TOWARDS ADAPTIVE MUSIC GENERATION BY REINFORCEMENT LEARNING OF MUSICAL TENSION

Sylvain Le Groux
SPECS
Universitat Pompeu Fabra
sylvain.legroux@upf.edu

Paul F.M.J. Verschure
SPECS and ICREA
Universitat Pompeu Fabra
paul.verschure@upf.edu

ABSTRACT

Although music is often defined as the "language of emotion", the exact nature of the relationship between musical parameters and the emotional response of the listener remains an open question. Whereas traditional psychological research usually focuses on an analytical approach, involving the rating of static sounds or preexisting musical pieces, we propose a synthetic approach based on a novel adaptive interactive music system controlled by an autonomous reinforcement learning agent. Preliminary results suggest an autonomous mapping from musical parameters (such as tempo, articulation and dynamics) to the perception of tension is possible. This paves the way for interesting applications in music therapy, interactive gaming, and physiologically-based musical instruments.

1. INTRODUCTION

Music is generally admitted to be a powerful carrier of emotion or mood regulator, and various studies have addressed the effect of specific musical parameters on emotional states [1, 2, 3, 4, 5, 6]. Although many different self-report, physiological and observational means have been used, in most of the cases those studies are based on the same paradigm: one measures emotional responses while the subject is presented to a static sound sample with specific acoustic characteristics or an excerpt of music representative of a certain type of emotions.

In this paper, we take a synthetic and dynamic approach to the exploration of mappings between perceived musical tension [7, 8] and a set of musical parameters by using Reinforcement Learning (RL) [9].

Reinforcement learning (as well as agent-based technology) has already been used in various musical systems and most notably for improving real time automatic improvisation [10, 11, 12, 13]. Musical systems that have used reinforcement learning can roughly be divided into three main categories based on the choice of the reward characterizing the quality of musical actions. In one scenario the reward is defined to match internal goals (a set of rules for instance), in another scenario it can be given by the audience (a like/dislike criterion), or else it is based on some notion of musical style imitation [13]. Unlike most previous examples where the reward relates to some predefined musical rules or quality of improvisation, we are interested in the emotional feedback from the listener in terms of perceived musical tension (Figure 1).

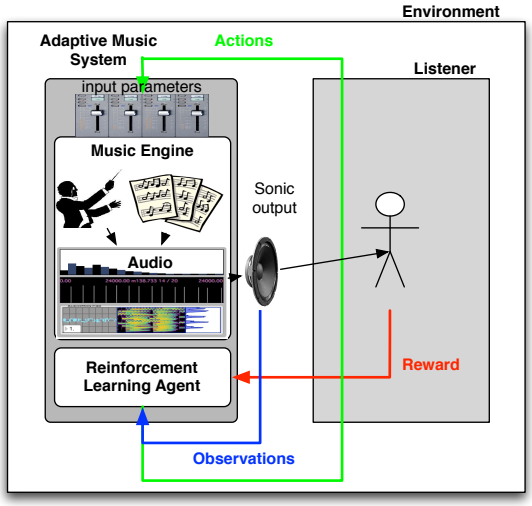

Figure 1. The system is composed of three main components: the music engine (SiMS), the reinforcement learning agent and the listener who provides the reward signal

Reinforcement learning is a biologically plausible machine learning technique particularly suited for an explorative and adaptive approach to emotional mapping as it tries to find a sequence of parameter change that optimizes a reward function (in our case musical tension). This approach contrasts with expert systems such as the KTH rule system [14, 15] that can modulate the expressivity of music by applying a set of predefined rules inferred from previous extensive music and performance analysis. Here, we propose a paradigm where the system learns to autonomously tune its own parameters in function of the desired reward function (musical tension) without using any a-priori musical rule.

Interestingly enough, the biological validity of RL is supported by numerous studies in psychology and neuroscience that found various examples of reinforcement learning in animal behavior (e.g. foraging behavior of bees [16], the dopamine system in primate brains [17], ...).

Copyright: ©2010 Sylvain Le Groux et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License 3.0 Unported, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

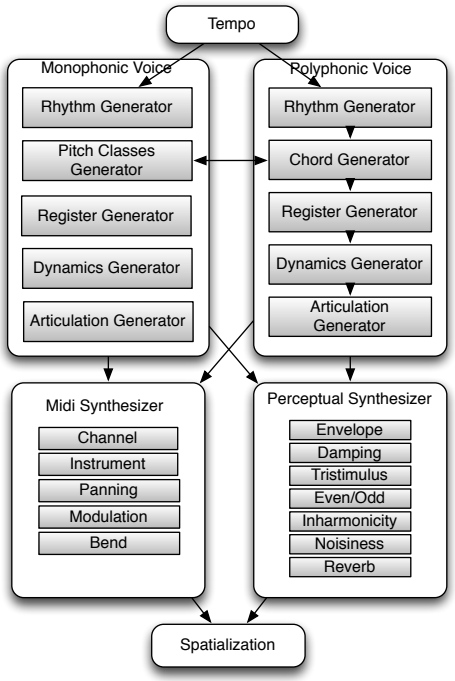

Figure 2. SiMS is a situated music generation framework based on a hierarchy of musical agents communicating via the OSC protocol.

2. A HIERARCHY OF MUSICAL AGENTS FOR MUSIC GENERATION

We generate the music with SiMS/iMuSe, a Situated Intelligent Interactive Music Server programmed in Max/MSP [18] and C++. SiMS's affective music engine is composed of a hierarchy of perceptually meaningful musical agents (Figure 2) interacting and communicating via the OSC protocol [19]. SiMS is entirely based on a networked architecture. It implements various algorithmic composition tools (e.g: generation of tonal, Brownian and serial series of pitches and rhythms) and a set of synthesis techniques validated by psychoacoustical tests [20, 3]. Inspired by previous works on musical performance modeling [14], iMuSe allows to modulate the expressiveness of music generation by varying parameters such as phrasing, articulation and performance noise.

Our interactive music system follows a biomimetic architecture that is multi-level and loosly distinguishes sensing (the reward function) from processing (adaptive mappings by the RL algorithm) and actions (changes of musical parameters). It has to be emphasized though that we do not believe that these stages are discrete modules. Rather, they will share bi-directional interactions both internal to the architecture as through the environment itself [21]. In this respect it is a further advance from the traditional separation of sensing, processing and response paradigm[22] which was at the core of traditional AI models.

In this project, we study the modulation of music by three parameters contributing to the perception of musical tension, namely articulation, tempo and dynamics.

While conceptually fairly simple, the music material generator has been designed to keep the balance between predictability and surprise. The real-time algorithmic composition process is inspired by works from minimalist composers such as Terry Riley *(In C, 1964)* where a set of basic precomposed musical cells are chosen and modulated at the time of performance creating an ever-changing piece.

The choice of base musical material relies on the extended serialism paradigm. We a priori defined sets for every parameter (rhythm, pitch, register, dynamics, articulation). The generation of music from these sets is then using non-deterministic selection principles, as proposed by Gottfried Michael Koenig [23]. (The sequencer modules in SiMS can, for instance, choose a random element from a set, or choose all the elements in order successively, choose all the elements in reverse order, or play all the elements once without repetition, etc.)

For this project we used a simple modal pitch serie [0, 3, 5, 7, 10] shared by three different voices (2 monophonic and 1 polyphonic). The first monophonic voice is the lead, the second is the bass line, and the third polyphonic voice is the chord accompaniment. The rhythmic values are coded as 16n for a sixteenth note, 8n for a eighth note, etc. The dynamic values are coded as midi velocity from 0 to 127. The other parameters correspond to standard pitch class set and register notation. The pitch content for all the voices is based on the same mode.

- Voice1:
  - Rhythm: [16n 16n 16n 16n 8n 8n 4n 4n]
  - Pitch: [0, 3, 5, 7, 10]
  - Register: [5 5 5 6 6 6 7 7 7]
  - Dynamics: [90 90 120 50 80]
- Voice2:
  - Rhythm:[4n 4n 4n 8n 8n]
  - Pitch: [0, 3, 5, 7, 10]
  - Register: [3 3 3 3 4 4 4 4]
  - Dynamics: [90 90 120 50 80]
- Polyphonic Voice:
  - Rhythm: [2n 4n 2n 4n]
  - Pitch: [0 3 5 7 10]
  - Register: [5]
  - Dynamics: [60 80 90 30]
  - with chord variations on the degrees [1 4 5]

The selection principle was set to "series" for all the parameters so the piece would not repeat in an obvious way [1]. This composition paradigm allows the generation of constantly varying, yet coherent, musical sequences. Properties of the music generation such as articulation, dynamics modulation and tempo are then modulated by the RL algorithm in function of the reward defined as the musical tension perceived by the listener.

[1] Samples: http://www.dtic.upf.edu/~slegroux/confs/SMC10

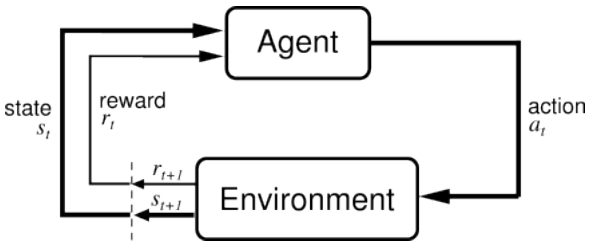

Figure 3. The agent-environment interaction (from [9])

3. MUSICAL PARAMETER MODULATION BY REINFORCEMENT LEARNING

3.1 Introduction

Our goal is to teach our musical agent to choose a sequence of musical gestures (choice of musical parameters) that will increase the musical tension perceived by the listener. This can be modeled as an active reinforcement learning (RL) problem where the learning agent must decide what musical action to take depending on the emotional feedback (musical tension) given by the listener in real-time (Figure 1). The agent is implemented as a Max/MSP external in C++, based on RLKit and the Flext framework [2].

The interaction between the agent and its environment can be formalized as a Markov Decision Process (MDP) where [9]:

- at each discrete time t, the agent observes the environment's state $s_t \in S$, where S is the set of possible states (in our case the musical parameters driving the generation of music).
- it selects an action $a_t \in A(s_t)$, where $A(s_t)$ is the set of actions available in state s_t (here, the actions correspond to an increase or decrease of the musical parameter value)
- the action is performed and a time step later the agent receives a reward $r_{t+1} \in R$ and reaches a new state s_{t+1} (the reward is given by the listener's perception of musical tension)
- at time t the policy is a mapping $\pi_t(s,a)$ defined as the probability that $a_t = a$ if $s_t = s$ and the agent updates its policy as a result of experience

3.2 Returns

The agent acts upon the environment following some policy π. The change in the environment introduced by the agent's actions is communicated via the reinforcement signal r. The goal of the agent is to maximize the reward it receives in the long run. The discounted return R_t is defined as:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} \quad (1)$$

[2] http://puredata.info/Members/thomas/flext/

where $0 \leq \gamma \leq 1$ is the discount rate that determines the present value of future rewards. If $\gamma = 0$, the agent only maximizes immediate rewards. In other words, γ defines the importance of future rewards for an action (increasing or decreasing a specific musical parameter).

3.3 Value functions

Value functions of states or state-action pairs are functions that estimate how good (in terms of future rewards) it is for an agent to be in a given state (or to perform a given action in a given state).

$V^\pi(s)$ is the state-value function for policy π. It gives the value of a state s under a policy π, or the expected return when starting in s and following π. For MDPs we have:

$$\begin{aligned} V^\pi(s) &= E_\pi\{R_t | s_t = s\} \\ &= E_\pi\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\} \end{aligned}$$

$Q^\pi(s,a)$, or action-value function for policy π, gives the value of taking action a in a state s under a policy π.

$$\begin{aligned} Q^\pi(s,a) &= E_\pi\{R_t | s_t = s, a_t = a\} \\ &= E_\pi\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s, a_t = a\} \end{aligned}$$

We define as optimal policies the ones that give higher expected return than all the others. Thus,$V^*(s) = max_\pi V^\pi(s)$, and $Q^*(s,a) = max_\pi Q^\pi(s,a)$ which gives $Q^*(s,a) = E\{r_{t+1} + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\}$

3.4 Value function estimation

3.4.1 Temporal Difference (TD) prediction

Several methods can be used to evaluate the value functions. We chose TD learning methods over Monte Carlo methods as they allow for online incremental learning. With Monte Carlo methods, one must wait until the end of an episode whereas with TD, one need to wait only one time step. The TD learning update rule for V^* the estimate of $V$is given by:

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

where α is the step-size parameter or learning rate. It controls how fast the algorithm will adapt.

3.4.2 Sarsa TD control

For the transitions from state-action pairs we use a method similar to TD learning called sarsa on-policy control. On-policy methods try to improve the policy that is used to make decision. The update rule is given by:

$$\begin{aligned} Q(s_t,a_t) &\leftarrow Q(s_t,a_t) + \alpha[r_{t+1} + \\ &... \quad \gamma Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)] \end{aligned}$$

3.4.3 Memory: Eligibility traces (Sarsa(λ))

An eligibility trace is a temporary memory of the occurrence of an event.

We define $e_t(s,a)$ the trace of the state-action pair s, a at time t. At each step, the traces for all states decay by $\gamma\lambda$ and the eligibility trace for the state visited is incremented. λ represent the trace decay. It acts as a memory and sets the exponential decay of a reward based on previous context.

$$e_t(s,a) = \begin{cases} \gamma\lambda e_{t-1}(s,a) + 1 & for\, s = s_t\,,\, a = a_t \\ \gamma\lambda e_{t-1}(s,a) & if\, s \neq s_t \end{cases}$$

we have the update rule

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha\delta_t e_t(s,a)$$

where

$$\delta_t = r_{t+1} + \gamma Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)$$

3.4.4 Action-value methods

For the action-value method, we chose a ϵ-greedy policy. Most of the time it chooses an action that has maximal estimated action value but with probability ϵ it instead select an action at random [9].

4. MUSICAL TENSION AS A REWARD FUNCTION

We chose to base the autonomous modulation of the musical parameters on the perception of tension. It has often been said that musical experience may be characterized by an ebb and flow of tension that gives rise to emotional responses [24, 25]. Tension is considered a global attribute of music, and there are many musical factors that can contribute to tension such as pitch range, sound level dynamics, note density, harmonic relations, implicit expectations, ...

The validity and properties of this concept in music have been investigated in various psychological studies. In particular, it has been shown that behavioral judgements of tension are intuitive and consistent across participants [7, 8]. Tension has also been found to correlate with the judgement of the amount of emotion of a musical piece and relates to changes in physiology (electrodermal activity, heart-rate, respiration) [26].

Since tension is a well-studied one-dimensional parameter representative of a higher-dimensional affective musical experience, it makes a good candidate for the one-dimensional reinforcer signal of our learning agent.

5. PILOT EXPERIMENT

As a first proof of concept, we looked at the real-time behaviour of the adaptive music system when responding to the musical tension (reward) provided by a human listener. The tension was measured by a slider GUI controlled by a standard computer mouse. The value of the slider was sampled every 100 ms. The listener was given the following instructions before performing the task: "use the slider to express the tension you experience during the musical performance. Move the slider upwards when tension increases and downward when it decreases".

The music generation is based on the base material described in section 2. The first monophonic voice controlled the right hand of a piano, the second monophonic voice an upright acoustic bass and the polyphonic voice the left hand of a piano. All the instruments were taken from the EXS 24 sampler from Logic Pro (Apple).

The modulation parameter space is of dimension 3. Dynamics modulation is obtained via a midi velocity gain factor between [0.0, 2.0]. Articulation is defined on the interval [0.0, 2.0] (where a value > 1 corresponds to a legato and < 1 a staccato). Tempo is modulated from 10 to 200 BPM. Each dimension was discretized into 8 levels, so each action of the reinforcement algorithm produces an audible difference. The reward values are discretized into three values representing musical tension levels (low=0, medium=1 and high=2).

We empirically setup the sarsa(λ) parameters, to $\varepsilon = 0.4, \lambda = 0.8, \gamma = 0.1, \alpha = 0.05$ in order to have an interesting musical balance between explorative and exploitative behaviors and some influence of memory on learning. ϵ is the probability of taking a random action. λ is the exponential decay of reward (the higher λ, the less the agent remembers). α is the learning rate (if α is high, the agent learns faster but can lead to suboptimal solutions).

5.0.5 One dimension: independant adaptive modulation of Dynamics, Articulation and Tempo

As our first test case we looked at the learning of one parameter at a time. For dynamics, we found a significant correlation ($r = 0.9, p < 0.01$): the tension increased when velocity increased (Figure 4). This result is consistent with previous psychological literature on tension and musical form [27]. Similar trends were found for articulation ($r = 0.25, p < 0.01$) (Figure 5) and tempo ($r = 0.64, p < 0.01$) (Figure 6). Whereas litterature on tempo supports this trend [28, 2], reports on articulation are more ambiguous [2].

5.0.6 Two dimensions: modulation of Tempo and Dynamics

When testing the algorithm on the 2-dimensional parameter space of Tempo and Dynamics, the convergence is slower. For our example trial, an average reward of medium tension (value of 1) is only achieved after 16 minutes of training (1000 s.) (Figure 7) compared to 3 minutes (200 s.) for dynamics only (Figure 4). We observe significant correlations between tempo ($r = 0.9, p < 0.01$), dynamics ($r = 0.9, p < 0.01$) and reward in this example, so the method remains useful for the study the relationship between parameters and musical tension. Nevertheless, in this setup, the time taken to converge towards a maximum mean reward would be too long for real-world applications such as mood induction or music therapy.

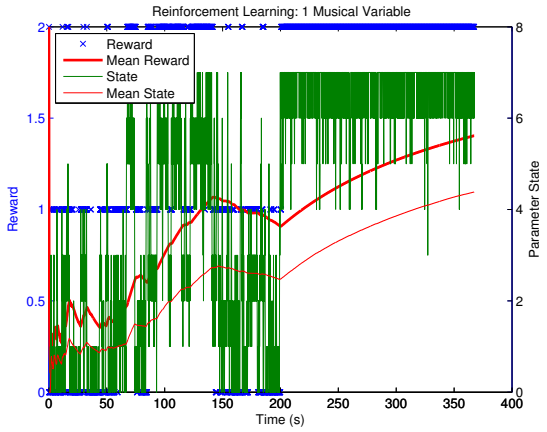

Figure 4. The RL agent automatically learns to map an increase of perceived tension, provided by the listener as a reward signal, to an increase of the dynamics gain. Dynamics gain level is in green, cumulated mean level is in red/thin, reward is in blue/crossed and cumulated mean reward is in red/thick.

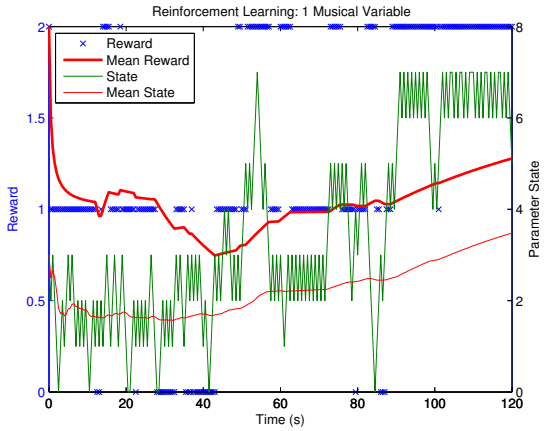

Figure 5. The RL agent learns to map an increase of perceive tension (reward) to longer articulations.

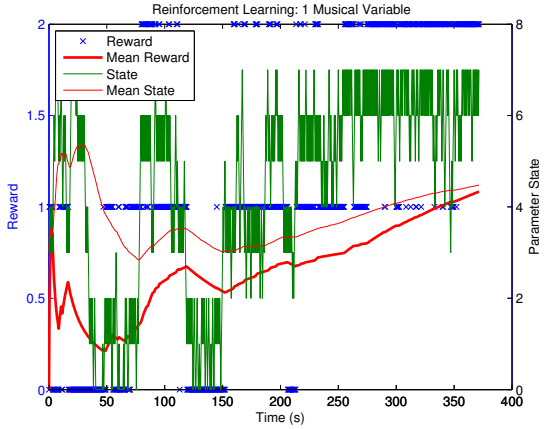

Figure 6. The RL agent learns to map an increase of musical tension (reward) to faster tempi.

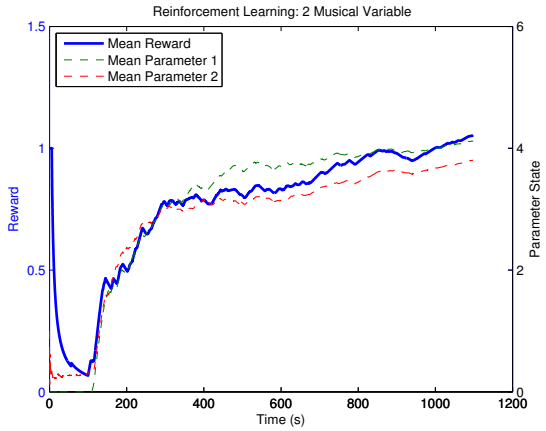

Figure 7. The RL agent learns to map an increase of musical tension (reward in blue/thick) to faster tempi (parameter 1 in green/dashed) and higher dynamics (parameter 2 in red/dashed).

5.0.7 Three dimensions: adaptive modulation of Volume, Tempo and Articulation

When generalizing to three musical parameters (three dimensional state space), the results were less obvious within a comparable interactive session time frame. After a training of 15 minutes, the different parameters values were still fluctuating, although we could extract some trends from the data. It appeared that velocity and tempo were increased for higher tension, but the influence of the articulation parameter was not always clear. In figure 8 we show some excerpt where a clear relationship between musical parameter modulation and tension could be observed. The piano roll representative of a moment where the user perceived low tension (center) exhibits sparse rhythmic density due to lower tempi , long notes (long articulation) and low velocity (high velocity is represented as red) whereas a passage where the listener perceived high tension (right) exhibits denser, sharper and louder notes. The left figure representing an early stage of the reinforcement learning (beginning of the session) does not seem to exhibit any special characteristics (we can observe both sharp and long articulation. e.g. the low voice (register C1 to C2) is not very dense compared to the other voices). From these trends, we can hypothesize that perception of low tension would relate to sparse density, long articulation and low dynamics which corresponds to both intuition and previous offline systematic studies [27].

These preliminary tests are encouraging and suggest that a reinforcement learning framework can be used to teach an interactive music system (with no prior musical mappings) how to adapt to the perception of the listener. To assess the viability of this model, we plan more extensive experiments in future studies.

6. CONCLUSION

In this paper we proposed a new synthetic framework for the investigation of the relationship between musical pa-

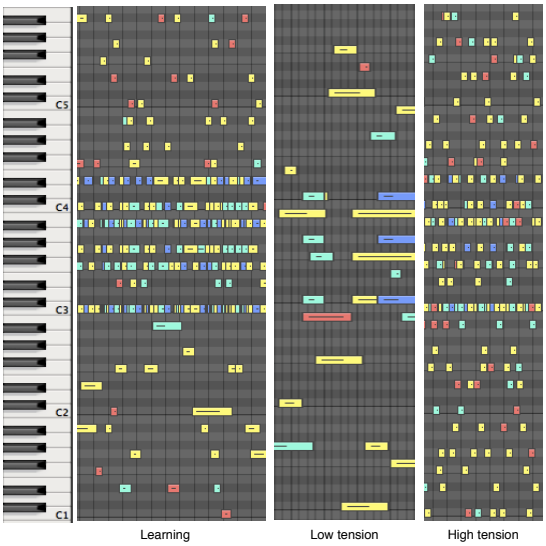

Figure 8. A piano roll representation of an interactive learning session at various stage of learning. At the beginning of the session (left), the musical output shows no specifc characteristics. After 10 min of learning, excerpts where low tension (center) and high tension reward is provided by the listener (right) exhibit different characteristics (cf text). The length of the notes correspond to articulation. Colors from blue to red correspond to low and high volume respectively.

rameters and the perception of musical tension. We created an original algorithmic music piece that can be modulated by parameters such as articulation, velocity and tempo, assumed to influence tension. The modulation of those parameters was autonomously learned in real-time by a reinforcement learning agent optimizing the reward signal based on the musical tension perceived by the listener. This real-time learning of musical parameters provides an interesting alternative to more traditional research on music and emotion. We could observe correlations between specific musical parameters and an increase of perceived musical tension. Nevertheless, one limitation of this method for real-time adaptive music is the time taken by the algorithm to converge towards a maximum average reward, especially if the parameter space is of higher dimensions. We will improve several aspects of the experiment in follow-up studies. The influence of the reinforcement learning parameters on the convergence needs to be tested in more details, and other relevant musical parameters will be taken into account. In the future we will also run experiments to assess the coherence and statistical significance of these results over a larger population.

7. REFERENCES

[1] L. B. Meyer, *Emotion and Meaning in Music*. The University of Chicago Press, 1956.

[2] A. Gabrielsson and E. Lindström, *Music and Emotion - Theory and Research*, ch. The Influence of Musical Structure on Emotional Expression. Series in Affective Science, New York: Oxford University Press, 2001.

[3] S. Le Groux, A. Valjamae, J. Manzolli, and P. F. M. J. Verschure, "Implicit physiological interaction for the generation of affective music," in *Proceedings of the International Computer Music Conference*, (Belfast, UK), Queens University Belfast, August 2008.

[4] C. Krumhansl, "An exploratory study of musical emotions and psychophysiology," *Canadian journal of experimental psychology*, vol. 51, no. 4, pp. 336–353, 1997.

[5] M. M. Bradley and P. J. Lang, "Affective reactions to acoustic stimuli.," *Psychophysiology*, vol. 37, pp. 204–215, March 2000.

[6] S. Le Groux and P. F. M. J. Verschure, "Emotional responses to the perceptual dimensions of timbre: A pilot study using physically inspired sound synthesis," in *Proceedings of the 7th International Symposium on Computer Music Modeling*, (Malaga, Spain), June 2010.

[7] W. Fredrickson, "Perception of tension in music: Musicians versus nonmusicians," *Journal of Music Therapy*, vol. 37, no. 1, pp. 40–50, 2000.

[8] C. Krumhansl, "A perceptual analysis of Mozart's Piano Sonata K. 282: Segmentation, tension, and musical ideas," *Music Perception*, vol. 13, pp. 401–432, 1996.

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, March 1998.

[10] G. Assayag, G. Bloch, M. Chemillier, A. Cont, and S. Dubnov, "OMax brothers: a dynamic yopology of agents for improvization learning," in *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, p. 132, ACM, 2006.

[11] J. Franklin and V. Manfredi, "Nonlinear credit assignment for musical sequences," in *Second international workshop on Intelligent systems design and application*, pp. 245–250, Citeseer, 2002.

[12] B. Thom, "BoB: an interactive improvisational music companion," in *Proceedings of the fourth international conference on Autonomous agents*, pp. 309–316, ACM, 2000.

[13] N. Collins, "Reinforcement learning for live musical agents," in *Proceedings of the International Computer Music Conference*, (Belfast), 2008.

[14] A. Friberg, R. Bresin, and J. Sundberg, "Overview of the kth rule system for musical performance," *Advances in Cognitive Psychology, Special Issue on Music Performance*, vol. 2, no. 2-3, pp. 145–161, 2006.

[15] A. Friberg, "pdm: An expressive sequencer with real-time control of the kth music-performance rules," *Comput. Music J.*, vol. 30, no. 1, pp. 37–48, 2006.

[16] P. Montague, P. Dayan, C. Person, and T. Sejnowski, "Bee foraging in uncertain environments using predictive hebbian learning," *Nature*, vol. 377, no. 6551, pp. 725–728, 1995.

[17] W. Schultz, P. Dayan, and P. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, p. 1593, 1997.

[18] D. Zicarelli, "How I learned to love a program that does nothing," *Computer Music Journal*, no. 26, pp. 44–51, 2002.

[19] M. Wright, "Open sound control: an enabling technology for musical networking," *Org. Sound*, vol. 10, no. 3, pp. 193–200, 2005.

[20] S. Le Groux and P. F. M. J. Verschure, "Situated interactive music system: Connecting mind and body through musical interaction," in *Proceedings of the International Computer Music Conference*, (Montreal, Canada), Mc Gill University, August 2009.

[21] P. F. M. J. Verschure, T. Voegtlin, and R. J. Douglas, "Environmentally mediated synergy between perception and behaviour in mobile robots," *Nature*, vol. 425, pp. 620–4, Oct 2003.

[22] R. Rowe, *Interactive music systems: machine listening and composing*. Cambridge, MA, USA: MIT Press, 1992.

[23] O. Laske, "Composition theory in Koenig's project one and project two," *Computer Music Journal*, pp. 54–65, 1981.

[24] B. Vines, C. Krumhansl, M. Wanderley, and D. Levitin, "Cross-modal interactions in the perception of musical performance," *Cognition*, vol. 101, no. 1, pp. 80–113, 2006.

[25] C. Chapados and D. Levitin, "Cross-modal interactions in the experience of musical performances: Physiological correlates," *Cognition*, vol. 108, no. 3, pp. 639–651, 2008.

[26] C. Krumhansl, "An exploratory study of musical emotions and psychophysiology," *Canadian Journal of Experimental Psychology*, no. 51, pp. 336–352, 1997.

[27] C. L. Krumhansl, "Music: a link between cognition and emotion," in *Current Directions in Psychological Science*, pp. 45–50, 2002.

[28] G. Husain, W. Forde Thompson, and G. Schellenberg, "Effects of musical tempo and mode on arousal, mood, and spatial abilities," *Music Perception*, vol. 20, pp. 151–171, Winter 2002.