice of pertinent features becomes key in affording meaningful creative exploration.

At this point it is not feasible (or possible) to fully simulate human audition such that an abstract learning agent can access musical abstractions from audio signals. However, through careful preprocessing this challenge is approachable. Our objective here is to transform sonic stimuli into relevant elemental components of music, such that the machine can relate these elements in meaningful ways and generate globally interesting music.

The reduction employed in this design is based on pitch and pitch classes (MIDI pitch modulo 12, or the letter names of each pitch). The pitch class of each incoming note is recorded, as is the interval between each input and the previous note. The interval is confined to a two octave range, from 12 half-steps down to 12 half-steps up, with larger intervals being clamped at 12. The direction of the interval (i.e. its sign) is further recorded as either down, unison, or up. The octave that the note is in (pitch / 12, rounded down) is stored, as is the size of the interval in octaves (interval / 12, rounded down). Finally, the interval (within an octave) and direction are recorded together in a sliding window, retaining the previous seven inputs. Thus the ART sees the following feature vector: With the exception of the interval and direction time window all of the parameters are encoded using a *spatial encoding* technique [3][4]. This is a simple neural network model that is based on the premise of sensory, attentional stimuli. As a given token is presented to the network a corresponding neuron (node) is activated. This simultaneously causes the other nodes in the network to be suppressed, effectively encoding the sequence of tokens as the activation levels of the network's nodes. The calculation is simply:

pitch class	interval	interval+direction window	direction sign	octave	interval octaves

Fig. 1. Input feature vector layout

$$\mathbf{x_t} = Ax_{t-1} \vee \mathbf{u_t} \tag{7}$$

where $\mathbf{x}$ is the network's state, A is the attenuation vector, and u is the input stimuli vector. The fuzzy MAX operator $\vee$ enables the excitation of individual nodes while allowing the network as a whole to forget (attenuate),

$$(\mathbf{x} \vee \mathbf{y}) = max(x_i, y_i) \tag{8}$$

The sliding window of interval and direction can be viewed as a list of the seven most recent inputs (with $A \approx 0.143$) . As a new input is observed it is recorded at the top of the list and the oldest item is removed from the bottom.

2.4 Reward

Over the course of processing and learning from a new input feature vector the ART's internal model changes, incorporating the new data or forming a new category. This change can be measured precisely and is termed *relative entropy.* Since only one encoded category is allowed to change during any given presentation we calculate the RL's intrinsic reward R as the magnitude of change between the state prior to the new input observation $\mathbf{w_{t-1}}$ and the resulting changed state $\mathbf{w_t}$,

$$R = \sqrt{\sum_i (w_t^i - w_{t-1}^i)^2} \tag{9}$$

This presents an irregular problem when a new category is encoded. For functional purposes empty categories are stored as undefined (all elements initialized to 1) and thus measuring the amount of change using (9) is generic and not useful. However, the encoding of new categories is a direct function of the vigilance parameter (set low enough it would enable one category to encompass all inputs) and if we consider a slightly wider vigilance setting we can obtain a useful reward measure. Thus we take the residual that would result from the expansion of the best matching category (i.e. closest, in feature space), despite vigilance verification. This allows any input to produce a measurable reward value.

Because the ART will readily incorporate any new input as a new category (or expansions of existing ones) we transform R to locate the maximal point on the Wundt curve (i.e. the most interesting and rewarding point between boredom and confusion). This is accomplished by determining a tolerance m for similarity (i.e. how quickly repetition becomes boring, controlling the width of the curve) and inverting R around this threshold. We also define a minimum threshold (typically 0.001) below which R is clamped to 0, to reduce floating-point error and drift.

$$R' = \frac{m}{|R - m| + m} \tag{10}$$

Thus inputs that produce insignificant changes, because they have been observed before, are deemed not rewarding and hence boring. Inputs that result in a change are progressively less rewarding the more chaotic they are and the less they match patterns that are already known. The RL's prediction phase will always select the most rewarding next stimuli R', and through (10) can ensure that the most satisfying choice is always made.

2.5 Prediction

By defining our problem in a limited domain space (i.e. confined pitch space and regular rhythm, at least initially) it becomes possible to evaluate every possible future input, for a limited number of steps, in a finite amount of time. At each time step we consider forty-eight possibilities (the chromatic pitches between C2 and B5) and prediction involves calculating the reward that would result from observing each of these possibilities. In standard operation our model produces the maximally rewarding choice, is allowed to play the associated pitch (sounded on a synthesizer for human audition), and observe the result (i.e. process the pitch as the next input). This results, theoretically, in the exploration of every local maxima but because these points loose value with each presentation (become more boring) the agent is never stuck.

3 Composition Lesson

Initially the RL is given a starting pitch (middle C4) and allowed to improvise, seeking to maximize its internal reward structure. For purposes of clarity and explication the RL is only allowed to play notes of a single duration, and is constrained to the range of C2 to B5. In effect it starts with a creative agenda (internal model), an instrument of sorts, and no context or fore-knowledge. Figure 1 depicts the start of this improvisation, as the RL begins to explore

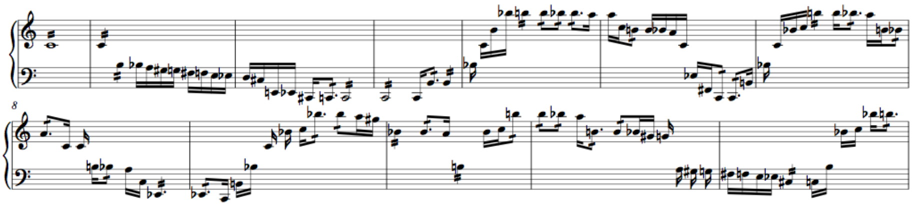

Fig. 2. Beginning of an improvisation with learning rate of 0.1, vigilance of 0.925, and maximal residual of 0.025

the available musical space. After repeating the initial pitch (C4) twelve times, it starts a chromatic descent, sounding all the available pitch classes in one octave. However, rather than arrive at C3 (which is perhaps too expected, or boring) it skips a major sixth down and descends from there to the lower limit, C2. A bar on this lowest pitch leads to a series of ascending and descending sweeps featuring combinations of major sevenths and half-steps, with a few major sixths along the way. The third descending sweep is interrupted with a return to the peak, B5 in bar 10, before an extended descent commences covering three octaves, followed by a final ascent.

Figure 2 was created with a learning rate β of 0.1, a vigilance setting p of 0.925, and a maximal residual m of 0.025. Decreasing the learning rate (to 0.01) allows the RL to find more interest in repetition (fig. 3), extending the initial Cs, in this case, before beginning the descending chromatic scale as before. However, here the scale is broken after three steps, causing a leap down to the next C, repeating the same descending pattern. Now an extended exploration of small intervals filled with chromatic steps begins, anchored on C2.

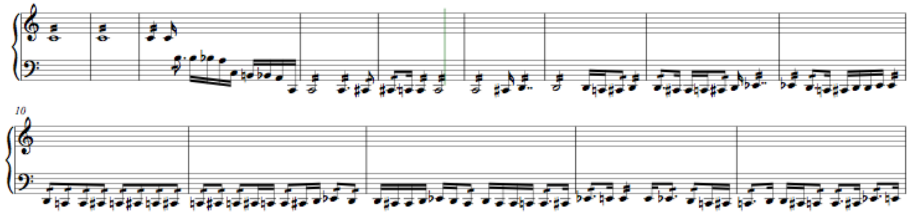

Fig. 3. Beginning of an improvisation with learning rate of 0.01

Increasing the maximal residual ($m = 0.2$) forces the agent to more aggressively discover new patterns (fig. 4). Now the descending scale starts sooner, but spends more time on each individual pitch during the two octave span (this is due to the increased tolerance for more complex patterns, as produced by the irregular repeating rhythms). These temporal extensions continue as the former ascending sweep is fleshed out with more step-wise motion, followed by a similar treatment of the descent.

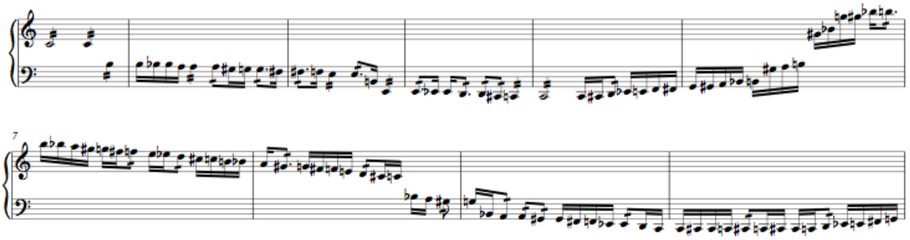

Fig. 4. Dual-layer improvisation

Already it can be seen that the precise parameter settings have a significant impact. The parameters values employed here were chosen as examples, however further treatment and refinement of precise values would be necessary to produce a given, aesthetically desirable result.

The music created thus far presents a relatively continuous exploration of pitch space, without any apparent global direction. While the surface texture presents elements of patterning and variations that are quickly perceptible, there is apparently no higher level organization. This results from the computational model described thus far, which only looks one note ahead and only desires small increments of change. What if the machine is given a notion of hierarchy and structure resulting from motivic and phrase relationships within its improvisation?

We address this need by adding two more layers of ARTs that encode and learn from sequences of categories identified by the underlying ARTs (inspired by the dual layered model in [4]). Just as the initial ART encodes and trains on pitches and pitch classes, the middle layer ART takes the identified category index of the first ART, encodes it in a short-term memory and learns to categorize this input. The third ART does the same for the second layer ART's output. In musical terms we now have motives (first layer), phrases (second layer), and phrase groups (third layer). Figure 5 depicts the improvisation with the first two layers active.

Immediately, we are struck by the directional reversal of this piece. Instead of choosing to descend, it now chooses a chromatic ascent, skipping a major sixth in the middle, before arriving at the high B. Otherwise it is similar in construction to that in fig. 2, with alternating ascents and descents. The start of bar 8 presents

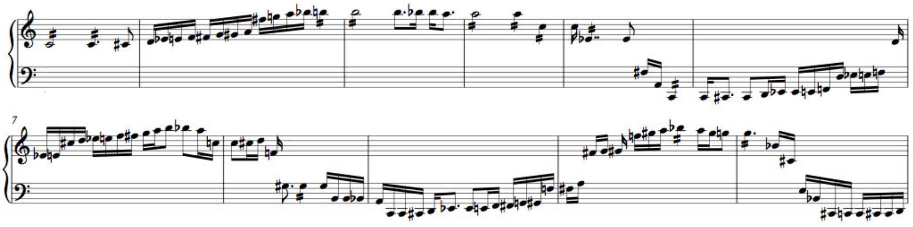

Fig. 5. Tri-layer improvisation

a nice variation, as the overall descent is filled with a brief climb. Also, the leap of a sixth, from bar 2, becomes an important feature in the foreshortened descents in bars 4-5, 7-8, and 11. Unlike the previous figures there is apparently more internal consistency, with more intervalic, pitch, and shape cohesion.

Adding a third layer of abstraction produces more refinement (see fig. 6). The start is identical to the previous figure, since the third layer apparently only determines macro level relationships. These are quickly seen in bar 8, where the descending sixths are broken by a return to the ascending pattern, once with a leap up and the second more fully fleshed out. The arrival on the high B is denied at the end of bar 9, resolving the neighbor-group of G-sharp and B-flat on the A, which precipitates a diminished-chord arpeggio, downward.

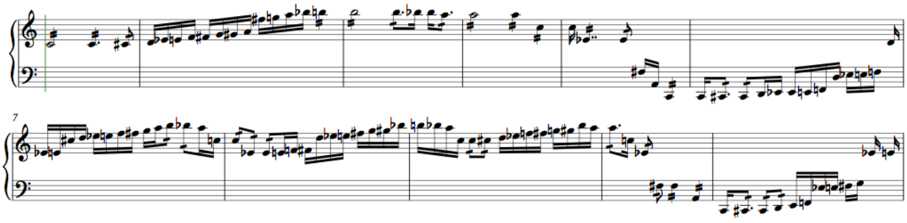

Fig. 6. Beginning of an improvisation with learning rate of 0.01

Now we have shown that some elements that inform musical composition and improvisation are intentionally (on the RL's part) present in the generated musical examples. However, what the RL produces is more akin to a baby's pre-lingual vocalizations, since it has only the tools to create music, a desire to do so, but no stylistic context or previous examples. Building the capability to fully process exemplary pieces, and generate extrinsic reward values based on how well the RL matches, is beyond the scope of this text. However, seeding the RL's memory with precomposed works is a ready next step.

The final example from our RL takes the *Prelude* from the first of J.S. Bach's six unaccompanied cello suites and fills the ARTs with the resulting categories. Then the RL is allowed to continue, seeking variations and developments, incrementally expanding on the given exemplars. Figure 6 depicts a passage improvised by the RL after listening to the Bach.

Figuration from the original is in evidence (broken chords, scaler patterns in the second and third bars, brief pedal points with alternating notes in the last three bars), although notions of functional harmony and strict motivic repetition, which are predominant in the Bach, are largely absent. Yet, the ability to take given patterns and expand upon them is seen in a sort of mechanical *stream of consciousness* fashion. Certainly the texture is far richer than the previous examples.

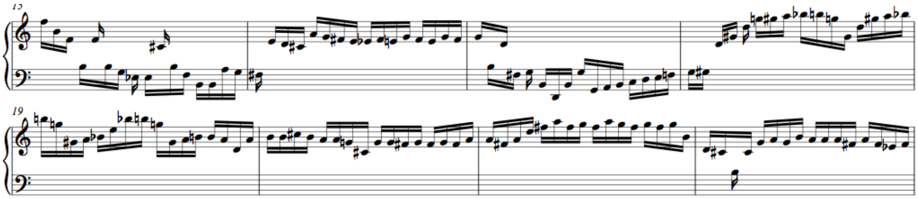

Fig. 7. Improvisation after listening to J.S. Bach's *Prelude* from the first suite for unaccompanied cello

4 Discussion

While capable of arguably creative generation in a limited fashion this prototype denies many aspects of music, which remain to be incorporated into this preliminary model. While it is capable of creating varied and potentially interesting musical lines, these are currently lacking in rhythmic variation, polyphony, dynamics, and articulation, to name a few. Along with rhythm the RL needs to develop a sense of metric patterning and conventions for treatment within its operating context or style. A larger sense of direction, goals, expectation development, and structure is similarly lacking. This may be addressed in a more complex, hierarchical system in which the ART-based RL plays a core role, or it may be a result of different feature choices. While the former would argue that music creation demands stylistic context and extensive exposure and training, the latter would present a strong case for the self-organizing properties of musical materials.

During the design process many parameter variants were considered and briefly evaluated. Most of these produced results that were distinct, but primarily in terms of aesthetics (for example, different learning rates cause it to focus on broken octaves while some focus on scale patterns). Our conclusion is that the appropriate parameters may have to be learned from an existing corpus of music, or deduced dynamically based on some other metric (such as audience engagement or a human ensemble's rate of change). Alternatively these parameter settings may be predetermined towards compositional goals.

This increase in complexity will require a more dynamic prediction model, possibly employing dynamic bayesian networks [6] and particle filtering. These machine learning techniques have been shown to perform well for systems with mixed discrete and continuous state variables (such as pitch versus dynamics)[11].

As the RL improvises its memory expands, requiring more and more computational resources. On a high end consumer desktop, with eight CPUs, the system reaches 50% load after twenty minutes of operation, and will begin to slow down sometime thereafter. Further optimizations, in implementation, or larger computational systems are required to enable a RL that could run persistently, learning over the course of years, as humans do. Alternatively, a model of forgetfulness may prove profitable, allowing the machine to destructively compress old, well understood, or less valuable categories and data.

We have demonstrated that a RL model can produce musical lines that betray interesting developments, variations and perhaps even *creativity.* The simplicity of the musical context under examination has provided a focused space in which to expose the functionality of the implementation and begin to explore the aesthetic implications resulting from the various parameters and settings. This introductory look at RL in a musical application displays proof of concept and promises more advances in automated creativity, perhaps leading one day to the creation of an aesthetically informed, automated, musical intelligence.

References

1. Boden, M.: The Creative Mind: Myths and Mechanisms, 2nd edn. Routledge, London (2004)
2. Carpenter, G.A., Grossberg, S., Rosen, D.B.: Fuzzy ART: Fast Stable Learning and Categorization of Analog Patterns by an Adaptive Resonance System. Neural Networks 4, 759–771 (1991)
3. Davis, C.J., Bowers, J.S.: Contrasting five different theories of letter position coding: Evidence from orthographic similarity effects. Journal of Experimental Psychology: Human Perception and Performance 32(3), 535–557 (2006)
4. Gjerdingen, R.O.: Categorization of musical patterns by self-organizing neuronlike networks. Musical Perception (1990)
5. McCormack, J.: Open problems in evolutionary music and art. Applications of Evolutionary Computing, 428–436 (2005)
6. Murphy, K.P.: Dynamic Bayesian Networks: Representation, Inference and Learning. Ph.D. in Computer Science. University of California, Berkeley (2002)
7. Newell, A., Shaw, J.G., Simon, H.A.: The process of creative thinking. In: Gruber, H.E., Terrell, G., Wertheimer, M. (eds.) Contemporary Approaches to Creative Thinking, pp. 63–119. Atherton, New York (1963)
8. Pearce, M. T. The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition. PhD. City University, London (2005)
9. Schmidhuber, J.: 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 (2009)
10. Smith, B.D., Garnett, G.E.: The Self-Supervising Machine. In: Proceedings of NIME 2011, Oslo, Norway (2011)
11. Swaminathan, D.: A Dynamic Bayesian Approach to Computational Laban Shape Quality Analysis. Advances in Human-Computer Interaction (2009)