Proceedings of the V Brazilian Conference on Neural Networks - V Congresso Brasileiro de Redes Neurais
pp. 457–462, April 2–5, 2001 - Rio de Janeiro - RJ - Brazil

CONTROLE DE SINAL FISIOLÓGICO HUMANO BASEADO EM APRENDIZADO POR REFORÇO - UMA ABORDAGEM COMPETITIVA

Iamar Zuza de Araújo[1], Weber Martins[1,2]
[1]Laboratório PIRENEUS,
Escola de Engenharia Elétrica, Universidade Federal de Goiás
[2]Mestrado em Psicologia, Universidade Católica de Goiás
Setor Universitário, Goiânia, GO, Brasil.
E-mails: iamar@cultura.com.br, weber@pireneus.ufg.br

Abstract

This paper shows a reinforcement learning model, that is mainly based on the adjustment of probabilistic transitions among states on a competitive way. The application controls the physiologic signal GSR [16,20] (Galvanic Skin Resistence) using musical estimulation by the measurement of GSC (Galvanic Skin Conductance), its inverse. The GSC signal have been studied as a measure of human physical tension. We have studied methods from reinforcement learning field and from competitive learning, on the search for system adaptation towards low galvanic conductance. We use a matrix (dice) musical structure where a transition from any cell on a column to any cell in the next column is valid, in such way that music remains continuous, that is, transitions are not sensed. The last column can be combined with the first column in such way that we have an endless source of music. The agent goal is to find musical sequences that gets lowers GSC values. Note that this is a case of non-stationary environment since the preference of musical sequences changes over time. The experiments, so far, have shown the desired effect (the decrease of GSC) by 30% when music is controlled intelligently by the Reinforcement Learning agent compared with the agent that performs music by random choices.

1. Introdução

O objetivo deste trabalho é apresentar uma nova aplicação para a proposta de Aprendizado por Reforço, enfocando o melhoramento de políticas através de abordagem competitiva, oriunda dos trabalhos da área conexionista. O uso desta abordagem visa obter uma convergência mais rápida do agente com razoável capacidade de exploração no sentido de promover equilíbrio na abordagem do dilema da exploração e aproveitamento [2, 3] para a aplicação desejada, por natureza, não estacionária.

Em várias arquiteturas de redes neurais, observamos métodos de competição que se utilizam da técnica "Winner Takes All" [12]. Tais métodos levam os sistemas a convergirem extremamente rápido. Tal efeito é atenuado pelo ajuste de vizinhanças [13, 14].

Nesta aplicação, vamos investigar o controle de respostas fisiológicas (condutância galvânica da pele [16, 20]) a estímulos arbitrários (música). Este é um passo importante na busca por respostas para situações onde observamos que o ser humano se beneficia de estímulos de natureza específica para uma melhoria, mesmo que temporária, de seu sistema físico e psicológico. Não se sabe, todavia, em que combinação, quantidade e seqüência isto acontece.

2. Aprendizado por Reforço

No Aprendizado por Reforço (AR), procura-se, no espaço de problemas, a solução sem conhecer respostas para os dados de entrada. Capta-se, por outro lado, sinais de crítica do ambiente que são correlacionados com os padrões de entrada, a fim de se criar uma representação das situações que coincidiram estatisticamente com as condições mais favoráveis. Neste método, procura-se conhecer o mapeamento da dinâmica do ambiente através da tentativa e erro. Para isso, são utilizados três fatores a serem modelados dentro de cada problema: estado, ação e recompensa, que serão detalhados a seguir.

2.1. Conceitos Básicos

O aprendizado por reforço é um método de aprendizado de máquina que se propõe a resolver problemas através da interação de um agente com o ambiente para atingir um objetivo que é determinado pelas críticas que o agente sofre do ambiente.

O AR se baseia em uma estrutura com vários componentes. Primeiramente temos o **agente** que é o aprendiz e tomador de decisões que deverá acumular experiências à medida que ele interage com o ambiente. O **ambiente** é o conjunto de todos os elementos externos ao agente com os quais ele interage. A **interação agente-ambiente** ocorre continuamente, o agente selecionando ações e o ambiente respondendo a estas ações e apresentando novas situações para o agente. O ambiente poderá dar **recompensas**, que são valores numéricos especiais. O agente deverá sempre tentar maximizar a soma total das recompensas recebidas ao longo do tempo (**retorno**). Quando definimos uma especificação completa do ambiente, podemos iniciar uma instância do aprendizado por reforço chamada de **tarefa** [1].

O agente e o ambiente interagem a cada passo em uma seqüência de passos discretos no tempo **t** = 0, 1, 2, 3, ... Em cada passo de tempo **t**, o agente recebe uma representação do estado do ambiente, $\mathbf{s_t} \in \mathbf{S}$, onde **S** é o conjunto de estados possíveis. Baseado neste estado $\mathbf{s_t}$, o agente seleciona uma ação $\mathbf{a} \in \mathbf{A(s_t)}$, onde $\mathbf{A(s_t)}$ é um conjunto de ações disponíveis no estado $\mathbf{s_t}$, como na Figura 1. Um passo de tempo mais tarde, em parte como uma conseqüência de sua ação, o agente recebe uma recompensa numérica, $\mathbf{r_{t+1}} \in \mathbf{R}$ e encontra um novo estado $\mathbf{s_{t+1}}$ [1].

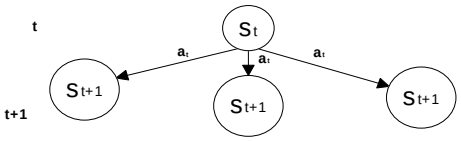

Figura 1: Três transições possíveis de um estado atual s_t para os próximos estados s_{t+1}.

A Figura 2 descreve a interação agente ambiente graficamente.

Em cada passo de tempo, o agente realiza um mapeamento de estados para probabilidades de selecionar cada ação possível. Este mapeamento é chamado de **política do agente** e é denotado por π_t, onde $\pi_t(s,a)$ é a probabilidade de $a_t = a$ se $s_t = s$. Refinando o conceito de AR do ponto de vista da política, temos que: o aprendizado por reforço especifica como o agente deve mudar sua política de acordo com os resultados de sua interação com o ambiente e com os objetivos desejados. O objetivo de um agente, grosseiramente falando, é maximizar o retorno, ou seja, o volume total de recompensa que ele recebe a longo prazo.

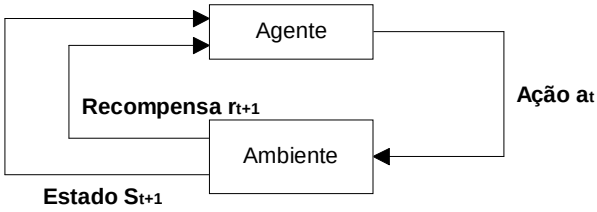

Figura 2: Interação entre o agente e o ambiente através dos sinais de estados, das ações e das recompensas.

As **ações** são meios que o agente possui de afetar o ambiente e transitar de um estado para outro. Elas podem assumir várias formas, de acordo com o nível de abstração em que o agente atua. Considere o problema de dirigir um carro, uma ação de baixo nível é aplicar energia a um motor para que o volante vire a direita, e uma ação de alto nível é a escolha de uma nova rota para se chegar ao destino.

Os **estados** são as informações a respeito do ambiente que sabemos que podem ser úteis na tomada das decisões. Parte da composição do estado poderia se basear em uma memória de sensações passadas ou mesmo estar em um nível mais alto de abstração sendo inteiramente mental ou subjetiva.

Quase todos algoritmos de aprendizado por reforço são baseados na estimativa de **funções valor** denotadas por V^{π}, funções de estado (ou de pares ação-estado), que estimam o quão bom é para um agente estar em um dado estado sob uma determinada política π.

Poderíamos dizer que um agente está adaptado a um problema quando ele conseguiu avaliar cada estado (*função valor*) e definir assim uma *política* dentro deste ambiente que lhe proporciona o melhor *retorno*, encontrando assim uma **política ótima**.

2.2. Programação Dinâmica (PD)

A programação dinâmica é uma coleção de algoritmos que podem ser usados para computar **políticas ótimas** dado um **modelo perfeito do ambiente** tal como um Processo de Decisão de Markov. Um problema pode ser representado como um Processo de Decisão de Markov [4] se os seus estados possuírem a propriedade de Markov, ou seja, o estado deverá conter em si toda a informação relevante necessária para que uma transição para o próximo estado só dependa dele mesmo de acordo com a Figura 3.

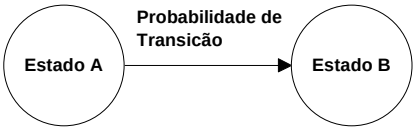

Figura 3: A transição entre estados markovianos, onde para chegar no estado B dependemos apenas das informações do estado A

Neste caso, o estado A deve considerar todos os atributos que influenciem na transição para o estado B como, por exemplo, se o estado A tivesse apenas a informação "dia chuvoso" e no estado B tivéssemos também "dia chuvoso" , seria possível deduzir que haveria alguma chance de chover amanhã só porque hoje está chovendo? Provavelmente teríamos que agregar mais informação tal como a época do ano ou direção do vento, etc. E como saber se realmente todas as variáveis estão realmente presentes?

Para computar uma política ótima, precisamos calcular a função valor ótima **V*** dos estados. Bellman [3] propôs um método para se calcular a função valor por um sistema de equações recursivas, onde o valor atual de um estado depende dos seus estados sucessores V*(s').

Captar todas as informações capazes de caracterizar os estados de um problema como estados de Markov pode ser extremamente difícil, principalmente em problemas complexos. Uma outra dificuldade é que a programação dinâmica necessita de varreduras completas no conjunto de estados para que se obtenha a política ótima. Segundo [1] para o jogo de gamão, que possui mais de 10^{20} estados, levaria-se mais de um milhão de anos para fechar uma única varredura. Apesar disso, a PD garante que a política ótima seja encontrada, propondo de maneira clara como calcular exatamente a função valor. No AR, inspira-se na teoria da programação dinâmica e procura-se estimar a função valor estado sem varrer todo o conjunto de estados.

2.3. Diferença Temporal

A Diferença Temporal (TD) utiliza-se apenas da experiência obtida pelo agente para resolver os problemas de aprendizado por reforço [4]. Ao contrário da Programação Dinâmica que assume varreduras completas em todo o conjunto de estados. A Diferença Temporal incrementa a estimativa da função valor a cada passo de tempo de acordo com Equação 2 [5].

$$V(s_t) = V(s_t) + \alpha [r_{t+1} + \gamma V(s_{t+1}) - V(s_t)] \quad (2)$$

Onde $V(s_t)$ é o valor do estado s_t, α é a taxa de aprendizado, r_{t+1} é o reforço recebido no passo t+1, γ é uma taxa de desconto entre 0 e 1 e $V(s_{t+1})$ é o valor do estado s_{t+1}.

Note que a estimativa da diferença temporal baseia-se, recursivamente em outra estimativa feita no passo de tempo anterior. Este fato é conhecido como "Bootstrap".

Dentre algumas formas mais elaboradas de implementação destes métodos, temos os algoritmos TD(λ) de Sutton [19] e Q-learning de Watkins[18], que se diferenciam, basicamente, pelo tipo de controle de política adotado.

2.4. Exploração x Aproveitamento

Quando um método se propõe a trabalhar apenas com a experiência obtida tal como no método da diferença temporal, o dilema da exploração e aproveitamento se torna bastante evidente. Isto ocorre porque torna-se muito importante definir o quanto deve-se aproveitar do conhecimento das experiências obtidas, até o momento, ou quando deve-se explorar o ambiente em busca de novas possibilidades de descobrir soluções melhores do que as já encontradas.

3. O Sistema Proposto

O sistema, proposto deverá adaptar um agente a uma relação entre duas situações complexas do mundo real. Uma delas é um sinal fisiológico captado na pele humana chamado de **GSC (Galvanic Skin Conductance)** que informa a condutância galvânica da pele. Este sinal, estudado em [16,20], está associado à tensão física do corpo humano. Quanto maior for a GSC, mais tensa, fisicamente, estará a pessoa. De outro lado, temos uma estrutura de trechos musicais que podem ser combinados com uma certa liberdade (dados musicais), dando-nos a possibilidade de "montar" uma música contínua por um tempo arbitrário. A instância utilizada desta estrutura é conhecida como **"Dados de Mozart"**[11,17].

O aprendizado por reforço nos permite trabalhar com esta proposta, indiferentemente de se conhecer quais alterações ocorrem em uma pessoa ao ouvir uma música, ou como a música poderia aumentar ou reduzir a GSC medida. No aprendizado por reforço, só precisamos definir claramente o que seria ação, estado e recompensa para este problema.

A seguir, vamos descrever a técnica de aprendizado por reforço que será utilizada e a modelagem proposta para resolver o problema dentro deste enfoque. Em seguida os resultados obtidos serão avaliados.

3.1. Método Usado

Como afirmou-se anteriormente, um agente é capaz de aprender por reforço, utilizando-se apenas três tipos de informação: estado, ação e recompensa. Alguns destes métodos ainda não possuem demonstração formal de sua convergência para a otimalidade, mas todos possuem aplicações práticas bem sucedidas como TD-Gammon [6,7], jogo de damas [8, 9] e a alocação dinâmica de canais [10].

Para muitos problemas complexos, onde não se têm o mapeamento prévio da dinâmica do ambiente, podemos reconhecer pelo menos parcialmente os estados, as ações e as recompensas envolvidas. Neste caso, o aprendizado por reforço será um grande aliado na busca por uma solução satisfatória para problemas ainda não resolvidos.

3.3. Aplicação Proposta

Representamos abaixo (ver Figura 4) a implementação proposta para os "Dados de Mozart". Cada circunferência representa um trecho de música clássica. Estas circunferências estão disposta em dezesseis colunas e onze linhas. Os segmentos que unem cada circunferência às onze circunferências do seu lado direito significam a possibilidade de se transitar deste trecho musical para qualquer um dos trechos à direita.

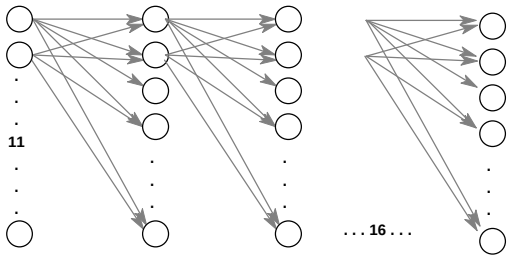

Figura 4: Estrutura dos Dados de Mozart. *Onze nodos em uma coluna se conectam totalmente a outros onze nodos da coluna seguinte, ao longo de 16 colunas.*

A idéia é privilegiar as montagens de trechos musicais que coincidiram com os valores mais baixos da GSC. Isto pode ser feito, seguindo a estrutura proposta de aprendizado por reforço, tratando este gerador de música como um agente que vai buscar uma adaptação com o ambiente, que seria o corpo humano, recebendo recompensas positivas ou negativas baseando-se no decréscimo ou crescimento da GSC respectivamente (ver Figura 5).

Dentro do paradigma de aprendizado por reforço, nossa proposta se encaixa com os seguintes elementos:

- **Estado**: é a posição atual dentro da matriz de trechos musicais.
- **Ação**: a execução de qualquer um dos trechos musicais válidos a partir de um determinado estado.
- **Recompensa**: crescimento ou decrescimento da condutância galvânica da pele. Em outras palavras, faremos a diferença entre a GSC anterior e a atual
- (Equação 3):

$$r_t = GSC_{t-1} - GSC_t \quad (3)$$

A música é gerada continuamente com 11^{16} possibilidades de combinação, em cada ciclo. Adotamos o método baseado na diferença temporal que procura otimizar a função valor estado através de medidas apenas entre a diferença de um valor de estado para outro.

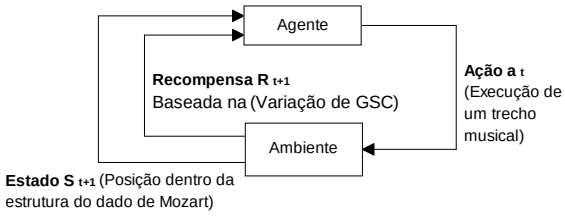

Figura 5: A arquitetura da aplicação proposta. *Encaixando a aplicação proposta na arquitetura básica do aprendizado por reforço.*

Para que uma transição ocorra será dado a cada nodo uma probabilidade de ocorrência proporcional ao seu valor de estado. Isto significa que o nó com maior valor terá mais chances de "vencer" a competição e o seu trecho musical será executado. A recompensa ou punição para o nó será distribuída com os "n" últimos vizinhos anteriormente executados na seqüência, calculado de acordo com x visto na Equação 4.

$$R_x = \gamma^{t-x} r_t \quad (4)$$

Onde x = t, t-1, t-2 ... t-n-1. O tamanho da vizinhança (o maior valor que "n" pode assumir) é um parâmetro que deve ser ajustado durante as experiências. O valor de γ deve estar entre 0 e 1 e determina a intensidade de participação da vizinhança nas recompensas ou punições. Com γ próximo a zero o sistema demostrou-se mais lento em conseguir a redução de GSC e, por outro lado, quando γ estava bem próximo de um, o sistema apresentou-se semelhante a um sistema que faz as escolhas aleatoriamente. Adotamos, então, a taxa de 0,5 para γ que nos possibilitou atingir os seguintes resultados.

3.4 Avaliação dos Resultados

O desempenho do agente foi medido através da comparação do modelo proposto com um outro modelo que age aleatoriamente, para avaliarmos a relevância estatística das reduções de sinal obtidas.

As coletas de dados foram feitas em sessões onde o indivíduo recebeu um mínimo de interferência possível como som, luz, imagem e temperatura para que a música seja seu principal alvo de concentração. Temos um grupo de cinco pessoas, chamado grupo de controle, onde cada indivíduo é submetido a duas sessões aleatórias e um grupo de nove pessoas, chamado de grupo experimental onde cada indivíduo é arbitrariamente submetido, ora a uma sessão com o agente inteligente, ora com um agente aleatório.

O primeiro grupo serve para avaliarmos as variações naturais entre as sessões aplicadas a um mesmo indivíduo. O segundo grupo permite avaliar as diferenças conseguidas pelo agente. Podemos visualizar melhor esta proposta na Figura 6.

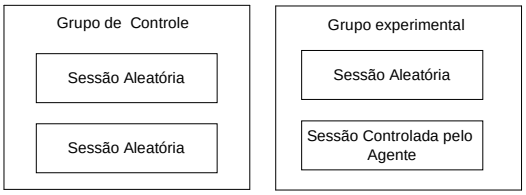

Figura 6: Projeto Experimental**.** Dois grupos para coletas de dados. O grupo de controle só recebe estímulos aleatórios e o grupo experimental é submetido a uma sessão aleatória e outra controlada pelo agente por reforço.

4 Resultados

O resultado típico obtido nas sessões experimentais apontam uma **redução média de 30% da condutância galvânica**, quando o processo é conduzido pelo agente inteligente. A Figura 7 ilustra tal fato. Entretanto no grupo de controle não observamos redução alguma de valores. Isto fortalece nossas confirmações de que o agente está promovendo a redução da GSC, através do controle dos estímulos musicais apresentados às pessoas.

Nas figuras 8 e 9, temos uma amostra dos histogramas que representam as freqüências de utilização dos trechos musicais de cada coluna dos "Dados de Mozart". Podemos observar que no processo aleatório temos uma distribuição mais homogênea da execução dos trechos musicais (Figura 8). Por outro lado o processo inteligente polariza nos trechos que conseguem maior redução de GSC (Figura 9).

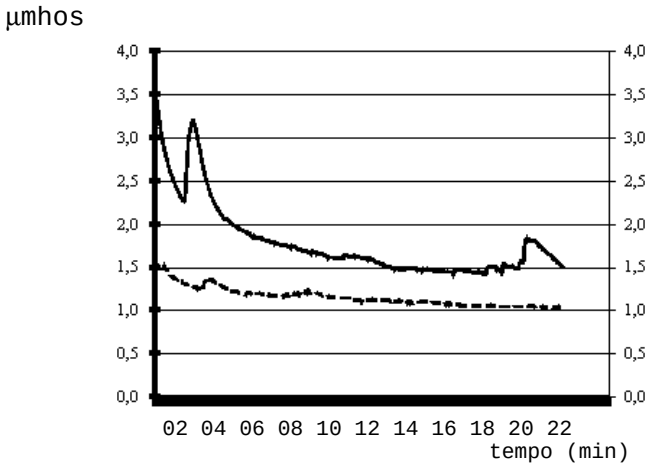

Figura 7: Medidas de GSC para um dos participantes do grupo experimental ao longo do tempo. *Em linha cheia as medidas da sessão aleatória e em linha pontilhada a sessão inteligente.*

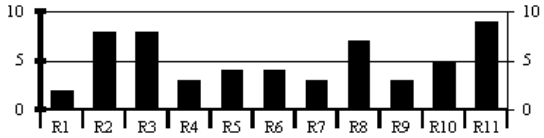

Figura 8: Histograma de uso aleatório das células do dados de Mozart. *Eixo y é a quantidade de execuções e o eixo x são as linhas do "Dado de Mozart".*

Figura 9: Histograma de uso inteligente das células do dados de Mozart. *Eixo y é a quantidade de execuções e o eixo x são as linhas do "Dado de Mozart".*

5 Conclusão

O Aprendizado por reforço nos permite trabalhar com problemas que a ciência ainda não desvendou a estrutura ou a dinâmica. O controle do sinal fisiológico GSC através da música se enquadra nesta categoria de problemas, pois não se sabe exatamente como a música pode exercer influências sobre esta medida.

Usamos o método da diferença temporal, que se mostrou bastante adequado para esta aplicação, pois o sinal de crítica estava disponível a todo instante através das medidas dos sensores.

O uso de paradigmas competitivos é importante para amenizar a predominância que um trecho musical pode assumir, o que poderia diminuir a capacidade de reação do sistema a mudanças provocadas pelas novas preferências que a pessoa pode estabelecer durante o processo.

As experiências realizadas mostraram que esta técnica tem uma capacidade promissora de reduzir a GSC, através do controle inteligente da estrutura musical "Dado de Mozart".

Outros trabalhos poderiam ajustar este modelo a uma estrutura musical mais elaborada para gerar música. Um exemplo deste tipo de estrutura pode ser visto em [15], onde temos uma decomposição de um banco de dados formado por músicas em regras gramaticais, para geração de música.

Referências

[1] Barto, A. G., Sutton, R. S, (1998), Reinforcement Learning an Introduction, The MIT press, Cambridge, Massachusetts.

[2] Kaelbling, L. P., Littman, M. L., and Moore, A.W. (1996). Reinforcement Learning: A Survey, *Journal of Artificial Intelligence Research* Volume 4, pp. 237-285.

[3] Bellman, R. (1957). *Applied Dynamic Programming*. Princeton University Press, Princeton, New Jersey.

[4] Ribeiro, C. H. C. (1999). Aprendizado por Reforço. V Escola de Redes Neurais, Conselho Nacional de Redes Neurais, ITA, São José dos Campos, SP pp. c028-c072.

[5] Tesauro, G, J. (1992) Practical issues in temporal difference learning. *Machine Learning*, 8:257-277.

[6] Tesauro, G, J. (1994) TD-Gammon a self-teaching backgammon program, achieves master-level play. *Neural Computation*. 6(2):215-219.

[7] Tesauro, G, J. (1995) Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38:58-68.

[8] Samuel A. L. (1967). Some studies in machine learning using the game of checkers. IBM Journal on Research and Development, 3:211-229. Reprinted in E. A. Feigenbaum and J. Feldman (eds.) Computers and Thought, pp.71-105. McGraw-Hill, New York, 1963.

[9] Samuel A. L. (1967). Some studies in machine learning using the game of checkers. II Recent Progress. *IBM Journal on Research and Development*, 11:601-617.

[10] Singh, S. P., and Bertsekas, D. (1997). Reinforcement learning for dynamic channel allocation in cellular telephone systems. In Advances in Neural Information Processing Systems: Proceedings of the 1996 Conference, pp. 974-980. MIT Press, Cambridge, MA.

[11] Schwanauer, Stephan M.(1993) Machine Models of Music, MIT Press, Cambridge, Massachusetts.

[12] Rumelhart, David E. Parallel distributed processing. 9. ed. University of California, San Diego

[13] Kohonen, T. (1982) "Self-organized Formation of Topologically Correct Features Maps" Biological Cybernetics, 43:59-69. Reprinted in Anderson and Rosenfeld [1988] pp. 511 – 521.

[14] Kohonen, T. (1990b). "The Self-Organizing Map" Proceedings of the IEEE, 78(9):1464.

[15] KOHONEN, Laine, Tirts, Torkkola. A Non Heuristic Automatic Composing Method, Music and Connectionism, Londres: MIT Press1991.

[16] Lykken D. T. & Vernables P. H. (1971) Direct Measurement of Skin Conductance: A Proposal for Standardization. Psychophysiology, 8, 656-672.

[17] Roads, Curtis, (1996), The Computer Music Tutorial, MIT Press, Cambridge, Massachusetts.

[18] C.J.C.H. Watkins and P. Dayan. Q-learning. Machine Learning. 8:279-292, 1992.

[19] R.S. Sutton. Implementation Details of the TD(λ) Procedure for the Case of Vector Predictions and Back-propagation. Technical Report TN87-509.1, GTE Laboratories Inc., 1989.

[20] Fredrickson, M. & Ohman, A (1979) Cardiovascular and Electrodermal Responses Conditioned to Fear-Relevant Stimuli, Psycophysiology,18,456-465.