CAPÍTULO 4

APLICAÇÕES DAS REDES NEURAIS ARTIFICIAIS À COMPOSIÇÃO MUSICAL

4. Aplicações das redes neurais artificiais à composição musical.

4.1 Apresentação.

Neste capítulo trataremos da aplicação das redes neurais artificiais à composição e a percepção musical. Em primeiro lugar, descreveremos as modelagens conexionistas da percepção auditiva, e em segunda os modelos conexionistas da composição musical. Quanto a modelagem da percepção, iniciaremos nosso caminho descritivo pelos modelos da percepção rítmica, passando à percepção de alturas, de melodias e, posteriormente, à percepção tonal. Na sequência abordaremos uma implementação desenvolvida para a composição musical. Por fim, discutiremos os limites e alcances das redes neurais artificiais aplicadas à composição musical.

4.2 Modelando a percepção e a composição musical – modelos conexionistas

Passaremos, agora, a averiguar como as Redes Neurais Artificiais são aplicadas à área musical, tanto para o estudo da percepção quanto para a geração de sistemas composicionais artificiais. Enquanto que, geralmente, nas abordagens da IA clássica encontramos um enfoque principal voltado ou para a geração automática de música ou para a análise (de caráter predominantemente sintático ou estatístico) de obras musicais, nas abordagens conexionistas teremos como ponto central o estudo da percepção auditiva. Esta diferença corrobora, em nossa opinião, os paradigmas *top-down* da IA em oposição ao *bottom-up* do conexionismo. No primeiro temos um conjunto de regras lógicas (de alto-nível) bem definidas e dadas a priore que são implementadas num algoritmo, que, numa perspectiva funcionalista (lógico-computacional) pouco tem a ver com a instanciação física que as sustentam. Enquanto que no segundo paradigma temos poucas regras de nível local que determinam o comportamento das unidades do sistema, mas não implicam diretamente em propriedades computacionais de níveis mais altos e abstratos; tais propriedades serão

observáveis apenas no conjunto de unidades básicas atuando como um único sistema. A relação que defendemos acima vem neste sentido: na IA aplicada à música temos um conjunto de regras de alguma teoria da musicologia que é implementada em algum algoritmo, podendo ou não ser validada após experimentação (tais teorias dizem respeito apenas a propriedades lógicas de alto-nível); nas RNAs aplicadas à música temos, ao contrário, um conjunto de regras que determinam o comportamento dos componentes do sistema, mas não de suas propriedades macroscópicas, como, por exemplo, a capacidades de reconhecimento de padrões ou de extrapolação de valores (tais propriedades podem explicar como o sistema perceptual funciona, com alguma plausibilidade biológica, sem determinar processos lógicos de alto-nível).

De maneira bastante semelhante à dicotomia Searleana entre IA forte e fraca, Bharucha (1991, p.84) postula a distinção entre o que chama de redes artificiais e redes antropo-símiles (*artificial networks* versus *human networks*, nas palavras do autor). O objetivo das redes artificiais vem da perspectiva do projeto de máquinas pela engenharia computacional, visando a construção de um sistema computacional que desempenhe alguma atividade, como a extração de alturas de um sinal acústico, mas sem explicar como tal atividade é realizada por seres humanos. Neste sentido, sistemas bastante diversos quanto à arquitetura computacional (e mesmo quanto ao paradigma em que se baseiam, podendo ser a IA clássica ou o conexionismo) podem executar a mesma tarefa, e a comparação entre eles será pertinente apenas em termos da eficiência na realização de tal tarefa. Por outro lado, existem as redes chamadas de antropo-símiles, que compartilham a metodologia de verificação de hipóteses das ciências experimentais. O objetivo de tais redes é responder a perguntas como: "*How (either physically or functionally) does the brain carry out this task?*" (BHARUCHA 1991, p.84). Tal objetivo leva a construção de modelos que expliquem o funcionamento de partes do cérebro envolvidas com alguma atividade específica, e se determinada localidade cerebral é conhecidamente responsável por mais de um tarefa, uma rede que a simule tem que ser capaz de realizar todas as tarefas tal e qual a localidade em questão. Isto é uma das formas de avaliação da eficiência de uma rede antropo-símile, além da comprovação empírica (com testes, em laboratório, especialmente designados) que corroborem as previsões teóricas feitas com tais redes.

Contudo, os dois tipos de RNAs estabelecidos por Bharucha (1991) ainda caem, em nossa perspectiva, no que Churchland e Sejnowsky (1992) chamam de redes simplistas ou abstratas do cérebro. Os exemplos de implementações que

discutiremos demonstrarão, acreditamos, que estão longe de uma compatibilidade com alta complexidade da fisiologia cerebral, mas isso não os impede de esclarecerem (talvez em aspectos mais funcionalistas do que fisicalistas) pontos importantes sobre a atividade perceptual e criativa, tanto em música quanto em geral.

Os modelos conexionistas, tanto artificiais quanto antropo-símile, de aspectos perceptuais são desenvolvidos para atividades específicas, como a percepção rítmica ou de tonalidade isoladamente, devido ao fato de que o conjunto de dados utilizado como *input* para rede é, normalmente, designado de maneira específica, em seu conteúdo, para cada fim. Ainda, segundo Bharucha (1991, p. 89) existe uma dissociação anatômica entre partes do cérebro responsáveis por certos processamentos envolvidos na atividade musical. Contudo, por outro lado o tipo de arquitetura utilizado para diversos fins pode ser bastante semelhante, e mesmo as funções de ativação o podem ser. Mas, de qualquer forma, a especificidade das RNAs parece poder se justificar pela objetividade no tratamento do problema em questão, visto que tendo apenas um aspecto a ser modelado e estudado pode-se verificar quais variáveis influenciam quais comportamentos da rede como um todo para aquela determinada tarefa; e, ainda, pela redução da dimensionalidade da rede, o que implica numa economia em termos computacionais e, conseqüentemente, em tempo de processamento. Vamos, pois, analisar algumas aplicações de RNAs às tarefas específicas de percepção auditiva (musical) e, posteriormente, de composição musical.

4.2.1 Modelando a percepção rítmica.

A percepção rítmica está relacionada à detecção de padrões temporais, pela determinação de periodicidades em um conjunto de eventos sonoros sobre o fluxo do tempo e sua organização. Quando nos limitamos à percepção rítmica em (alguns tipos de) música temos que evocar outras noções como metro e pulso, por exemplo. Large e Kolen (1994, p. 68), para situarem sua implementação, descrevem pulso como uma série de eventos percebidos que marcam subjetivamente unidades iguais no contínuo temporal; por metro eles entendem a medida do número de pulsos entre acentos recorrentes. Tais acentos são diferenças fenomenológicas dentro de uma seqüência de

eventos sonoros, relacionados com as manipulações de parâmetros físicos, principalmente a intensidade, mas também reforçados por outros parâmetros, como freqüência e duração. A percepção rítmica (e musical) tem na métrica um de seus mais importantes aspectos, e para percebemos uma métrica precisamos estabelecer uma pulsação sobre o conjunto de eventos sonoros. Vários autores postulam que a organização métrica existe em múltiplas escalas temporais (LERDAHL e JACKENDOFF, 1983; COOPER e MEYER, 1960). A percepção da estrutura métrica, para a teoria gerativa de LERDAHL e JACKENDOFF (1983) ocorre sobre vários níveis de pulsos dentro de uma estrutura musical, e onde temos mais pulsos sincrônicos nos vários níveis temos pontos que são percebidos como mais fortes.

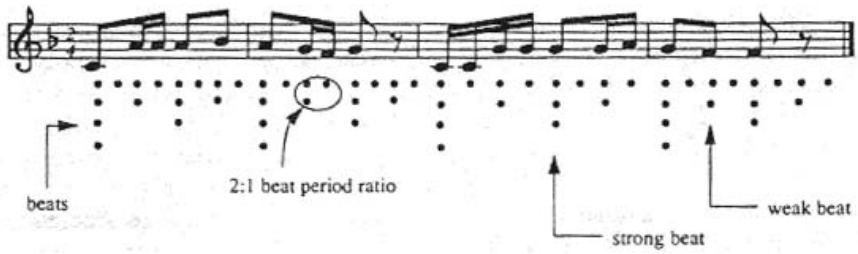

Fig. 4.1. Escalas métricas multi-temporais para a determinação do metro (In: LARGE e KOLEN, 1994, p.69).

Uma RNA que se proponha a reconhecer padrões rítmicos terá que atuar sobre estas noções (metro e pulso) para realizar sua tarefa. Existem diversas implementações concebidas com este objetivo e vamos brevemente descrever aqui algumas delas.

4.2.1.1 A quantização do tempo musical

Dentre as diversas implementações relacionadas à percepção rítmica, uma das mais discutidas na literatura é a de Desain e Honing (1991). Os autores afirmam que quantização é processo de separar "*the discrete and continuous components of musical time*" (DESAIN e HONING 1991, p.150). Os componentes contínuos do que os autores chamam de tempo musical estão relacionados ao que é comumente designado de expressividade (*expressive timing*), como *accelerando*, *rubato*,

ralentando etc. São aspectos importantes da execução musical, que enfatizam certas estruturas musicais, mas que, no entanto, normalmente apresentam pouca precisão na notação musical, além de que variações randômicas sobre o tempo, que são causadas pelos limites de precisão do sistema motor e dos processos mentais de regularidade temporal, estão associados a qualquer execução instrumental. Para que algum sistema, seja ele artificial ou biológico, reconheça um padrão rítmico, ele precisa separar adequadamente os componentes temporais discretos e contínuos presentes na execução, através do processo de quantização, e, estando tais aspectos separados, cada um deles pode servir como *input* para o outro processo. (DESAIN e HONING 1991, p.150). Desain e Honing afirmam que:

> "(…) *apart from its importance for cognitive modeling, a good theory of quantization has technical applications. It is one of the bottlenecks in the automatic transcriptions of performed music, and is also important for compositions with a real-time, interactive components where the computer improvises or interacts with a live performer. Last, but not least, a quantization tool would make it possible to study the expressive timing of music for which no score exists, as in improvised music*." (DESAIN e HONING 1991, pp.150-151)

O modelo de quantização de Desain é "*a connectionist network designed to converge from nonmetrical performance data to a metrical equilibrium state. This convergence is hard-wired into the system, and no learning takes place*" (DESAIN e HONING 1991, p.151). A arquitetura proposta é composta de células básicas e de células interativas, intercaladas entre duas células básicas. Uma célula interativa é conectada bidirecionalmente a duas células básicas, alterando os estados internos de uma em direção a um múltiplo inteiro do estado interno da outra. Os estados internos das células básicas correspondem aos intervalos de *inter-onset* entre duas notas (valores de duração). Células de soma são adicionadas à arquitetura para a representação de intervalos de tempo maiores, somando os valores de duas células básicas. As células de soma também são bidirecionalmente conectadas às células básicas, de forma que uma alteração de valor pode caminhar em qualquer uma das duas direções.

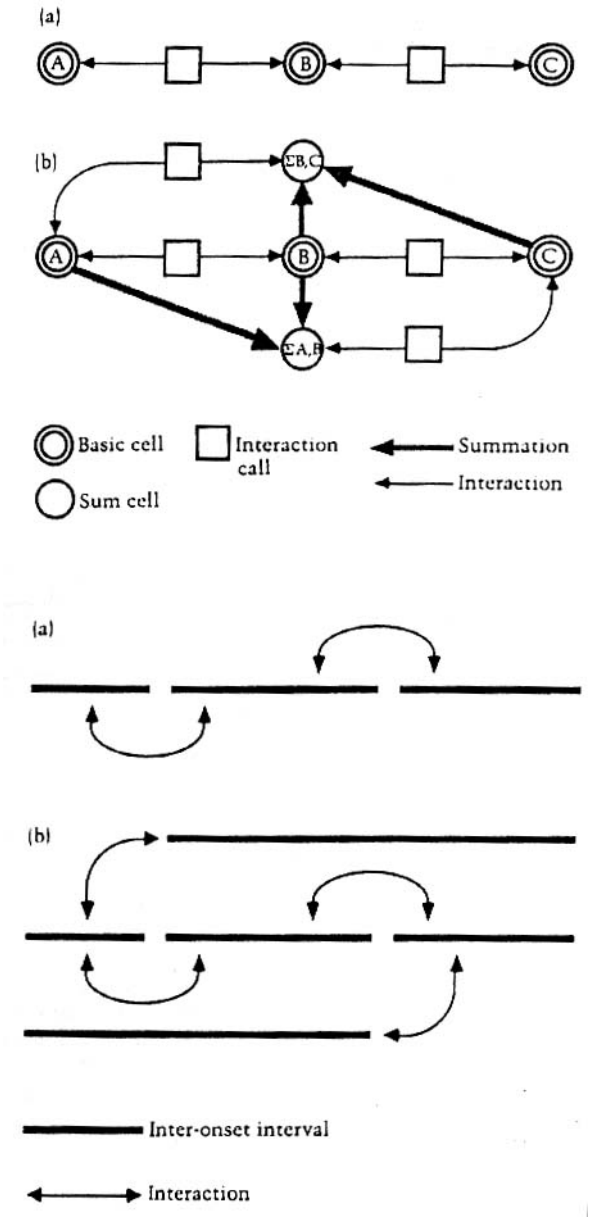

Fig. 4.2. Topologia de uma rede composta e intervalos de tempo interativos em uma rede composta (In: DESAIN e HONING, 1991, p.152).

Sem as células de soma, uma relação de valores tais como 1.1, 2.0 e 2.9 irá convergir, após *n* iterações, para 1.2, 2.4 e 2.4, longe do valor esperado de 1, 2 e 3. Com a adição das células de soma à arquitetura, a rede converge corretamente para os valores desejados (1, 2 e 3), porque o valor da última célula será determinado pela interação entre 3.1 (soma dos dois primeiros intervalos) e 2.9, da mesma forma que o primeiro valor será determinado pela interação entre os valores de 1.1 e 4.9. Uma rede composta para reconhecer uma seqüência de *n* intervalos consistirá de *n* células

básicas, $[(n + 1)\cdot(n - 2) / 2]$ células de soma, e $[n(n^2 - 1) / 6]$ células de interação (DESAIN e HONING, 1991, p.154).

Uma rede composta, com 14 células básicas, 90 células células de soma e 455 células de interação é testada pelos autores para o seguinte padrão rítmico:

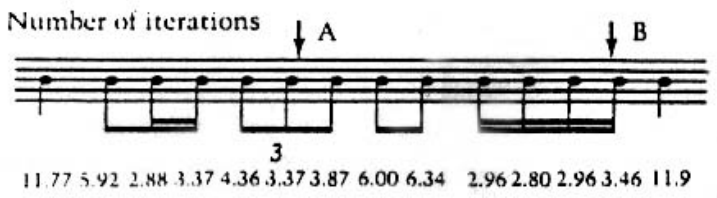

Fig. 4.3. Padrão rítmico, com os respectivos intervalos *inter-onset* de duração, para teste da rede composta. (In: DESAIN e HONING, 1991, p.152).

A classificação de padrões pela rede após 30 iterações é mostrada no gráfico abaixo:

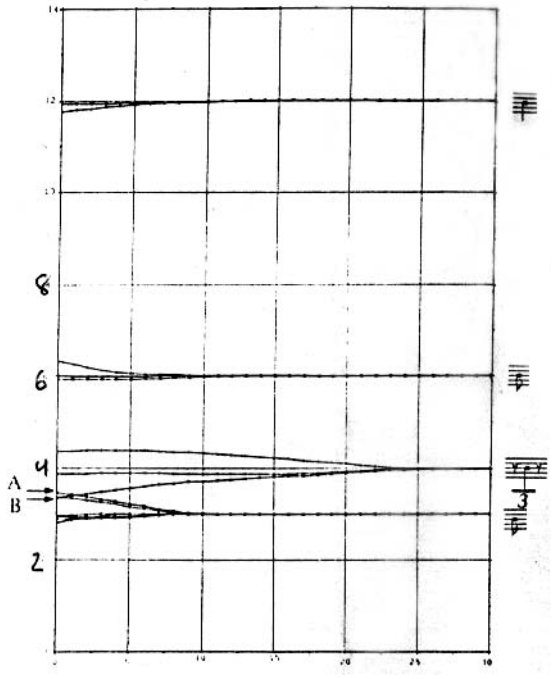

Fig. 4.4. Estados da rede em 30 iterações (In: DESAIN e HONING, 1991, p.154).

Cabe-se aqui observar que devido à interação local entre células vizinhas, valores como os das notas chamadas de A e B são classificados corretamente, agrupados dentro de um padrão musicalmente coerente. Podemos entender o comportamento de uma rede como a exposta por Desain e Honing (1991) através de

um gráfico das trajetórias dos vetores num espaço de estados em *n* iterações. Tal gráfico exibe o comportamento da rede como um todo e os pontos estáveis de atração.

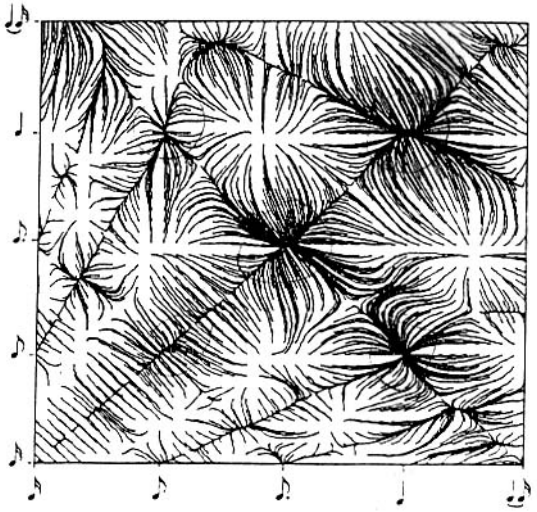

Fig. 4.5. Trajetórias num espaço de estados de um rítmo de 3 notas dentro de um compasso ternário (In: DESAIN et al., 1991, p.161).

Desain et al (1991, p. 161) afirmam ainda que

> "*Diagrams as such (...) can form the basis for experiments to test the validity of the connectionist quantizing methods as a cognitive model for rhythm perception. For example, we can plot the analogous diagram for human listeners performing a categorical perception experiment on part of the rhythm space and compare it with the output of the quantizer method*."

4.2.1.2 Métrica e ressonância.

Large e Kolen (1994) apresentam um modelo alternativo para a percepção rítmica, que em vez de estar baseado no processo de quantização, baseia-se no processo sincronização (*entrainment*) entre o sistema e o estímulo. Esta modelagem procura se aproximar dos processos perceptuais pela noção de expectativa, que segundo os autores é fundamental em atividades de percepção musical.

Modelos conexionistas da percepção são muitas vezes baseados em tarefas de extrapolação de seqüências, mas normalmente a preocupação central está localizada

sobre ‘o que virá?’ e não sobre ‘quando-virá?’ (LARGE e KOLEN 1994, p.66). Neste sentido, a distinção entre processamento de seqüências e processamento temporal deve ser considerada. “*[A] temporal processing system must predict when future events are likely to occur (‘When next?’) by exploiting knowledge of temporal structure*” (LARGE e KOLEN 1994, p.66). A estrutura métrica fornece uma referência ao percebedor que possibilita a construção da expectativa para eventos futuros. Porém, a estrutura métrica envolve múltiplas escalas temporais e, ainda, a organização temporal de uma obra apresenta desvios sistemáticos da regularidade temporal da estrutura métrica. Considerando estes aspectos, Large e Kolen fornecem um modelo matemático do *entrainment*, apropriado para a modelagem (conexionista) da a percepção da estrutura métrica.

No modelo matemático do *entrainment*, consideraremos dois ou mais osciladores que, devido a um acoplamento, entrarão em sincronia pela alteração recíproca de suas fases, de seus períodos intrínsecos, ou de ambos. Large e Kolen (1994, p.75) ainda ressaltam que:

> “(…) *rather than using coupled oscillations to describe a neural strategy for performing an implementational-level operation such as feature binding, we will use synchronization to describe how the brain may execute the relatively high-level cognitive function of meter perception. Consequently, the oscillatory units we propose will represent higher levels of neural abstraction than individual neurons*.”

Apesar de estarem buscando uma modelagem abstrata, funcionalista dos processos de reconhecimento de padrões, existem vários indícios fisiológicos sobre oscilação e sincronia desde um nível local de um único neurônio até uma larga população neuronal. Porém, a rede de osciladores proposta, como visto na citação acima, não deseja apresentar uma plausibilidade biológica de nível local, mas representar processos cognitivos de alto-nível. Cada unidade por ser considerada “*as the emergent behavior of a wide range of possible brain structures from simple neuronal substructures to large networks of oscillatory neurons*” (LARGE e KOLEN 1994, p.90).

No intuito de trabalhar sobre a percepção métrica, algumas considerações são feitas para a devida adequação. Por exemplo, em sistemas de alinhamento de fase (*phase-tracking systems*), quando o efeito do oscilador controlador é removido

mesmo que por um ciclo, o oscilador controlado volta a sua periodicidade intrínseca. Contudo, em ritmos musicais, nem sempre existem eventos em cada tempo (pulso), de forma que a pulsação, em música, não pode ser modelada apenas por sincronização por alinhamento de fase (*phase-tracking entrainment*). É necessário, então, um oscilador que responda à alinhamento de freqüência, pois em tais unidades, mesmo quando o sinal do oscilador controlador é removido, o oscilador controlado mantém a freqüência anteriormente imposta do controlador. No modelo de Large e Kolen um sinal controlador (um padrão rítmico) perturba tanto a fase quanto o período intrínseco do oscilador controlado, causando uma mudança quase permanente no comportamento da unidade. Ainda, uma unidade altera sua fase e freqüência apenas em certos pontos no padrão rítmico, isolando um elemento periódico do estímulo.

> "*With these assumptions, it will be possible to model the perception of metrical structure as a self-organizing process. What looks like a single macroscopic temporal pattern, a metrical structure, may emerge as the collective consequence of mutual entrainment among many constituent processes.*" (LARGE e KOLEN, 1994, p.78)

Cada unidade oscilatória tem um output periódico, sendo que cada pulso define o campo receptivo da unidade, onde ela pode receber influência do sinal controlador (padrão rítmico como estímulo). A função de ativação de cada unidade é dada por:

$$a(t) = \cos\frac{2\pi}{p}(t - t_o) - 1 \qquad (1)$$

onde t é o tempo, p o período de oscilação e $t\text{-}t_o \pmod p$ é a fase. O *output* é dado por:

$$o(t) = 1 + \tanh(\gamma a(t)) \qquad (2)$$

onde γ é o ganho do *output* (determina a largura do pulso).

Através do procedimento de gradiente descendente, a unidade ajusta sua freqüência e fase em relação para diminuir o erro entre seu pico de amplitude (ponto onde a unidade tem maior expectativa de receber um estímulo) e o instante em que o estímulo de fato ocorre.

A função de erro é dada por:

$$E(t) = s(t)\,(1 - o(t)) \qquad (3)$$

$E(t)$ tem um valor diferente de zero quando existe um estímulo presente ($s(t)$=1), e um valor mínimo quando a força do output é máxima ($o(t)$=1). Para o comportamento de alinhamento de fase, o erro é minimizado por gradiente descendente em t_o, gerando a seguinte regra delta:

$$\Delta t_o = -\eta_1 s(t) p \operatorname{sech}^2 \gamma a(t) \sin \frac{2\pi}{p}(t - t_o) \qquad (4)$$

Para o comportamento de alinhamento de freqüência o procedimento é semelhante, porém é mais adequado limitar o período do oscilador dentro de um alcance entre p_{min} e p_{max}, da seguinte forma:

$$p = p_{min} + 0.5(p_{max} - p_{min})(1 + \tanh \Omega) \qquad (5)$$

onde Ω é o parâmetro de controle da freqüência. Como Ω determina p, a função de erro é minimizada por gradiente descendente em Ω:

$$\Delta \Omega = -\eta_2 s(t) \operatorname{sech}^2 \gamma a(t) \sin \frac{2\pi}{p}(t - t_o) \frac{\partial p}{\partial \Omega} \qquad (6)$$

onde η_2 é a força do acoplamento para alinhamento de freqüência.

Cada uma destas unidades osciladoras sincroniza seu pulso de *output* com uma seqüência periódica de eventos discretos (*notes onsets*), ajustando a fase e freqüência de oscilação. O pulso de *output* é uma janela temporal onde a unidade pode responder a um estímulo (eventos de estímulos que ocorram for a desta janela temporal são ignorados pela unidade) (LARGE & KOLEN 1994, p.81). O gráfico abaixo mostra uma sincronia numa razão simples de 1:1 de uma destas unidades:

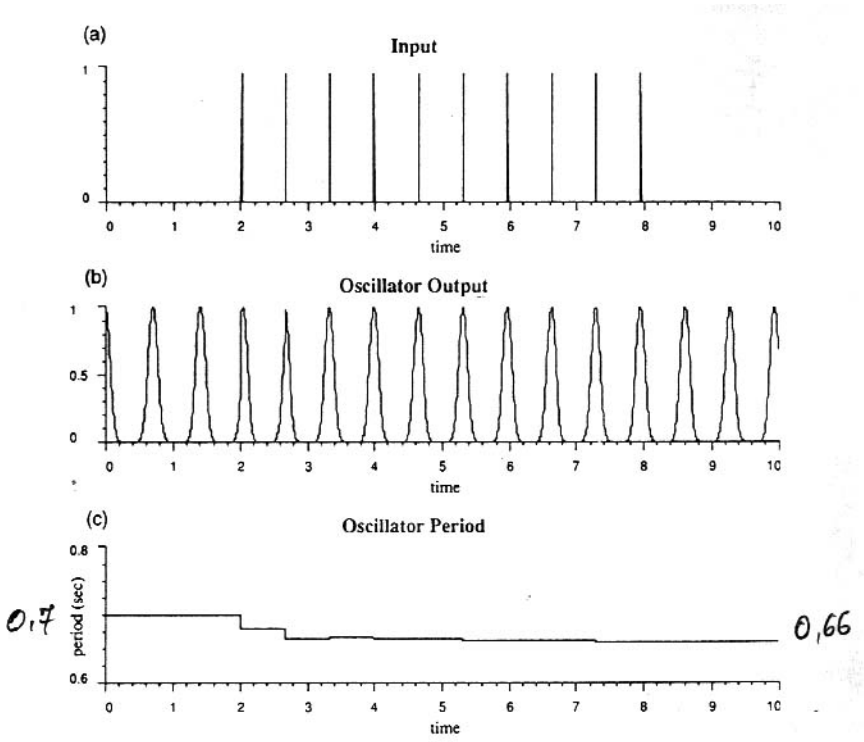

Fig. 4.6. Um oscilador respondendo a uma estimulação periódica de 660ms. Inicialmente, o período do oscilador é 700 ms. Após alguns ciclos com o estímulo presente, o oscilador ajusta seu período para 660 ms. (a) estímulo periódico; (b) resposta do oscilador; (c) período do oscilador. (In: LARGE e KOLEN, 1994, p.81)

No entanto, para a modelagem da percepção rítmica não basta que a unidade responda adequadamente a um estímulo periódico, é necessário que ela também responda adequadamente à estimulação de padrões mais complexos. E, como a unidade isola apenas um alcance de pulsações, uma pequena faixa de freqüência, é necessário uma rede delas, cada uma com um alcance de freqüência diferente, para se conseguir capturar as múltiplas escalas temporais da estrutura métrica presente em padrões rítmicos complexos.

Uma rede composta por seis osciladores foi implementada, com um alcance total indo de 600 a 2560 ms por período. Assumindo-se que o início do padrão rítmico (estímulo) restaura a fase de todos osciladores, e que todos têm o mesmo valor de força de acoplamento (η_1=0.159 e η_2=3.1416), o seguinte padrão rítmico foi apresentado à rede (apenas os valores de *note-on*, sem nenhuma informação de intensidade e altura):

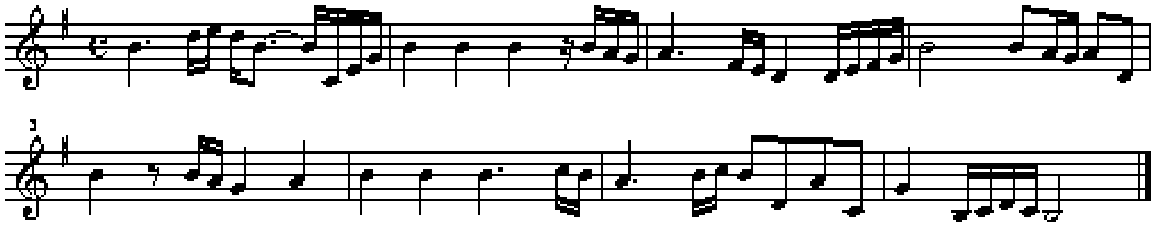

Fig. 4.7. Transcrição da melodia apresentada à rede. (In: LARGE & KOLEN 1994, p.87)

Os resultados de Large e Kolen (1994, p.87) apontam que dois dos seis osciladores, os de número 1 e 4, entraram em estabilizaram sincronicamente com o estímulo (*stable mode-locks*), enquanto que os demais nunca atingiram a estabilização.

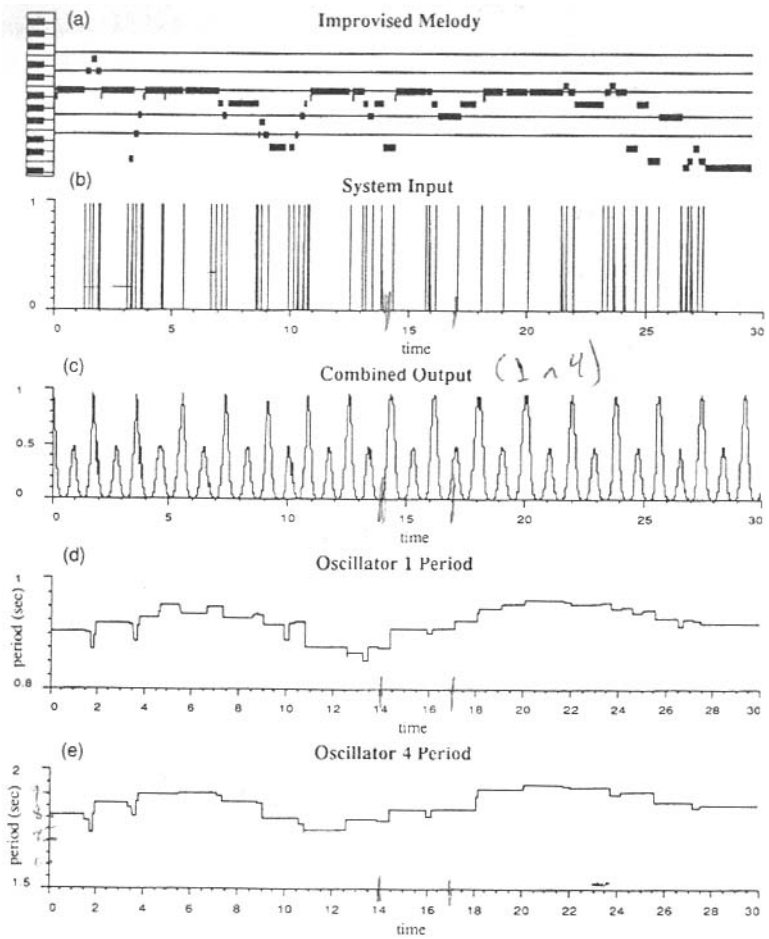

Fig. 4.8. Dois osciladores respondendo à melodia: (a) notação de *piano roll*, (b) *input* dos osciladores, (c) soma ponderada do output dos osciladores 1 e 4, (d) curva de período do oscilador 1, (e) curva de período do oscilador 4 (In: LARGE e KOLEN, 1994, p.86).

O processo de ajuste de fase e período pode ser visto com mais clareza numa no seguinte figura:

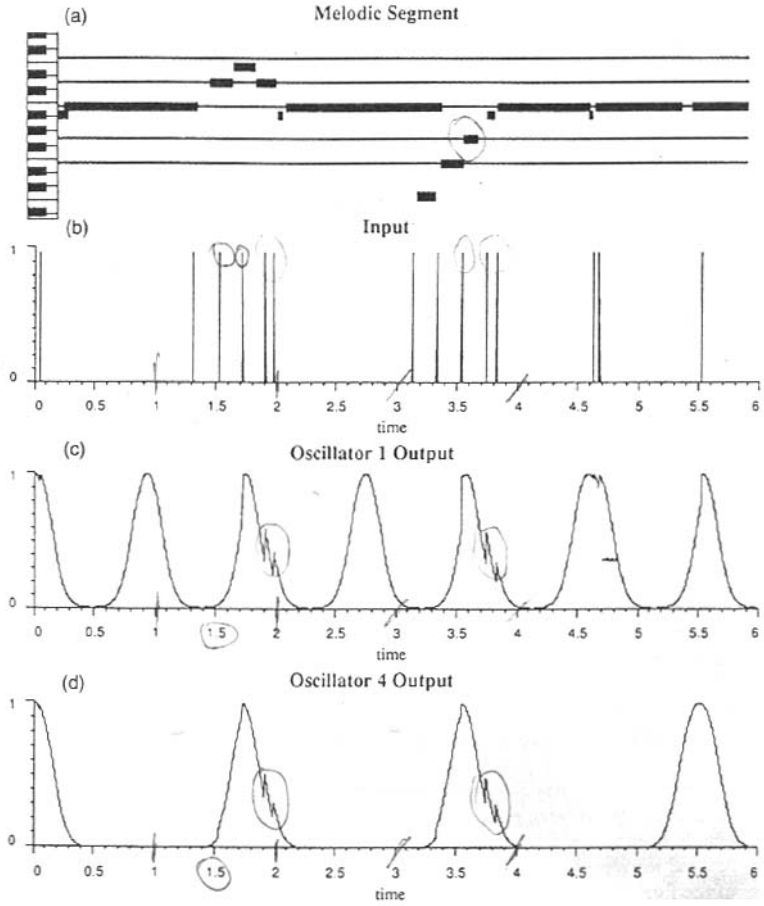

Fig. 4.9. Resposta de dois osciladores durante os primeiros segundos de uma performance improvisada: (a) notação *piano-roll* da melodia, (b) *input* dos osciladores, (c) *output* do oscilador 1, (d) *output* do oscilador 4 (In: LARGE e KOLEN, 1994, p.88).

Os autores interpretam os resultados afirmando que:

> "*First, each of these oscillators is isolating a periodic component of the complex rhythm without any phenomenical accent information. The global response, as can be seen from the combined output of the two units in figure [19c], shows that a stable metrical interpretation of the input rhythm emerges rather quickly, with strong and weak beats clearly observable. According to our metrical interpretation of this performance (see figure [18]), these two oscillators are correctly responding to the metrical structure at the quarter-note and half-note levels. Also, as figures [19d and e] show, the oscillators are tracking the performance over rather large changes in tempo*." (LARGE e KOLEN, 1994, p.87)

As novidades apresentadas pela implementação de uma rede de osciladores (não conectados uns aos outros) para a modelagem da percepção rítmica são que: a) a rede captura componentes periódicos de um padrão complexo sem utilizar acentos fenomenológicos para isto; b) unidades com um alcance diferente de freqüência capturam diferentes níveis da estrutura métrica multi-temporal; c) sem ter a rede qualquer tipo de restrição quanto a constituições adequadas (*well-formedness*), gramática e sintaticamente, de estruturas métricas, pode-se sugerir que a estrutura de padrões rítmicos, mesmo em execuções complexas, possui mais informação do que anteriormente pensado (LARGE e KOLEN, 1994, pp.89-90).

Algumas suposições são feitas sobre o comportamento de redes de osciladores interconectados (LARGE e KOLEN, 1994, pp.91-92). A influência de unidades umas sobre as outras, num processo auto-organizado, pode levar ao estabelecimento de uma pulsação predominante dentro da estrutura métrica multi-temporal. A interação entre unidades pode instanciar restrições de boa constituição métrica, de acordo com postulações teóricas de ordem sintática, onde uma taxa de pulsação é a mais adequada dentre todas as possíveis numa escala multi-temporal. Podemos ressaltar que numa abordagem *bottom-up* como esta (apesar da afirmação que esta rede é uma abstração de um processo de alto-nível) as regras sintáticas (no caso, sobre boa constituição métrica) são um reflexo da atividade das unidades. As regras não existem aprioristicamente e não explicam a percepção (como numa abordagem tipicamente *top-down*), mas o comportamento das unidades, e da rede como um todo o fazem. E, este comportamento pode, por sua vez, inclusive justificar tais regras, enquanto postulações analíticas e interpretativas da musicologia, e não da percepção. Trata-se de inverter a ordem tradicional, passando a percepção a justificar as regras musicológicas; mostrando que estas regras são um reflexo da atividade perceptual, e não que processos perceptuais operam devido a regras sintáticas pré-existentes em algum lugar de um cérebro (visto computador serial digital). Large e Kolen (1994, p.93) afirmam, neste sentido, que ressonância, é uma metáfora muito mais adequada ao estudo da percepção do que o computacionalismo clássico, principalmente depois dos trabalhos de Gibson (1966 e 1979), Treffner e Turvey (1993), Kelso e deGuzman (1988).

4.2.2 Modelando a percepção de altura e de tonalidade.

Antes de prosseguirmos a uma investigação sobre as RNAs aplicadas ao reconhecimento de altura e tonalidade vamos brevemente expor alguns pressupostos envolvidos (da psicofísica, psicologia cognitiva e neurociência) em tais modelos e, ainda, elencar alguns dos possíveis tipos de representação para altura (*pitch*).

A psicofísica é a área científica usualmente relacionada com o estudo da percepção, e Bharucha (1991, p.86) a define como "*the sutdy of how dimensions that characterize the physical world are transformed into dimensions that characterize perception*". Para conseguir relacionar o mundo físico com o perceptual, o método, experimental, utilizado por esta área envolve um sistema composto de três partes: o sistema sensorial de um indivíduo (ou animal), os estímulos físicos que são apresentados a este indivíduo, e as respostas que expressam sensações psicológicas relatadas pelo indivíduo, assumindo-se uma relação causal entre estímulo (e indivíduo) e resposta (ROEDERER, 1998, p.27). As respostas dos sujeitos submetidos a experimentos de psicofísica são, normalmente, expressas por afirmações (que supostamente devem refletir de alguma forma os critérios utilizados no processamento perceptual) como 'mais forte', 'mais brilhante', 'maior', 'menor' etc.

Altura, para a psicofísica, é a sensação (psicológica) relacionada principalmente à freqüência de vibração (parâmetro físico) de um som que estimula as células ciliadas internas de determinada região da membrana basilar. Bharucha (1991, p.86) afirma que a teoria da psicofísica atualmente mais aceita para explicar a percepção de altura (*pitch*) postula que a altura percebida é devida a um sistema de reconhecimento de padrões que executa uma comparação das freqüências componentes do espectro do estímulo com um padrão espectral anteriormente aprendido. Trata-se de um padrão de disparo das células ciliadas ao córtex auditivo, envolvendo tanto informações tonotópicas (localização na membrana basilar) quanto temporais, que é processado pelo sistema de reconhecimento de padrões (ROEDERER, 1998, p.89).

Na opinião de Bharucha (1991, p. 86-87) um dos problemas da psicofísica é que ela toma a altura como um parâmetro unidimensional, enquanto que a psicologia cognitiva a considera como um parâmetro multidimensional. Psicólogos cognitivos

referem-se a tal dimensão (da psicofísica) como *pitch height*, mas substituem-na pelo que chamam de *pitch-class*, que leva em consideração a equivalência entre oitavas. Além desta alteração, a psicologia cognitiva considera outras dimensões na percepção de altura. Uma destas outras dimensões do espaço das alturas é a relação consonância-dissonância entre sons de diferentes freqüências. Outra dimensão é o contexto em que determinado(s) som(ns) é (são) ouvido(s), por exemplo, na música tonal ocidental, alturas como as de uma tríade maior são mais proximamente relacionadas do que outras notas fora desta relação, ou notas de uma escala diatônica são mais facilmente confundidas do que notas fora desta escala (BHARUCHA, 1991, p.87). Estes fatores determinam (ou restringem) a percepção de altura e a relação perceptual (categorial) entre diversas alturas.

Outro aspecto que a psicologia cognitiva considera (e acreditamos que também a psicofísica o faz) é o processamento mental que computa as relações entre alturas. Neste sentido, as RNAs são ferramentas bastante adequadas, principalmente aquelas chamadas de redes antropo-símile. Fatores multidimensionais relacionados ao processamento mental da altura podem ser comprovados experimentalmente. Segundo Bharucha (1991, p.87), ouvintes ocidentais têm representações elaboradas, por exemplo, as relações tonais na harmonia, mesmo sem um treinamento formal (baseado em regras) em música. Isto pode ser comprovado pela velocidade e precisão com as quais um evento sonoro é processado, em função do contexto musical precedente. A velocidade e precisão de processamento variam nomologicamente com a distância, dentro do ciclo das quintas, do evento anterior. Para uma explicação melhor deste fenômeno é necessário aqui breve elucidação de aspectos de aprendizagem e inatismo envolvidos.

Para a postulação de um sistema de reconhecimento de altura, e de fenômenos perceptuais em geral, que não necessite de aprendizado é necessário a defesa de uma perspectiva inatista forte, que leva a postulação, inclusive, de categorias universais. Mas, para Bharucha (1991, p.87), "*a mechanism can be innately specified but can fail to develop if the necessary environmental conditions are not present*". Da mesma forma, o autor afirma, que algumas categorias podem ser universais simplesmente porque certas propriedades ambientais são universais, como espetros complexos (condizentes com a série harmônica). Por exemplo, a relação entre alturas distanciadas por oitavas é universalmente julgada como similar pela forte presença de relações deste tipo em sons complexos. A similaridade de oitava não pode ser,

portanto, nem fruto apenas de propriedades inatas nem apenas de aprendizagem, mas sim de um mecanismo de localização ao longo da membrana basilar que permite a detecção da relação entre os parciais da série harmônica em sons inevitavelmente presentes no ambiente, envolvendo aprendizado pela constante exposição. Dessa forma, seria pertinente afirmar que a percepção auditiva (pelo menos de altura) depende da constituição física do aparato perceptivo, da estruturação do evento sonoro e da constante exposição a este. Em acordo com isto a teoria do *virtual pitch* afirma que o espectro complexo de um acorde musical induz um certo números de alturas (*virtual pitchs*) (TERHARDT et al., 1982), que determinam a relação tonal. De forma semelhante, Parncutt (1989) defende a visão de que a relação harmônica entre acordes dentro de uma relação tonal é fruto da quantidade de notas compartilhadas por estes acordes, compartilhando a visão sobre a extração de alturas virtuais da estrutura espectral complexa, sem a necessidade de mecanismos mentais de alto-nível.

Bharucha (1991, p.88) critica as abordagens do *virtual pitch* e das notas compartilhadas com um contra-exemplo na área da harmonia. Segundo o autor, a teoria de alturas virtuais induzidas de espectros complexos e a explicação das relações harmônicas tonais por compartilhamento de notas entre acordes é infundada, visto que o encadeamento mais importante no sistema tonal é apresentado no ciclo de quintas, onde os acordes mais próximos harmonicamente não tem notas em comum. Porém, o ciclo de quintas não é apenas uma construção teorética, ele emerge das respostas de sujeitos em experimentos psicológicos (BHARUCHA, 1991, p.88), e podemos estender a afirmação de Bharucha entendendo que ele emerge devido à aprendizagem (perceptual e não sintática) pela inevitável exposição de qualquer sujeito ocidental ao tipo de relação encontrado no ciclo das quintas. De fato, se relações tonais não são fruto de propriedades inatas nem de simples propriedades acústicas, são um ótimo aspecto para uma investigação conexionista da percepção musical, que necessariamente envolve aprendizagem por exposição.

Por outro lado, os modelos conexionistas, além de estabelecerem um diálogo com áreas como psicofísica e psicologia cognitiva, relacionam-se fortemente com os postulados da neurociência, especificamente aquelas arquiteturas conexionistas que buscam implementar uma rede do tipo antropo-símile. Bharucha (1991, pp.88-89) aponta alguns pontos confluentes entre a neurociência e as RNAs, relacionados à

sintonia neuronal a certas características de estímulos sonoros e à dissociação anatômica das funções cerebrais relacionadas com a música.

Inicialmente, o autor se refere a certas evidências sobre a plasticidade de neurônios no córtex auditivo refletindo o aprendizado associativo, o que corrobora a hipótese da mudança de pesos de uma RNA como comportamento que possibilita o aprendizado. Essa mudança de pesos relaciona-se com a sintonia que uma unidade pode estabelecer com o padrão de entrada, ou características dele. Apesar de existirem algumas evidências da resposta de alguns neurônios ao contorno de pitch, pouco se sabe sobre a existência de neurônios com características mais complexas de sintonia no córtex auditivo.

> "*Neurons in the early stages of auditory processing have innately specified tuning characteristics. Because of the place coding of the cochlea, a neuron that receives excitation from the basilar membrane is tuned to a particular frequency (*its characteristic frequency*). Adjacent neurons have slightly different characteristic frequencies, with overlapping receptive fields. Collectively, these neurons constitute a tonotopic representation of the audible frequency range, scaled logarithmically (...).*
>
> *A tonotopic mapping of log frequency is also found in the auditory cortex (Lauter et al. 1985), albeit with less clarity. The presence of a tonotopic mapping throughout the auditory system and the absence of robust evidence of the existence of more abstract tuning characteristics should not preclude the postulation of abstract representational units in neural net models. Indeed, neural net models can play a valuable role in making predictions for the neuroscientific study of response selectivity, provided the models are of human and not artificial networks."* (BHARUCHA, 1991, p.89) (grifo do autor)

O autor ainda afirma que se conhece pouco sobre outros tipos de mapeamentos além do tonotópico é devido à utilização generalizada de sons senoidais nos experimentos da neurociência (como, em geral, na psicofísica e na psicologia experimental).

A segundo ponto confluente, na perspectiva de Bharucha (1991, p.89), entre a neurociência e as RNAs, a dissociação anatômica entre as diversas funções cerebrais relacionadas à musica, tem como maior consenso que os processamentos relacionados com aspectos temporais e atemporais que são resultados da atividade de partes dos hemisférios esquerdo e direito, respectivamente. Outro exemplo é oferecido pelo estudo de casos envolvendo lesões cerebrais. Pacientes que perderam todo o córtex auditivo primário (assim como boa parte de áreas não primárias) são incapazes de

detectar mudanças espectrais em eventos sonoros (acordes), mas são capazes de perceber diferenças entre os acordes, detectando relações harmônicas entre eles. Assim, se encontra evidência de que o conhecimento tácito sobre a relação entre acordes é dissociada dos mecanismos que permitem comparações entre espectros sonoros. Quando tratamos de RNAs aplicadas à modelagem da percepção auditiva, como estávamos e prosseguiremos vendo, esta confluência, de fato, se comprova pela literatura. As implementações propostas envolvem algum aspecto específico da atividade perceptual em detrimento de outros, como o estudo da percepção de padrões rítmicos normalmente ignora aspectos da percepção e de altura, ou vice versa. No entanto, antes de verificarmos algumas das implementações para o estudo da percepção de altura e de tonalidade, vamos brevemente descrever as possíveis formas de representação e organização de alturas adequadas às RNAs, a saber, representação espectral, por *pitch-height*, por intervalo, por *pitch-class* e por *pitch-class* invariante (BHARUCHA, 1991, pp.90-93).

A representação espectral é a mais próxima do sinal acústico, e a existência de representações tonotópicas no sistema auditivo, incluindo o córtex, pode ser tomada como evidência de que a percepção e a codificação de altura é acompanhada por uma representação espectral. Mas, como a representação espectral codifica um tom por seu espectro, nenhuma distinção é feita entre as muitas freqüências componentes e a altura, única, resultante, mesmo na ausência da fundamental no espectro. Nesse sentido, Bharucha (1991, p.90) afirma a necessidade de duas representações distintas, uma de espectro e outra de altura, sendo a segunda causada pela primeira. Experiências de laboratório podem fornecer evidência para esta necessidade de representações diferentes, quando testes demonstram que pode ser atribuída uma altura maior para um som com espectro mais grave que um outro som precedente. "(…) *the spectral representation alone is not representation of pitch at all but rather an elaborated representation of the signal, from which pitch is extracted*" (BHARUCHA, 1991, p.90).

Uma das representações que podem ser extraídas da espectral é a *pitch-height*, onde se estabelece uma unidade de medida para determinadas alturas. Dentro de um contínuo de alturas pode se estabelecer um número infinito de unidades, mas normalmente se estabelece um limite finito de alturas que cobrem o alcance

freqüêncial humano, que pode ser dividido por unidades que representam JNDs[1], por exemplo. Pode ser postulado, por outro lado, uma escala *pitch-height* que represente as doze categorias dentro de cada oitava (*categorial pitch-height*), equivalente às doze notas cromáticas. Parece haver, para Bharucha, pouca dúvida de que nós possuímos representações mais abstratas do que as espectrais, como representações de *pitch-height*, mas mesmo com elas ainda permancem questões. Primeiro, tal representação ignora a equivalência de oitavas (mesmo porque se a unidade de discretização não for as categorias cromáticas sem sempre a relação de oitavas irá aparecer no escalonamento). Segundo, se a única representação que altura que possuímos e de *pitch-height*, as alturas absolutas originais de uma melodia sempre seriam lembradas com precisão sempre que lembrássemos de tal melodia, como em pessoas com ouvido absoluto. Terceiro, a representação *pitch-height* ignora a invariância sob transposição, ou seja, a capacidade de reconhecer uma mesma seqüência executada em diferentes alturas. Bharucha (1991, p.91) observa que representações *pitch-height* têm pouca plausibilidade psicológica, porque não temos uma associação direta e absoluta entre as categorias previstas num escalonamento e o contínuo freqüêncial audível (tanto que é necessária uma referência física como o diapasão para uma localização precisa da nota A4); em contra partida, no caso da percepção visual, apesar das fronteiras imprecisas sobre as cores, as três categorias principais (vermelho, verde e azul) são associadas a pontos específicos sobre o contínuo de comprimento de ondas.

Outra das possíveis formas de representação para alturas é a representação de intervalos, onde uma seqüência de alturas é representada em termos dos intervalos, em semitons ou freqüências logaritmas, sucessivos entre tons subseqüentes. Bharucha (1991, pp.91-92) oferece algumas objeções quanto validade psicológica das representações intervalares, pelo menos se tomadas sozinhas. Suponha que dois eventos são apresentados com uma pequena lacuna temporal entre eles. Parece bastante pertinente afirmar que a impressão perceptual é a audição de dois eventos, e muita evidência contrária seria necessária para se sustentar a posição de que a representação (mimética) é fundamentalmente diferente da representação perceptual, de tal forma que somente a relação entre os eventos seria codificada na memória, e não os eventos propriamente ditos. Ainda, "*a [musical] piece can be recognized by its very first event if the combination of tonal, temporal, and timbral cues is sufficiently*

[1] *Just Noticiable Difference*, escala que mede as variações mínimas de freqüências percebidas.

unique" (BHARUCHA, 1991, p.91). Mas podemos destacar que para este reconhecimento ser possível temos que ter uma peça com elementos tonais, temporais e timbrísticos únicos, contudo, sabemos que na tradição tonal da música ocidental o elemento valorizado composicionalmente sempre foi a altura, e a relação intervalar entre as alturas. De forma que podemos mudar radicalmente a orquestração de uma peça musical e ouvintes provavelmente ainda a reconhecerão como a mesma peça. No entanto, sem os demais elementos este reconhecimento só será possível após alguns intervalos. Portanto, podemos concluir que o sistema tonal, de certa forma, moldou ou restringiu a atividade perceptual para a percepção de altura, e seus intervalos em detrimento de outros parâmetros sonoros que poderiam, como afirma Bharucha, estabelecer um reconhecimento sonoro mais rápido e eficiente. Da mesma forma, quando transpomos uma música para outro tom, podemos afirmar que a reconhecemos como a mesma obra devido ao fato da relação entre os intervalos, ou sua ordenação numa escala estar mantida. Contudo, outro problema que a representação por intervalo apresenta é que ela prevê quais intervalos serão empregados, mas não a ordenação entre eles. Para tanto, seria necessário que a representação incluísse os graus da escala, o que torna, assim, a representação intervalar desnecessária. Bharucha (1991, p.92) ainda afirma que "*an interval representation predicts that a single mistake will cause the key to transpose sunddenly, making recovery difficutl*". Dessa forma se um instrumentista errar uma nota de uma melodia o ouvinte perceberia dois erros, uma transposição entre a nota errada e a nota anterior e outra modulação entre a nota errada e a nota subseqüente, de forma que se o erro intervalar não for corrigido perceptualmente nos pareceria uma execução menos errônea. Nossa experiência auditiva pode facilmente nos mostrar que tal hipótese é bastante improvável.

Uma das formas mais comuns de representação de altura é a conhecida por *pitch-class*, que é a codificação de representação de *pitch-height* sobre relações de oitavas, mas preservando a altura absoluta dentro das oitavas. Segundo Bharucha (1991, p.92), o *pitch-class* resolve o problema da equivalência de oitavas, mas não resolve o problema da não existência de uma memória de longo termo para alturas absolutas (visto que se tivéssemos tal tipo de representação teríamos todos ouvidos absolutos) e nem o da invariância sob transposição. Para o último problema, o autor afirma que a relação entre tons numa modulação pode ser entendida através da distância entre tais tons dentro do ciclo das quintas, e que tal ciclo pode ser

computado em termos de *pitch-class*. Se, assim, a relação entre tons pode ser estabelecida por *pitch-class* (tal representação não é suficiente, mas necessária, afirma Bharucha), informações sobre os níveis de *pitch-class* da primeira seqüência devem estar de alguma forma disponíveis na memória quando a segunda seqüência é ouvida. Por isso, em sua RNA, Bharucha utiliza uma representação em *pitch-class* que ressoa por um curto período de tempo (*feedback* decrescente) e permite a computação de tonalidades.

Com representações do tipo *pitch-class* invariante, o problema da invariância sob transposição é eliminado. Bharucha (1991, pp.92-93) afirma que este tipo de representação é necessário para a codificação mimética de longo termo de seqüências de alturas. Todo o alcance de oitavas possibilitado por nosso sistema perceptual é reduzido a uma única oitava, com por exemplo, 12 unidades (rotuladas de 0 a 11), onde a tônica é sempre a unidade 0. Porém, este tipo de representação descarta toda e qualquer informação além do grau (tonal) de uma altura. Por exemplo, o registro em que uma nota é executada é descartado, e o tipo de memória que Bharucha se refere ao falar da representação intervalar, apresentado logo acima, que utiliza informação sobre timbre, tempo além do *pitch-height* (ou *pitch-class*), também não é possível com a representação de *pitch-class* invariante. Se o *pitch-height* não considera equivalência entre oitavas e invariância sob transposição, o *pitch-class* invariante não considera o registro.

Podemos perceber com a descrição oferecida por Bharucha (1991) que a escolha de uma determinada maneira de se representar a altura num projeto conexionista tem mais a ver com questões objetivas de implementação do que com possíveis correlatos psicológicos. Cada uma das formas descritas aqui apresenta vantagens em alguns sentido e desvantagens em outros. Como veremos na seqüência, cada arquitetura vai tratar com um tipo de representação mais adequada, na opinião de seus projetistas, à atividade de percepção de altura e tonalidade em RNAs. Inclusive, acreditamos, mesmo nas RNAs chamadas de antropo-símile, a plausibilidade psicológica, em termos do tipo de representação, é bastante superficial e tendenciosa. Mas, isto não impede que tais sistemas elucidem aspectos importantes sobre a percepção auditiva e a atividade musical, de caráter musicológico e/ou composicional, mesmo porque os pressupostos psicológicos assumidos implicitamente pela musicologia tradicional estão em grande desacordo com os estudos atuais sobre a mente humana.

4.2.2.1 Um modelo para percepção de altura.

Passamos, agora, à descrição de uma implementação conexionista para a percepção de altura desenvolvida por Sano e Jenkins (1991). A rede neural projetada pelos autores tem como objetivo o reconhecimento de alturas definidas (em *pitch-class*) a partir de estímulos formados por sons complexos. Trata-se, então, de uma rede que realiza as tarefas de, primeiro, reduzir a dimensionalidade do padrão de estímulo, e, segundo, classificar tal padrão em categorias pré-definidas. Tal arquitetura investiga também a discrepância de sensibilidade (resolução) entre as células ciliadas e as JND[2] (*just noticiable difference*), mínimas variações de freqüências percebidas.

> "*We propose a neural network model to examine the sensivity discrepancy in general and pitch perception in particular, with emphasis on the neural representation of pitch perception. The resulting model concentrates on the preprocessing of the auditory stimulus, reducing it to a simple classification problem*." (SANO e JENKINS, 1991, p.42)

O ponto de partida para esta modelagem é um estudo da fisiologia da audição, porém concentrado apenas no ouvido interno, e mais especificamente nas células ciliadas internas (órgão de Corti). Cada uma destas células se comporta como um filtro passa-banda linear de baixa-ordem, como uma largura de banda de aproximadamente 10 por cento da freqüência característica, gerando um valor[3] $Q = 10$ para freqüências acima de 500 Hz. A discrepância observada pelos autores é de que enquanto cada célula tem sensibilidade para uma largura de banda de 10% da sua freqüência característica, a JND é de 0.3% por cento para freqüências entre 500 e 2000 Hz. Isto leva à afirmação de que *"the brain obviously processes the low-grade information received form the ear to greatly increase its spectral resolution"* (SANO

[2] Barucha (1991, p.86) coloca que a utilização de algumas escalas psicofísicas, como JND ou mel, não é pertinente no estudo da percepção porque foram obtidas pela estimulação de sujeitos à sons senoidais. Tais escalas ainda falham, na sua visão, ao não levarem em conta o aspecto da similaridade entre notas separadas por oitavas.

[3] Para se obter o valor Q toma-se a freqüência característica pela largura de banda.

& JENKINS 1991, p.44). Apesar do valor JND ser de 0.3% apenas na faixa de freqüência entre 500 e 2000 Hz, para efeito de simplificação do modelo, este valor será adotado para toda a faixa das freqüências audíveis por seres humanos.

A arquitetura da rede neural de Sano e Jenkins pode ser visualizada abaixo:

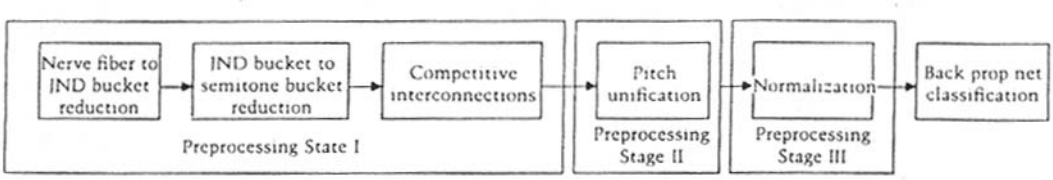

Fig. 4.10. Diagrama da rede neural para percepção de altura. (In: SANO e JENKINS, 1991, p.48)

São utilizados três estágios de pré-processamento para reduzir o nível de informação do input. O primeiro estágio tem três camadas de neurônios, sendo a primeira um modelo do *output* da cóclea e as demais modelos de processamento neuronal, apesar da plausibilidade biológica não apresentar necessariamente uma precisão próxima da realidade orgânica. Existem 28000 fibras nervosas que carregam o *output* da cóclea para o córtex, sendo que 7000 delas correspondem à faixa de freqüência 500-2000 Hz, utilizada nesta modelagem. A primeira camada do primeiro estágio executará a mesma tarefa que estas fibras nervosas, tendo neurôdos binários com resolução de freqüência de 0.2% (razão de 1.0002 entre as freqüências de dois neurôdos), de acordo com a seguinte equação:

$$\begin{aligned} 500\,\text{Hz} * x^{7000} &= 2000\,\text{Hz} \\ x &= \sqrt[7000]{4} = 1.0002 \end{aligned} \qquad (1)$$

O *output* desta camada serve de *input* para a segunda camada do primeiro estágio. A segunda camada é um modelo que possibilita uma resposta em acordo com a curva de resposta de JND, que tem sua resolução estipulada em 0.3%. Portanto, para se obter esta resolução (0.3%) teremos que obter o número *n* de neurôdos dentro de um *range* de 500-2000 Hz:

$$
\begin{aligned}
500\,\text{Hz} * (1.003)^{n} &= 2000\,\text{Hz} \\
1.003^{n} &= 4 \\
n &= \frac{\log 4}{\log 1.003} \approx 463
\end{aligned}
\qquad (2)
$$

Os 463 neurôdos JND também são binários, com um threshold que possibilita o seu disparo se apenas mais da metade de suas conexões de entrada forem excitadas. Na terceira camada do primeiro estágio existem 24 neurôdos, representando os 24 semitons das duas oitavas correspondentes ao alcance de 500-2000 Hz. O que leva a um espaçamento freqüêncial pela razão de 6% entre os semitons:

$$
\begin{aligned}
2\,\text{octaves} &= 24\ \text{semitones} \\
500\,\text{Hz} * x^{24} &= 2000\,\text{Hz} \\
x^{24} &= 4 \\
x &= \sqrt[24]{4} \approx 1.06
\end{aligned}
\qquad (3)
$$

O processamento nesta terceira camada é realizado por interconexões competitivas, onde cada unidade conta quantos *inputs* excitatórios recebeu, e por competição entre ela e suas vizinhas se determina qual é a vencedora para determinado padrão de entrada.

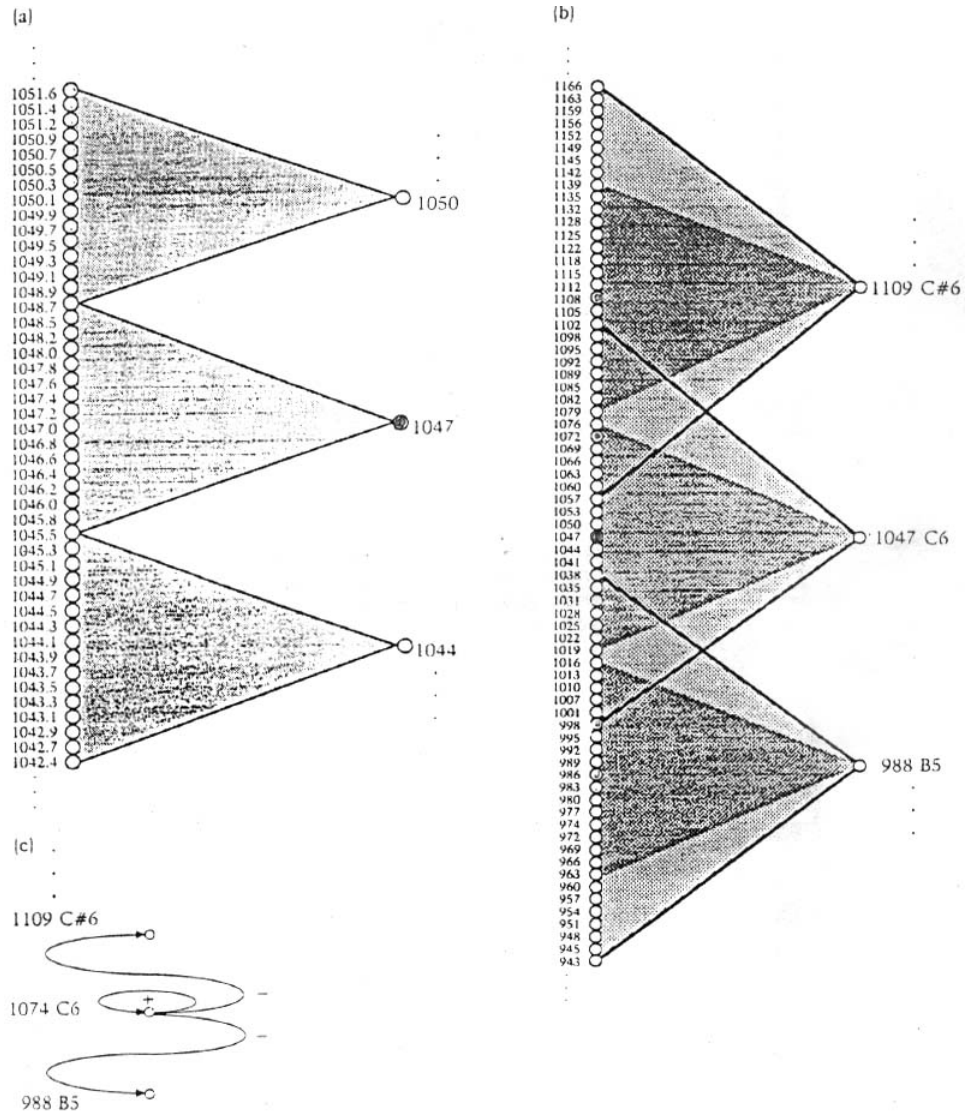

Fig. 4.11. Estrutura do estágio de pré-processamento em três camadas. (a) conexões não sobrepostas da camada "nervos cocleares" para a camada JND; (b) conexões entre a camada JND e a camada semitons, com sobreposição nas áreas claras; (c) interconexões competitivas entre neurôdos vizinhos. (In: SANO e JENKINS, 1991, p.46)

Existem as razões aproximadas de 15:1 entre a primeira e a segunda camada (nervos cocleares para JND) e de 19:1 entre as duas últimas (JND para semitons). Mas, como vimos anteriormente, cada célula ciliada pode ser entendida como um filtro linear passa-banda com Q=10, o que gera uma considerável imprecisão. Um som com uma freqüência de 1047 Hz faz as células que respondem dentro de um *range* de 998-1102 dispararem. Contudo, com uma relação de 19:1 entre as camadas segunda e terceira do primeiro estágio, apenas as unidades entre 1019-1076 Hz serão ativadas. Para a incorporação da imprecisão das células ciliadas, foi adicionando uma sobreposição (de 40%) nas conexões entre estas duas últimas camadas, para

possibilitar o alcance de 998-1102 Hz, introduzindo não-linearidade no sistema e suavizando a região de transição entre os semitons (SANO e JENKINS, 1991, p.47). Ainda, as interconexões competitivas da terceira camada possibilitam que apenas uma unidade seja a vencedora, mesmo quando os neurôdos excitados da camada anterior estejam na região de fronteira entre dois semitons.

Entre a primeira e última camada deste primeiro estágio temos uma redução aproximada na quantidade de unidades para representar um estímulo de 285:1. Esta redução parte da representação de sons complexo em décimos de Hz (com 0.02% de precisão) para um representação em *pitch-class* compreendendo duas oitavas.

Esta representação em *pitch-class* (dentro de duas oitavas) será utilizada pelas etapas dois e três da arquitetura de Sano e Jenkins. A segunda etapa executa o modo sintético de unificação de altura. Tal processo é realizado pela associação entre os parciais de um som complexo com sua fundamental, onde os parciais estão na última camada da primeira etapa e a representação das fundamentais na camada da segunda etapa. Tendo um som complexo sido representado como uma única nota, a terceira etapa normaliza as fundamentais dentro de uma classificação de altura, com 12 categorias, e outra classificação, separada, para a oitava de tal fundamental. A duas etapas são ilustradas abaixo:

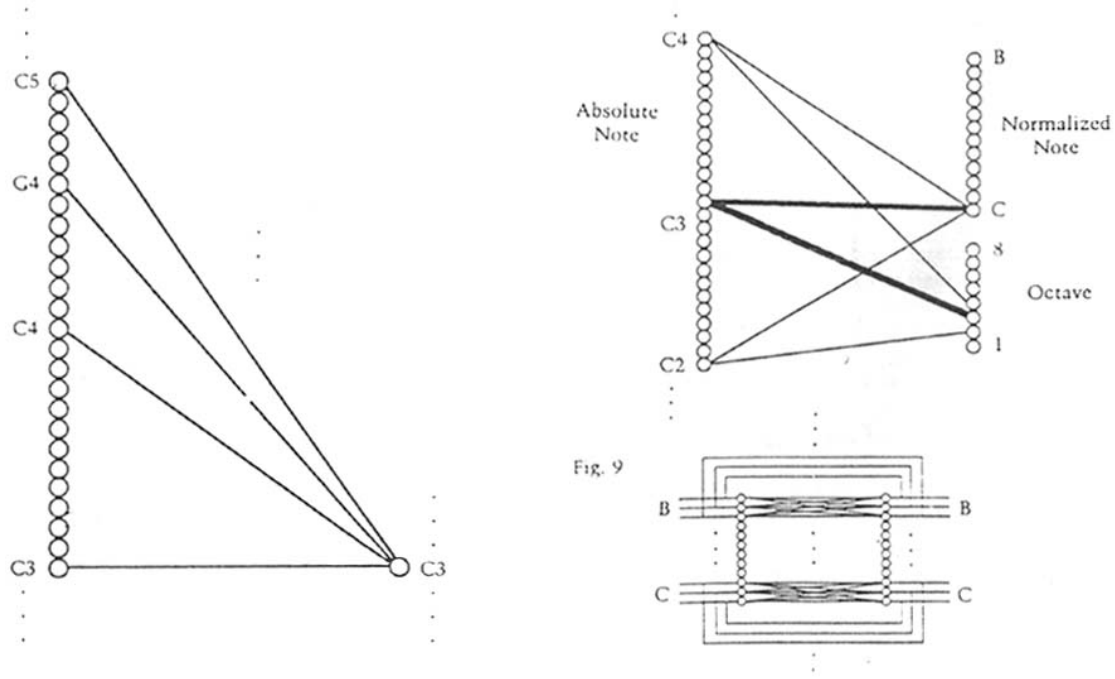

Fig. 4.12. Etapa 2 e 3 do pré-processamento da rede. (In: SANO e JENKINS, 1991, p.48)

Após todas as etapas do pré-processamento, as saídas separadas de altura e de oitava são transmitidas a duas redes neurais *back-propagation*, auto-associativas e totalmente conectadas, uma para processar notas e outra para oitavas. Após a fase de treinamento a rede pode reconhecer as 12 notas dentro de uma escala musical (SANO e JENKINS, 1991, p.47).

4.2.2.2 A percepção melódica por redes neurais artificiais – *an ear for melody*.

Katz (1994) propõe uma RNA para a operacionalização (conexionista) do ideal estético de unidade e diversidade. Existe um pressuposto de que o afeto em música deve-se às tais noções de unidade e diversidade, que são responsáveis pelo prazer contemplativo artístico. Em música estas noções relacionam-se aos conceitos de tema e variação, que são características universais sobre todas as culturas e períodos históricos (KATZ, 1994, p.200). Ou seja, existe aqui uma equivalência entre afeto em música e prazer estético. No entanto, para a aplicação de uma RNA sobre estes pressupostos, é necessária uma forma de medição do grau de unidade e/ou diversidade. Contudo, é necessária uma medição que leve em consideração a habilidade particular de um observador em unificar, por exemplo, estímulos sonoros, de forma que *"any method to measure this quantity must be embedded whithin a model of cognition"* (KATZ, 1994, p.201).

A topologia da rede de Katz consiste em várias sub-redes *feedfoward* de acordo com figura abaixo.

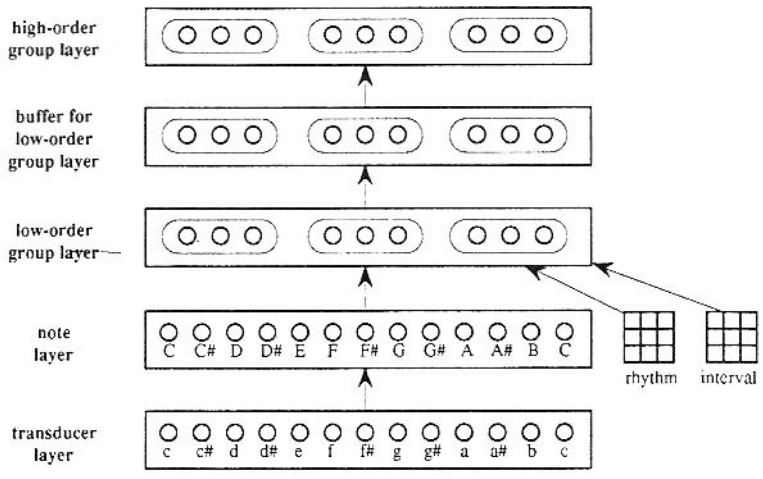

Fig. 4.13. Topologia do modelo de Katz (1994, p.204).

As duas camadas inferiores são responsáveis pelo sistema de reconhecimento de notas, onde um banco de filtros passa-banda simula o funcionamento da cóclea. Com funcionamento semelhante existem camadas que extraem informação rítmica e de intervalo. Ambas alimentam a camada intermediária (*low-order group layer*), que consiste de um grupo não-supervisionado de aglomerações competitivas de neurôdos, que devem representar grupos de eventos que pertençam a mesma categoria. A saída desta camada intermediária vai para um buffer, que alimenta a camada superior (*high-order group layer*), que executa uma classificação não-supervisionada dos grupos de baixa-ordem.

As duas camadas inferiores realizam o reconhecimento de alturas e, como veremos, assumem muitas das suposições de Sano e Jenkins (1991), como é explicitamente apontado pelo autor (1994, p.206). O intuito destas camadas é a simulação dos processos realizados pela cóclea para a elaboração de uma representação de *pitch-class* (associada a uma representação de intervalo) do sinal acústico. A camada de transdução opera como um banco de filtros passa-banda (com Q=10) ativado pelo sinal acústico. Katz (1994, p.205) aponta que o principal motivo deste sistema é a sobreposição das curvas de respostas de cada filtro, sendo a sobreposição entre filtros relativos a freqüências separadas por meio-tons de 49%, 19% quando separadas por tom inteiro, e 7% para distâncias de terça menor. Desta forma, cada unidade de *pitch-class* responde para várias unidades de transdução, todas as quais respondem, de acordo com a curva de filtragem, a uma dada freqüência. A figura abaixo ilustra a atividade destas duas camadas.

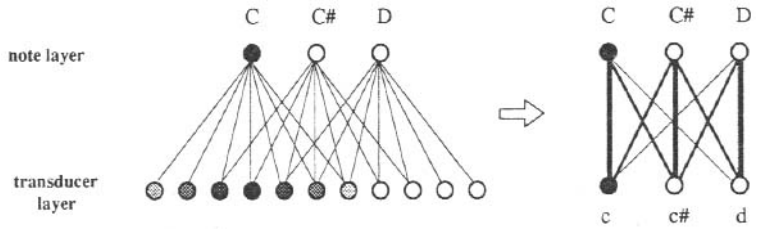

Fig. 4.14. Atividade das camadas inferiores da rede de Katz (1994, p.206).

A diferença entre o modelo de Katz (1994) e o de Sano e Jenkins (1991), é que o do primeiro autor utiliza apenas duas camadas para a tarefa de estabelecer uma representação de *pitch-class*, a de transdução e a de *pitch-class* propriamente dita.

Enquanto que o modelo de Sano e Jenkins apresenta uma cada intermediária que reflete as JND, em acordo com dados experimentais da psicofísica. Pode ser o caso de que o espaçamento dos filtros propostos por Katz já reflita o limiar de JND, apesar disto não ser explicitado pelo autor.

Katz aponta que a necessidade desta camada intermediária na RNA de Sano e Jenkins é porque eles usam filtros com janelas retangulares, enquanto que em seu "ouvido para melodia" utiliza janelas com funções exponenciais (Roex, ou *rounded –exponential functions*). As fibras nervosas dentro de um filtro com janelas retangulares disparam de maneira igual para qualquer freqüência dentro de seus limites, o que ativaria de igual maneira todas as unidades conectadas a elas, necessitando da camada de JND. Enquanto que com filtros com janelas exponenciais as fibras como um todo respondem de maneira não-linear, de forma resulta em estímulos diferentes indo para cada unidade de *pitch-class*. De qualquer forma, pelo tipo da sobreposição de células da camada de entrada da rede para as células de *pitch-class*, já seria suficiente para evitar que duas unidades *pitch-class* recebessem a mesma informação mesmo com filtros de janelas quadradas, pela diferença no número de conexões ativadas para cada uma delas; a não ser que as unidades de *pitch-class* fossem conectadas a todas as unidades inferiores ou a um número muito grande delas, mas isto seria de pouca pertinência biológica.

Contudo, Katz afirma (1994, p.206) que para efeito de simplicidade do modelo, sua rede trabalha apenas com estímulos senoidais, de forma que uma arquitetura como a citada acima seria dispensável. Neste sentido, ele elimina a camada de transdução e a substitui por uma cada já em *pitch-height*, de forma que os pesos entre as camadas de entrada e sua camada superior (*note layer*) junto à conexão de cada uma das unidades da camada inferior a três unidades superiores refletem o processo de sobreposição de filtros passa-banda, conforme ilustrado na figura 25. Existe ainda um mecanismo de fade exponencial da ativação destas unidades da camada de notas para que a camada superior, de grupos, possa considerar várias notas ao mesmo tempo para a tarefa de agrupamento melódico. Existe uma conexão excitatórias recorrente em cada unidade na camada de notas; enquanto que existem conexões inibitórias entre unidades vizinhas nesta mesma camada. "*The net effect of this arrangement is to produce an activation boost between successive notes inversely proportional to the size of the interval between the notes*" (KATZ, 1994, p.213). Existe o pressuposto que movimento melódico por grau conjunto induz um aumento

do afeto positivo, apoiado em procedimentos tradicionais de resolução de dissonância na harmonia e contraponto e na maior quantidade deste tipo de movimento melódico em músicas tradicionais (KATZ, 1994, p.206).

Então, voltando à arquitetura da rede, sobre a camada das notas vêm as camadas de grupos (*low-order* e *high-order group layer*) que devem agrupar as notas individuais em grupos, e estes grupos em grupos maiores, e possibilitando a detecção de diferenças e semelhanças entre diversos grupos. A categorização em grupos é feita por um mecanismo não-supervisionado de aprendizagem por competição. Se duas seqüências de notas, seja uma chamada de A seguida de outra B, que compartilhem características geram um ganho na ativação da camada de grupo de baixa-ordem, porque as unidades que responderam para a seqüência A continua a receber ativação durante a apresentação da seqüência B. Vamos ver em mais detalhes como esta dinâmica ocorre pela descrição de aspetos envolvidos, relacionados ao agrupamento melódico, forma de representação e categorização.

Katz (1994, p.208) entende por grupo a menor unidade, acima do nível de notas individuais, da estrutura musical. Ele apoia-se nas regras de Lerdahl e Jackendoff (1983) para divisão de melodias em grupos, que são de dois tipos:

> "(…) *those that determine whether groups are well formed, and preferences rules indicating where the likely boundaries between groups should occur. The former rules are direct to ensuring that groups have at least to members, and that group boundaries do not overlap. The latter rules, inspired by Gestalt principles, are meant to mimic human perception preferences. Thus, boundaries are formed whenever there is a relatively large gap between successive notes, whether this is a gap in pitch height, or a gap in time.*"

Tais regras podem ser implementadas num sistema conexionista. Como esta rede lida com estruturas musicas muito simples, podem ser determinadas as fronteiras entre grupos pela simples observação das frases musicais, e para o agrupamento de grupos, de ordem mais alta, da mesma forma podem ser detectados facilmente. Nos teste realizados com sua orelha para melodias, marcas tanto de baixa quanto de alta-ordem foram artificialmente estabelecidas a intervalos regulares de tempo (KATZ, 1994, p.208).

Quanto à forma de representação envolvida nesta RNA, existe um sistema que destaca simultaneamente a altura, o ritmo e o intervalo melódico, como ilustra a Figura 26. Existe um primeiro conjunto de unidades na camada de notas (*note layer*)

que codifica a altura das notas na forma de *pitch-height*. Outro conjunto de unidades codifica a informação de ritmo, associando uma duração para cada nota da melodia. Apesar da ilustração abaixo mostrar uma matriz de 4x5, existem mais unidades verticais para uma melhor resolução rítmica e mais unidades horizontais para permitir a formação grupos com mais do que quatro notas. Ainda, existem outras unidades que codificam a informação de intervalo melódico numa seqüência de notas, de maneira semelhante à maneira como a informação rítmica é representada. Existem cinco categorias para intervalos: salto ascendente; grau conjunto ascendente; mesma nota; grau conjunto descendente; e salto descendente. A representação de intervalos é necessária para permitir a distinção de dois grupos que contenham as mesmas notas e durações, mas não a mesma ordem de sucessão de notas; ou para perceber-se a semelhança quando dois grupos têm perfis melódicos similares mais não possuem exatamente as mesmas notas.

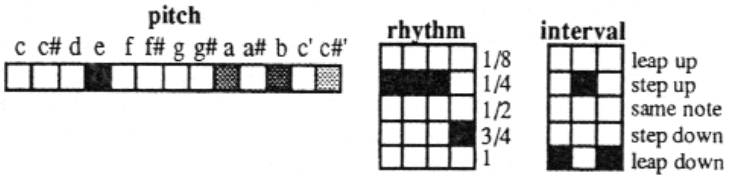

Fig. 4.15. Representação de um grupo de quatro notas na camada de notas. (In: KATZ, 1994, p.209)

A categorização é o procedimento realizado pelas camadas superiores, através da competição entre grupos de unidades em tais camadas. Katz (1994, p.210) explica que:

> "*An ART-like (Grossberg, 1980) mechanism was chosen in order to control de degree to which a group must be similar to another group if it is to be placed in the same category. Learning occurs only at the detection of a group boundary, permitting the network to acquire categories that correspond to psychological significant sets of stimuli. Learning for each group consists of choosing a unit in a competitive, winner-take-all cluster of units, and adjusting the weights to and from this unit.*"

Um cluster de unidades competitivas irá considerar dois grupos como semelhantes ou não, e coloca-los na mesma categoria, de acordo com um parâmetro de vigilância ρ. Se ρ apresenta um valor baixo, próximo a 0.1, será necessário muita

similaridade entre grupos para que uma mesma unidade responda e eles, enquanto que com um valor alto, acima de 0.8, torna-se difícil estabelecer uma única unidade vencedora se dois grupos sucessivos forem similares. Para as simulações realizadas, Katz (1994, p.211) configurou valores entre $0.5<\rho<0.8$. Além de categorizar grupos melódicos, a rede também categoriza grupos de grupos, tentando, assim, descobrir estruturas musicais de ordens mais altas. Para tanto, um *buffer*, conforme mostra a Figura 24, armazena os resultados de todas as categorizações até que uma fronteira de alta-ordem seja detectada, quando os valores deste *buffer* alimentam a camada superior da rede, que opera da mesma forma que a camada de grupos de baixa-ordem (KATZ, 1994, p.212).

Nestes últimos parágrafos estivemos descrevendo a estrutura arquitetônica da rede de Katz (1994), os tipos de representação envolvidos, e o funcionamento de cada uma das camadas da rede. Podemos passar, agora, a análise de simulações que o autor realizou para verificar o desempenho de sua rede e, ainda, verificar também se tal rede corrobora suas idéias dobre afeto positivo como uma função da similaridade entre peças musicais. Vale ressaltar que o comportamento da rede não é avaliado apenas pela camada de saída, mas por todas as camadas individualmente (da de nota até a de grupos de alta-ordem), já que cada uma delas apresenta operações sobre diferentes níveis da estrutura musical.

Os teste são quatro ao todo, sendo o primeiro relacionado ao comportamento da rede perante músicas de diferente complexidade em comparação a resultados obtidos com seres humanos; o segundo trata da relação entre familiaridade e afeto; o terceiro avalia se o modelo conexionista pode prever quando deve existir transferência de afeto positivo entre melodias similares; e o quarto teste analisa como a rede reage sobre formas degradadas de boas melodias (KATZ, 1994, p.212). Todos os parâmetros da rede foram mantidos inalterados em todos os testes.

O primeiro teste, complexidade e afetividade, visa testar o pressuposto de que afeto positivo apresenta a forma de um *U* invertido em função da complexidade musical. Complexidade aqui é entendida como o range de possibilidades sobre o qual um sistema estocástico pode escolher notas para formar uma seqüência. Neste sentido, foram criadas seqüências de 16 notas, onde cada uma foi selecionada por um processo randômico sobre uma determinada região da escala cromática. As seqüências foram artificialmente agrupadas em grupos de 4 notas cada, e estes grupos formaram grupos de alta-ordem com oito notas cada. Formou-se um conjunto de 15 melodias com

vários níveis de complexidade, e após a apresentação de todo o conjunto de treinamento ser apresentado a rede na fase de treinamento, cada melodia foi reapresentada e a ativação média (perante a apresentação da melodia inteira) de cada camada foi medida. Os resultados compõem o gráfico da figura abaixo:

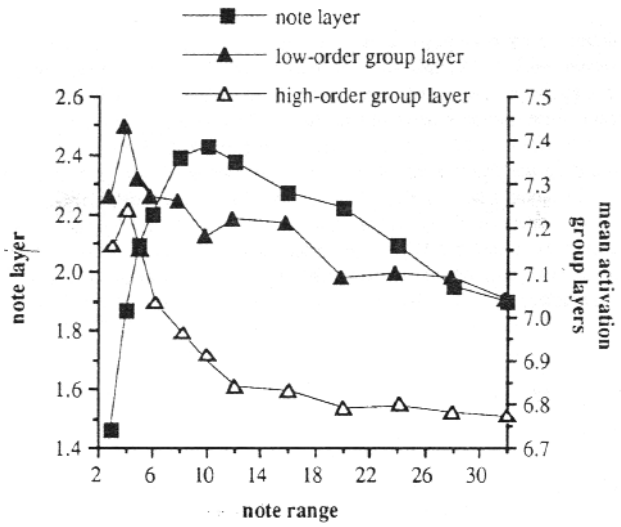

Fig. 4.16. Ativação média em função do alcance de notas em melodias geradas randomicamente. (In: KATZ, 1994, p.214)

Como podemos verificar, as suposições de Katz se confirmam, e a ativação média em todas as camadas apresenta a forma de um *U* invertido. Esta forma é decorrente da preferência por movimento de grau conjunto, pois se o alcance das possibilidades de alturas é pequeno existem mais movimentos deste tipo; por outro lado, com muitas possibilidades de alturas, a proporção de movimento por notas vizinhas é reduzida em melodias geradas randomicamente. Além de prever a existência desta forma de *U* invertido relacionada ao afeto em função da complexidade, segundo Katz (1994, p.214) outra importante avaliação deste gráfico reflete o recorrente problema da falta de estruturas de alta-ordem em música estocástica. Acompanhando a curva relativa à camada de grupo de alta-ordem, veremos que acima de 3 notas de *range* a ativação média caí significativamente, devido a falta de similaridade entre os grupos de notas; enquanto que, em geral, melodias compostas por seres humanos apresentam uma queda menos expressiva mesmo quando construídas sobre uma grande quantidade de notas diferentes.

O segundo teste, verifica se o comportamento da rede corrobora a idéia bastante intuitiva de que a familiaridade com uma peça musical aumenta o afeto, da

mesma forma que exposição excessiva a uma mesma peça acarreta numa queda deste afeto. Katz (1994, p.215), baseado em estudos da psicofísica e psicologia, afirma que a fase de aumento no afeto se deve a descoberta de propriedades e características de uma peça nas primeiras audições; enquanto que a repetição excessiva causa a desatenção pela ausência de novidade. Parece plausível que novamente uma curva em forma de *U* invertido será observada na ativação média da rede em função da quantidade de repetição de uma mesma peça. Mas algumas mudanças na rede são incorporadas para simular estes fatos. Katz (1994, pp.215-216) explica suas alterações:

> "*The rise in positive affect over a small number of repetitions, and the corresponding increase in sensivity to the thematic aspects of the music, follows directly from the adaptative characteristics of the current model.* (…) *To explain decrease affective response to over-exposure requires an explicit habituation mechanism. It is also necessary that this mechanism involve long-term effects* (…)*. The simplest possible habituation mechanism which reduces response to non-novel stimuli is Kohonen's (1984) anti-Hebbian rule* (…) *which forms an inhibitory connection between units that tend to be simultaneously active. This rule is applied in the model to inter-cluster units in the group layer.* (…) *These connections when sufficiently large, will suppress the activity in this layer when the same group of notes is presented to the network*."

Para o teste foram utilizadas três melodias de 16 notes com diferentes níveis de complexidade (*range* de 4, 8 e 16 notas), de acordo com o procedimento da simulação anterior. Cada curva representa o valor médio sobre 25 épocas (cada exposição do conjunto de treinamento). A figura abaixo apresenta o resultado para a camada *low-order group layer*, a camada de alta-ordem apresentou resultados muito próximos, e a camada de notas não foi considerada neste estudo porque as alterações só ocorreram sobre as camadas com aprendizagem por competição.

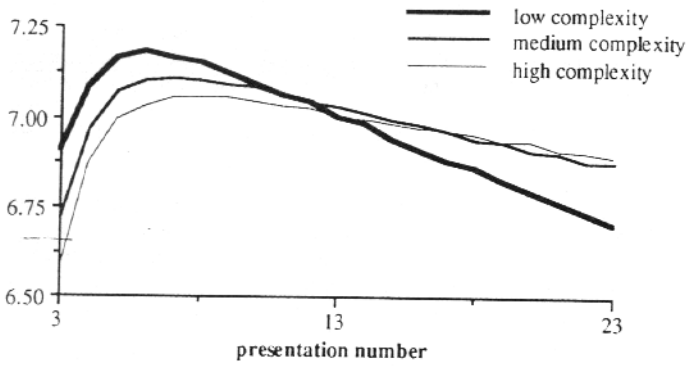

Fig. 4.17. Ativação em função do número de exposição para melodias de baixa, média e alta complexidade. (In: KATZ, 1994, p.216)

Todas as melodias apresentaram uma curva na forma prevista de *U* invertido. O aumento inicial, segundo Katz (1994, p.216) deve-se ao aumento do input da camada de grupo sobre as primeiras exposições, e a queda posterior deve-se ao aumento da atuação das conexões inibitórias da mesma camada. Outro fato observável é que a queda é mais expressiva para melodias menos complexas, o que confirma a suposição inicial.

Outro fato investigado por Katz (1994, pp.216-217) ainda nesta simulação, é que, segundo autores como Stevens e Latimer (1991), a repetição intercalada e não sucessiva de uma melodia reduz a queda do afeto. Um fator de desabituação foi adicionado a regra anti-hebbiana de Kohonen, zerando o valor inibitório das conexões quando elas não estão ativadas simultaneamente. Novamente o experimento foi repetido em dois modos, um com apenas a apresentação de uma melodia e outro com a alternância de 4 melodias de mesma complexidade. O gráfico aponta a diferença entre o valor máximo de ativação da camada de grupo de baixa-ordem após 6 apresentações e o valor final da ativação decorridas 20 apresentações. Os resultados apontam que existe uma menor queda de ativação quando as melodias são apresentadas alternadamente em todas as complexidades; mas, principalmente para melodias de alta complexidade, a queda é ainda menor.

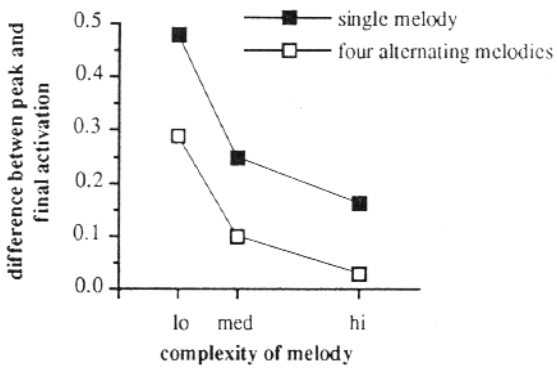

Fig. 4.18. Declínio da ativação na camada de grupo de baixa-ordem para apresentações de uma única melodia e de um grupo de 4 melodias. Diferença medida entre o valor máximo de ativação passadas 6 apresentações e o valor da ativação após 20 apresentações. (In: KATZ, 1994, p.217)

O terceiro teste visa averiguar o efeito da exposição a um gênero musical. Katz (1994, p.217) aponta que se pode prever que a exposição a um gênero musical específico gera uma maior apreciação de outras músicas dentro desse gênero. Podemos entender tal fato como uma familiaridade com as convenções técnicas de um estilo musical, que por sua vez proporciona uma maior e mais fácil compreensão de obras que compartilhem tais convenções. Tal familiaridade liga-se certos estilos étnico-musicais que indivíduos de uma cultura são mais expostos do que para estilos de outras culturas. Para se testar esta hipótese de familiaridade cultural com sua orelha para melodias, Katz (1994, p.217) constitui quatro conjuntos de treinamento, sendo um com 4 músicas jamaicanas, outro com 4 *negro spirituals*, mais 4 músicas irlandesas e 4 músicas chinesas. Todas as melodias foram selecionadas randomicamente dentre de coletâneas. Cada melodia foi estudada em dois aspectos: o efeito da primeira exposição dela fora de seu contexto étnico-musical; e o efeito da primeira exposição dentro de seu contexto étnico-musical. A rede foi treinada com três peças de um conjunto e depois exposta a outra peça deste conjunto e, também, a uma das peças de cada um dos outros conjuntos. Em cada uma das 16 etapas do teste a ativação média da camada de grupos de baixa-ordem foi medida, conforme demonstra a tabela abaixo:

Training set	Test set			
	Jamaican	Spiritual	Irish	Chinese
Jamaican	**0.39**	0.04	0.04	0.07
Spiritual	0.02	**0.09**	0.03	0.06
Irish	0.03	0.05	**0.17**	0.05
Chinese	0.05	0.06	0.05	**0.26**

Tabela 1. Aumento da ativação média para estilos melódicos em função da primeira exposição. (In: KATZ, 1994, p.218)

Claramente existe, no gráfico acima, um aumento de ativação intra-gênero, mesmo que mais expressivo para alguns estilos como o jamaicano do que em outros como o *negro spirituals*, mas que confirma a hipótese de Katz sobre a facilitação de apreensão dentro de um mesmo gênero. "*The primary reason for positive transfer within a genre is the similarity between groups in different songs of the same genre*" (KATZ, 1994, p.218). O efeito é gerado, segundo Katz (1994, p.218), porque a primeira exposição aumenta a resposta para outra melodia com um alto grau de similaridade, da mesma forma que algumas poucas repetições literais, como apontou o teste anterior, também a aumentam.

O último teste visa elucidar o efeito que melodias tradicionais em uma forma degradada causam à rede. O intuito aqui é verificar a resposta da rede para melodias reais perante a resposta de humanos. Poderia-se, então, calibrar todos os parâmetros da rede para simular a resposta de humanos, o que seria uma tarefa bastante longa. Dessa forma, Katz (1994, p.219) elabora um teste alternativo sobre melodias alteradas por permutações randômicas; ao invés de se alterar manualmente os parâmetros da rede alteram-se as melodias.

> "*One can then also show that the model's response to degraded versions of these melody decreases with the degree of degradation. If one makes the assumption that subject's affective response will also be less to a stochastically degraded version of a good melody, then the argument for a agreement between the model's judgment and that of human subjects can be further strengthened*" (KATZ, 1994, pp.219-220)

As duas melodias utilizadas neste teste são *Polly Put the Kettle On* e *Greenleeves*. Tais melodias foram degradadas pela substituição de algumas de suas

notas (de 0 a 25), tanto para alturas quanto durações, por outras notas determinadas randomicamente dentro de suas tessituras originais e dentro das durações originalmente utilizadas. Como se pode esperar, os resultados mostram uma queda na ativação média em função da quantidade de notas alteradas, o que sugere que ouvintes apresentariam menos afeto ao ouvirem músicas que tendem a um comportamento estocástico do que para músicas com alta similaridade e unidade entre seus diversos níveis e elementos dentro destes níveis. Os gráficos abaixo apresentam os resultados para as duas melodias.

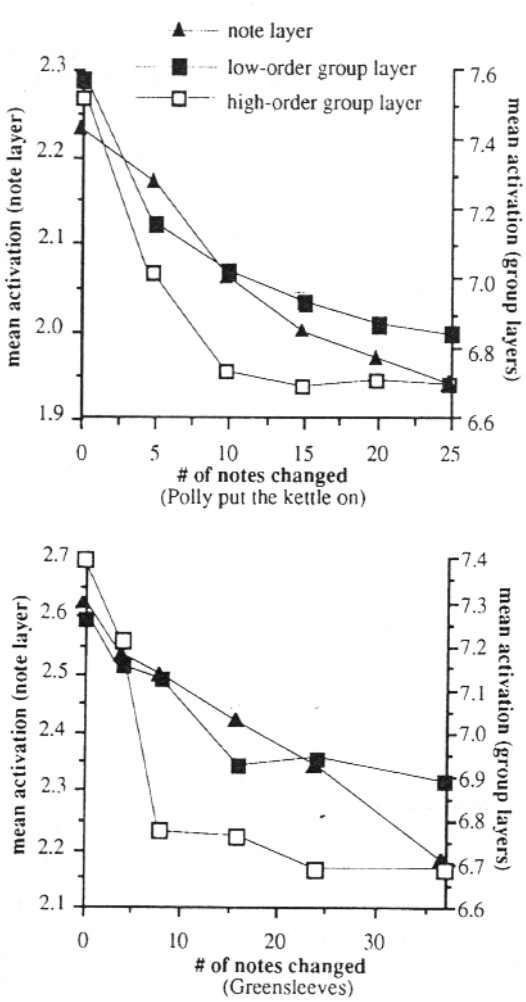

Fig. 4.19. Avaliação de duas melodias em função do grau de degradação. (In: KATZ, 1994, p.221)

Com estes experimentos Katz (1994) buscou estabelecer bases para justificar sua teoria do afeto musical como resposta de um sujeito perante a exposição de determinadas estruturas musicais, especialmente àquelas que apresentam similaridade

e unidade em sua constituição. Ele também buscou demonstrar "(…) *that the model captures at least some aspects of what a melody needs in order to be enjoyed*" (KATZ, 1994, p.220). No entanto, ele ressalta (KATZ, 1994, p.223):

> *"Moreover, this approach is not psychologically realistic. Although people are influenced by the judgments of others in their taste of music, in general, music does not come pre-labelled; evaluation is based on one's internal, affective response to the music. Developing an ear for melody will depend on understanding the basis for this response; this paper has argued that this can be accomplished by a connectionst operationalization of the principle of unity in diversity".*

4.2.2.3 Um modelo para a percepção tonal.

Em 1991, Marc Leman publica o artigo intitulado "*The Ontogenesis of Tonal Semantincs: Results of a Computer Study*", sobre o qual vamos nos concentrar nesta seção. Como o próprio título sugere, utiliza-se aqui uma rede neural artificial para o estudo da ontogênese da semântica tonal. De Outra forma, podemos dizer que o estudo visa estabelecer como um sistema auto-organizado desenvolve relações significativas entre acordes num contexto musical tonal. Leman (1991, p.100) define semântica tonal como "*a system of relations and meanings between tones within a context*". Neste sentido, para um estudo da semântica tonal deve-se considerar aspectos sensoriais (perceptuais), psicológicos, culturais, de aprendizagem, e sintático-musicais. Este fato é corroborado pela direção, apontada por Leman (1991, p.100), a uma musicologia cognitiva[4] quando propõe que atualmente temos meios mais adequados para considerar todos estes níveis de análise – "*Due to recent developments in psychology, neurobiology, and computer science, we now have a more powerful means to test this hypothesis in a more scientific way*" (LEMAN, 1991, p.100).

[4] Estamos entendendo o termo musicologia cognitiva no sentido de que Otto Laske (1992) lhe confere. Como, muito resumidamente, uma área que interessa tanto à ciência cognitiva quanto à musicologia, no sentido que, para a primeira, se mostra como um campo onde "*it might be able to elucidate, in a emprirical way, the limitations of the contemporary cognitive science*" (LASKE, 1992, p.4); e para a segunda, ela pode suplantar o que tradicionalmente faltou à musicologia: "*What has been lacking is a core set of methods shared by all its inquiries, as well as adequated tools for testing hypothesis*" (LASKE, 1992, p.5)

Leman (1991, p.101) aponta, que superando o reducionismo físico típico das abordagens iniciais (partindo de Helmholtz), um caminho levando a consideração de aspectos cognitivos relacionados à tonalidade passou a dominar a perspectiva destes estudos, assumindo-se uma interpretação qualitativa de dados quantitativos. Tal interpretação, típica em áreas como a psicoacústica, relaciona medições de parâmetros manifestados comportamentalmente com "*the listener's musical knowledge representations and cognitive information processing*" (LEMAN, 1991, p.101). Leman, contudo, analisa tais pressupostos epistemológicos e metodológicos:

> "*It should be noted here that the hypothesis fo an internal representation of tonal organization is based on an analysis of stimulus-response observations and not typically on an analysis of the sensory aspects of the acoustic signal and the ear. The concept of an internal representation of tonality, often used in the literature, therefore only makes sense within an broader paradigm of cognitive research – one in which mechanisms at lower levels are assumed although not taken into consideration to explain how this representation might come into existence.*
>
> *Yet there is some support in favor of a hypothesis that cognitive consonance might be due to the internalization of the statistical distribution of tones in the musical environment.*" (LEMAN, 1991, p.102)

Então, devemos buscar por um modelo cognitivo para a tonalidade, que explique as funções e relações tonais dentro de um paradigma cognitivo, levando em conta aspectos perceptuais, ambientais (contextuais), representacionais, e que pressuponha mecanismos de baixo-nível que suportem de alguma forma a emergência de tais funções e relações tonais. Leman (1991, pp.102-103) levanta alguns problemas relacionados aos modelos tradicionalmente usados pela ciência cognitiva para a explicação da tonalidade. Primeiro, quanto aos modelos computacionais baseados em regras (IA), não se pode dizer que eles expliquem a emergência da organização tonal na memória de um ouvinte. Tais modelos envolvem uma programação direta de modelos sintáticos da organização tonal – "*This approach is very ad hoc and questionable from an epistemological point of view. There in no learning involved, and so there is no theory about the ontogenesis of tonality functions*" (LEMAN, 1991, p.103). Quanto aos modelos baseados em aprendizagem supervisionada, nas quais existe um algoritmo que busca por uma configuração ideal dos pesos das conexões da rede estabelecendo uma associação entre *input* e *output*, não existe para Leman (1991, p.103) um postulado definitivo referente ao que tais modelos esclarecem sobre as

representações internas da organização e funções tonais. Novamente, existe uma determinação ad hoc e arbitrária entre *input* e *output*, entre quais as relações que a rede deve apresentar perante os acordes e/ou tons sucessivos e como deve classificar tais relações.

Outro ponto gerador de complicações que Leman (1991, p.103) aponta, em concordância com outros pesquisadores como Barucha (1991), é o tipo de representação normalmente utilizado nas modelagens conexionistas de atividades musicais. Tal tipo de representação normalmente empregado em RNAs, como já pudemos verificar, é o *pitch-class*, ou tipos correlatos deste como o *pitch-height*.

> "*The world in which these models [with pitch-class-like representations] operate is atomistic and Cartesian. The input representation is typically characterized by a local representation of pitch classes (much in the sense of a symbolic-based representation) and does not promise a very easy elaboration towards the processing of music "as it sounds"*." (LEMAN, 1991, p.103) (aspas do autor)

Leman (1991, p.103) aponta que a investigação cognitiva deve abandonar estas abordagens problemáticas apontadas nos parágrafos acima, principalmente investigações de aspetos como a ontogênese de funções tonais, ou mesmo outras questões relacionadas à música, que envolvem ou deveriam envolver aspectos perceptuais. Isso implica em utilizar-se representações realmente sub-simbólicas, não predefinidas; dessa forma o sistema pode desenvolver um comportamento próprio de resposta perante os estímulos externos e gerar representações por conta própria, motivadas pela interação estabelecida com o ambiente de acordo com suas restrições e possibilidades físicas. Acreditamos que a citação abaixo é bastante esclarecedora dos pontos de vista epistemológico e metodológico de Leman:

> "*This criterion embodies the idea that a system develops tonal semantics only in virtue of the response of the system to the environment. Stated differently, the tones encountered acquire meaning solely because they are relevant for the action of the organism in the environment*.
>
> *This further involves (a) that tonal functions are built up by a process of self-organization on the basis of the detection of invariant features in the environment (there is no external programmer except the environment), and (b) that the meaning of the system's response can only be know by virtue of the information given in the environment. The first statement is a rather general epistemological point: knowledge is built up by organizational*

> *principles inherent in the system and stimulus information provided by the environment. The second statement is a methodological one. The methodology implied is ecological, meaning that the system can be know only by virtue of the environment in which it is embedded. There is no way to understand the system just by looking at its memory. Together these statements propose that after the system has adapted itself to invariant information in the environment, it should be tested in order to discover the map of its self-organized output responses.*" (LEMAN, 1991, p.103)

Estando clara a postura de Leman sobre as implicações epistemológicas e metodológicas que o conexionismo deve apresentar para o estudo de atividades e fenômenos musicais, vamos ver em mais detalhes sua própria proposta de modelagem para a investigação da ontogênese das funções tonais. Em seu modelo, Leman (1991, p.103) baseia-se no pressuposto que funções tonais podem resultar de um mapa cortical auto-organizado, entendido como uma representação topográfica gerada por filtros neurais (ou unidades sintonizadas a certas características do estímulo) numa memória distribuída[5]. Estes mapas funcionam como um tipo de processo de ressonância de um sistema respondendo a estímulos ambientais, sendo que os sinais físicos adquirem significado porque são relevantes para guiar a ação do organismo num ambiente (LEMAN, 1991, p.l04). Grosso modo, podemos dizer que a proposta conexionista de Leman (1991) é inspirada no paradigma da percepção-ação, dentro de uma compreensão ecológica da percepção[6].

Tendo-se em conta o pressuposto dos mapas corticais auto-organizados, uma arquitetura conexionista em especial parece ser adequada para a investigação de Leman (1991): Os mapas auto-organizados de Kohonen (SOM – *Self-organizing Maps*). As redes SOM são chamadas por Leman (1991, p.104) de *Kohonen Feature Map* (KFM), enquanto que por Haykin (1994, p.408) são chamadas de Self-organizing Feature Maps (SOFM). Contudo, em concordância com as idéias de Churchland e Sejnowski (1991, p.136-137), Leman (1991, p.104) afirma que:

> "*The KFM method, however, is far from being an attempt to model real neural dynamics. As is the case with most neural networks models, the network mechanisms adopted are still too general and too abstract to count as a real model of the brain. Still, artificial neural networks approaches like the KFM are attractive because*

[5] Estes mapas também são conhecidos como mapas de características ou mapas cognitivos (LEMAN, 1991, p.104)

[6] Trataremos em mais detalhes a perspectiva ecológica da percepção e o paradigma da percepção-ação no capítulo terceiro desta dissertação.

> *they can more readily relate to cortical information processing and empirically-based brain research (...).*"

Vamos apresentar uma sucinta descrição de tal arquitetura não-supervisionada para vermos os motivos da adequação mencionada acima, utilizando por convenção a nomenclatura de Leman (1991) de KFM. O objetivo principal de um KFM é estabelecer uma representação dimensionalmente reduzida do conjunto de entrada da rede, sendo que tal redução acarreta na eliminação de dados redundantes e de ruído dos vetores de entrada. Existe um mapeamento de um espaço *n*-dimensional num espaço bidimensional. Topograficamente o KFM é um arranjo bidimensional de *n* x *n* unidades, todas conectadas a todas as unidades de entrada da rede. A conexão entre as duas camadas (entrada e a KFM propriamente) possui um peso de conexão. A ativação de cada unidade é a soma ponderada das unidades de entrada multiplicada pelos pesos sinápticos, cujo valor resultante deve passar de um limiar preestabelecido para ativar determinada unidade. Com essa estrutura, quando um vetor de entrada é apresentado à rede, um conjunto de unidades vizinhas irá responder a ele, mas apenas uma delas será a vencedora, sendo ela aquela que estiver no centro da área de resposta. Através de conexões de inibição lateral, o tamanho da vizinhança ativada por um vetor de entrada irá reduzir-se, até que apenas uma unidade vencedora represente aquele vetor de entrada.

Após a apresentação de todo o conjunto de padrões de entrada, a rede topologicamente representa a categorização de tal conjunto, onde cada unidade responde para cada padrão ou vetor de entrada. Unidades localizadas proximamente representam padrões similares, e vice-versa. Todo este processo é realizado por um procedimento algoritmo não-supervisionado; e, por não apresentar uma determinação para qual deve ser a correspondência direta a ser estabelecida entre entrada e saída da rede classifica-se esta como auto-organizada. Tem-se, então, um mapa (auto-organizado) topologicamente de acordo com características do conjunto de entrada. Nas palavras de Leman (1991, p.104):

> "(…) *a particular reduced dimensionality topological organization of the input data can be discovered, similar to the spatial organizations found by multidimensional scaling. Finally, this organization can be hypothesized as a possible psychological structure in the minds of experimental subjects and thus human music listeners.*"

Tal afirmação está em concordância com Haykin (1994, p.419):

> "*The topological ordering property of the SOFM [KFM] algorithm (...) makes it a valuable tool for the simulation of computational maps in the brain. Indeed, the self-organizing feature maps are perhaps the simplest model that can account for the adaptative formation of such topographic representations* (…)."

A rede projetada por Leman (1991, p.106) é um arranjo bidimensional de 20 x 20 unidades, num total de 400, implementada num sistema Transputer com quatro processadores em paralelo.

Além de aspectos relacionados à arquitetura da RNA utilizada por Leman (1991), devemos fazer alguns esclarecimentos referentes à forma de representação envolvida neste estudo. São considerados aspectos sensoriais e culturais na forma de representação dos dados. Os sensoriais estão baseados na teoria psicoacústica de Terhadt et al. (1982); enquanto que os culturais refletem o tipo de dado envolvido assim como sua distribuição estatística na música ocidental, de acordo com a teoria de Bhrun (1988).

Assume-se como pressuposto psicoacústica a existência de padrões subharmônicos, que consistem de vários subharmônicos gerados por processos perceptuais de análise freqüêncial sobre componentes senoidais extraídos de um som complexo. Ao se ouvir um som complexo, uma única altura subjetiva é percebida, pela comparação dos padrões de subharmônicos encontra-se qual subharmônico ocorre mais freqüentemente em resposta a um som complexo. Esta altura subjetiva é chamada de altura virtual (*virtual pitch*).

Esse processo de extração e comparação de subharmônicos é inserido na representação de altura envolvida no estudo de Leman (1991, p.107). O vetor distribuído que representa um acorde é computado pela combinação dos padrões de subharmônicos para cada nota do acorde, num sistema de *pitch-class*. Existem 12 unidades de entrada da rede, cada uma para uma nota da escala cromática, cuja ativação é determinada pela soma ponderada dos subharmônicos correspondentes a cada nota. A Figura abaixo ilustra a representação distribuída de um acorde de sétima de dominante.

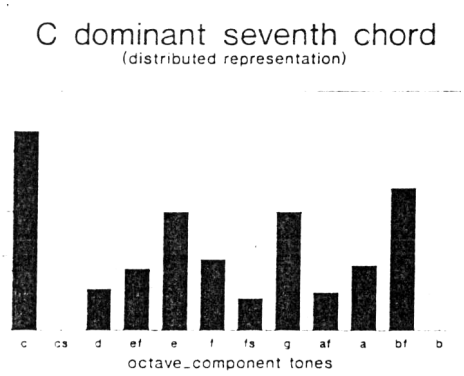

Fig. 4.20. Representação distribuída de um acorde de sétima de dominante, baseada na teoria de altura virtual. (In: LEMAN, 1991, p.108).

Podemos, agora, passar a descrição de três estudos que Leman realizou para investigar a ontogênese de funções tonais num sistema auto-organizado. O primeiro estudo utiliza um conjunto de 115 acordes, preparados pela computação das alturas virtuais, que são: 12 tríades maiores; 12 tríades menores; 12 tríades diminutas; 4 tríades aumentadas; 12 tétrades maiores com sétima maior; 12 tétrades menores com sétima; 12 tétrades de sétima de dominante; 12 tétrades meio-diminutas; 12 tétrades aumentadas com sétima; 12 tétrades menores com sétima maior; e 3 tétrades diminutas com sétima diminuta. A ordem de tais acordes no conjunto de treinamento foi randomicamente altarada a cada apresentação para a rede. O padrão de comportamento global (padrões de ativação) da rede foi estabelecido pela representação topográfica a cada época do treinamento, que abarcou um total de 180 épocas. Abaixo podemos ver tal padrão global para um acorde de dó maior, nas épocas 1, 10 e 180.

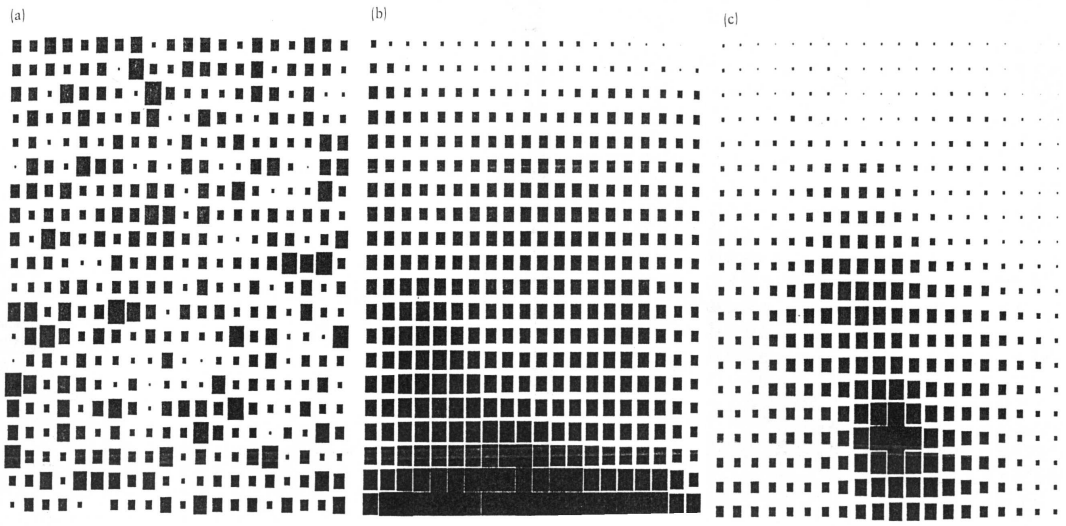

Fig. 4.21. Padrões de ativação da rede para um acorde de dó maior, após (a) a primeira época, (b) 10 épocas e (c) um acorde de fá maior após 180 épocas. O tamanho de cada unidade é diretamente proporcional ao seu valor de ativação. (In: LEMAN, 1991, pp.110-111)

Logo após a inicialização da rede, suas conexões estão ainda muito próximas da distribuição randômica inicial, mas conforme o conjunto de acordes vai sendo repetidamente apresentado à rede, uma estrutura topográfica organizada começa a ser observável, pela constante adequação dos pesos das conexões e da área de vizinhança do neurônio vencedor para cada acorde. "*If we think of each neuron in the KFM grid as a kind of filter on the inputs, then all the neurons in the response region can be thought of as being tuned more or less roughly to the particular input pattern*" (LEMAN, 1991, pp.110-111). Duas noções centrais estão envolvidas na análise do comportamento da rede, pelos padrões de ativação: o neurôdo característico (NC) e a região de resposta (RR). O NC é a unidade com maior valor de ativação para cada padrão de entrada; enquanto que a região de unidades ativadas por um padrão de entrada é chamada de RR. Se rotula-se cada NC de acordo com o padrão de entrada ao qual ele responde, obtém-se o mapa global de reposta da rede para todo conjunto de acordes.

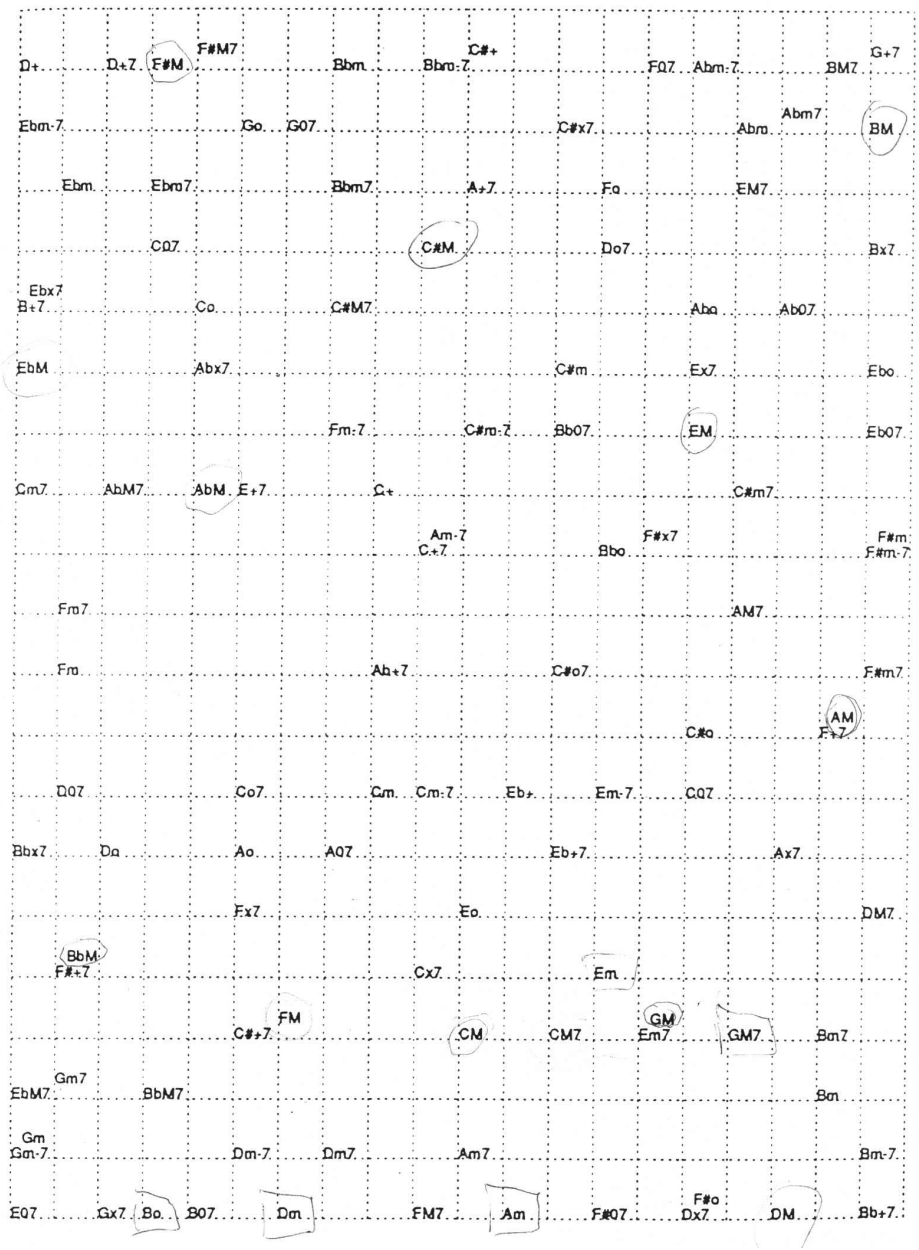

Fig. 4.22. Padrões globais de resposta do KFM para o conjunto de 115 acordes, após 180 épocas. (In: LEMAN, 1991, p.115)

Na análise da Figura acima, deve-se ter em mente que "*similarity means smaller distance*" (LEMAN, 1991, p.112). Vamos, por exemplo, verificar a análise de Leman das disposições dos acordes maiores neste mapa global de resposta. Primeiro, a organização resultante aproxima-se bastante daquela do círculo de quintas, se observarmos os NC dos acordes maiores, assim como dos menores, das dominantes etc. Contudo, esta relação com o círculo de quintas não é uma necessidade absoluta para a explicação das funções tonais, afirma Leman (1991, p.112), mas ela surge freqüentemente nas simulações realizadas (LEMAN, 1991, p.117). Tais circularidades

de distribuição de NC dependem da forma das RRs surgidas na rede. E, as RRs podem, por exemplo, apresentar-se divididas em duas partes sobre lados opostos da topologia da rede, sendo o mapa um contínuo pela união dos lados opostos. Leman (1991, p.117) afirma que:

> "*The fact that the neural network came up with some kind of organization anyhow should therefore be explained on the basis of the tonal stability of the input patterns, which is reflected in the error between the CNs and their corresponding input vectors. Tonal stability means that the input patterns can be clearly distinguished from each other. But this implies as well that some chords are more stable (more distinct) than others.*"

O fato do círculo de quintas ter freqüentemente aparecido nas simulações sugere, talvez, que as relações entre acordes em tal ciclo tenham justificativas psicológicas e neurológicas. Da mesma forma, as demarcações topográficas das RRs podem justificar a facilitação perceptual entre certos acordes; cada RR inclui unidades que também respondem a outros acordes. Algumas RRs de acordes tonalmente relacionados apresentam uma área de sobreposição, o que pode ser interpretado como uma explicação para como um contexto tonal pode ser estabelecido e como certos tons neste contexto adquirem suas funções tonais (LEMAN, 1991, p.118). Quando acordes tonalmente relacionados são tocados seqüencialmente, as unidades incluídas nas duas RRs (área de sobreposição) ficarão atividades continuamente, possibilitando a facilitação perceptual para o reconhecimento de relações tonais entre os dois acordes. A disposição topográfica dos acordes e as relações entre os acordes por ela elucidadas podem ser entendidas como um mapa psico-neurológico do contexto tonal.

O segundo estudo é bastante semelhante ao primeiro, porém incluindo no conjunto de acordes uma distribuição estatística da ocorrência de cada tipo de acorde na música clássica e romântica. Desta forma, o conjunto apresenta a seguinte distribuição probabilística: tríades maiores (43%); tríades menores (15%); tríades diminutas (4%); tríades aumentadas (1%); tétrades maiores com sétima maior (1%); tétrades menores com sétima (1%); tétrades de sétima de dominante (26%); tétrades meio-diminutas (2%); tétrades aumentadas com sétima (1%); tétrades menores com sétima maior (1%); e tétrades diminutas com sétima diminuta (7%). Por exemplo, as 12 tríades maiores ocorreram 43 vezes cada no conjunto de treinamento, as 12 tríades

menores 15 vezes cada, e assim por diante, resultando num conjunto total de 1156 acordes.

A rede foi, como no primeiro estudo, apresentada ao conjunto completo de acordes (em ordem randômica) 180 vezes. Os resultados apresentados foram extremamente próximos daqueles do estudo anterior. Vale, contudo, ressaltar que a medição do erro pela diferença entre os vetores sinápticos dos NCs pelos seus vetores de entrada revelou ser menor para aqueles acordes que são estatisticamente mais comuns.

> "*The errors reflects the stability of the network's response; thus, in general we observe that those chords that have a high frequency of occurrence have a stable response in the network. Of course, one should mention here that these common chords are stable and clearly distinguished from a traditional tonality point of view as well.*" (LEMAN, 1991, p.119)

O terceiro estudo visa clarificar a relação entre o mapa topográfico e o tipo de representação utilizada para a elaboração dos padrões de entrada. Enquanto que no primeiro estudo os padrões foram elaborados num processo 'manual' sobre a teoria de Terhardt et al. (1983); neste terceiro estudo eles foram confeccionados a partir de dados acústicos, pelo processo desenvolvido por Parncutt (1989). Foram computados os perfis de probabilidade croma (*chroma probability*), que consiste na probabilidade de um tom em particular (*pitch-class*) ser percebido numa passagem de acordes. Em mais detalhes, "*the output of the computation leads to values for pitch classes that are quite similar to (though slightly different from) the values obtained by our first study based on artifical data*" (LEMAN, 1991, p.119). O conjunto de acordes totalizou 91 deles, os 115 do primeiro estudo menos as 12 tétrades aumentadas com sétima e as 12 tétrades menores com sétima maior. A inspeção dos mapas gerados após 180 épocas de apresentação do conjunto de acordes revela alterações, como era de se esperar. A relação topográfica circular presente anteriormente não apareceu nesta simulação, apesar dos valores de erro serem bastante similares àqueles apresentados em simulações anteriores (LEMAN, 1991, p.121). Como as relações entre NCs nos mapas topográficos só podem efetivamente ser consideradas sobre distâncias curtas, talvez esta última simulação apresente algum tipo de relação entre CNs mais afastados geometricamente, mas que não pode ser tomada como garantida.

Após realizar os três estudos por nós descritos, Leman (1991, p.122) ressalta que:

> "*The importance of this work is in showing that aspects of tonality can in principle be accounted for internal representations that develop through self-organization from invariant features in the musical environment. The current model shows that it is possible to adopt an epistemological and methodological approach to the modeling of tonality that is in close agreement with Gestalt psychology and ecological theories of perception.*
>
> *There is also a predictive aspect to our results. If, disregarding the abstractions inherent in neural network simulations, self-organization does indeed makes sense, we think this give impetus to look for chordal functions at a neuronal level. Experimental research on human cortical functions would be needed to prove the existence of a circle-of-fifths map in the brain*."

4.2.3 Modelando a composição musical.

Para ilustrar a utilização de RNAs para a composição musical vamos descrever o artigo "*Neural Network Music Composition by Prediction: Exploring the Benefits of Psychoacoustic Constraints and Multi-scale Processing*", de Mozer (1994). O autor define sua proposta como uma extensão da aplicação de tabelas de transição (ou cadeias de Markov) para a composição musical, utilizando uma arquitetura mista *feedfoward* e *feedback*. A rede foi batizada de CONCERT (*CONnectionist Composer of ERudite Tunes*) e sua tarefa é a criação de melodias com acompanhamento harmônico e permitir uma análise da eficiência dos métodos de geração nota-a-nota baseados em tabelas de transição. CONCERT é um tipo de rede supervisionada que, como uma tabela de transição, precisa ser exposta a um conjunto de obras para realizar um levantamento estatístico das probabilidades de ocorrência de um evento de acordo com os eventos anteriores. As informações musicais são representadas nos aspectos à altura, duração e estrutura harmônica, de acordo com estudos psicológicos da percepção humana (MOZER, 1994, p.228).

Tabelas de transição são aplicadas à composição automática de música desde os primeiros programas da IA para este fim, como já apontamos no primeiro capítulo. Mozer (1994, pp.228-230) aponta alguns problemas típicos das composições geradas por tabelas de transição, como a escolha da ordem apropriada para a tabela (a

quantidade de eventos anteriores utilizados para a determinação probabilística de eventos subseqüentes). Para garantir-se um output com coerência motívica pode ser utilizada uma tabela de ordem alta (por exemplo, terceira-ordem); ao contrário, se a ordem for baixa a tabela não será sensível ao contexto estatístico das obras analisadas e o resultado será próximo da geração randômica. Mas, ordens altas implicam em alguns problemas. Primeiro, o tamanho da tabela aumenta exponencialmente conforme se aumenta a ordem, podendo chegar a quantidades além daquelas razoavelmente manipuláveis por computadores de pequeno porte. Segundo, tabelas de transição de alta ordem deixam de lado as estruturas de baixa-ordem de *corpus* de obras analisado. Mozer (1994, p.230) aponta ainda que sistemas simbólicos, apesar de poderem ser construídos sobre processos probabilísticos, devido a sua natureza (simbólica) não são sistemas adequados para lidar com propriedades estatísticas de dados e nem podem generalizar a partir dos dados probabilísticos obtidos. Outra complicação é decorrente do fato de que tabelas de transição não podem estabelecer relações entre elementos que não são vizinhos numa seqüência, e normalmente m música nem sempre as relações entre elementos vizinhos são as mais importantes, já que muitas notas têm um caráter mais ornamental do que estrutural. Com relação a todos estes aspectos, Mozer (1994, p.230) afirma:

> "*The connectionist approach, however, is far more flexible in principle: the form of the transition function can permit the consideration of varying amounts of context, the considerations of non-contiguous context, and the combination of low-order and high-order regularities.*
>
> *The connectionist approach also promises better generalization through the use of distributed representations* (…). *In a local representation, where each note is represented by a discrete symbol, the sort of statistical contingencies that can be discovered are among notes. However, in a distributed representation, where each note is represented by a set of continuous feature values, the sort of contingencies that can be discovered are among features.*"

Em resumo, podemos dizer que a proposta de Mozer (1994) é aliar a análise estatística das tabelas de transição com os mais flexíveis poderes computacionais gerativos das RNAs, levando a uma verificação última da eficiência da composição autômata de música por geração nota-a-nota. Mozer (1994, p.231) afirma que apesar da música apresentar vários níveis hierárquicos desde as notas até obras como um todo, passando por níveis de frases, períodos etc., todos estes níveis constiuem-se de

um conjunto finito relativamente pequeno e não-ambíguo de elementos, que apresentam muita regularidade estilística e psicoacústica. Dessa forma, um sistema de geração nota-a-nota pode inferir menos alguns destes níveis de estruturação por uma técnica linear de geração. Contudo, para a verificação da eficiência desta geração, em vez de uma análise em termos computacionais objetivos, Mozer (1994, p.231) defende uma análise pela audição da música gerada automatamente, um procedimento que podemos entender como um teste de Turing musical.

Podemos prosseguir agora numa descrição mais detalhada da arquitetura de CONCERT, e depois da representação da informação musical empregada nesta rede neural. CONCERT é uma rede *feedfoward* com quatro camadas, sendo que uma delas (a camada contexto) possui conexões *feedback*. A camada de entrada representa notas musicais apresentadas seqüencialmente, pelos parâmetros de altura, duração e acompanhamento harmônico. A informação da camada de entrada alimenta a camada de contexto, que é responsável pela realização dos processos estatísticos das tabelas de transição de *k*-ésima ordem, de acordo com as configurações da rede. A este respeito Mozer (1994, pp.232-233) afirma que:

> "(…) *the architecture is more general than a transition table because [the function employed] is not limited to implement a stack and the map from the context layer to the output need to be an arbitrary look-up table. From the myriad possibilities, the training procedure attempts to find a set of connections that are adequate for performing the next-note prediction task. This involves determining which aspects of the input sequence are relevant for making future predictions and constructing the function f appropriately. Subsequently, the context layer will retain only task-relevant information*."

As camadas superiores, como mostra a Figura 35 ,são responsáveis pela saída da rede explicitando o espaço probabilístico do próximo evento. A ativação da camada de contexto é encaminhada para a camada NND (*next-note-distributed*) que apresenta numa representação distribuída as possíveis notas subseqüentes; enquanto que a camada NNL (*next-note-local*) realiza uma tradução da informação distribuída (em NND) para categorias mais fácil análise, com unidades para cada altura, duração e acorde. A arquitetura geral pode ser vista abaixo.

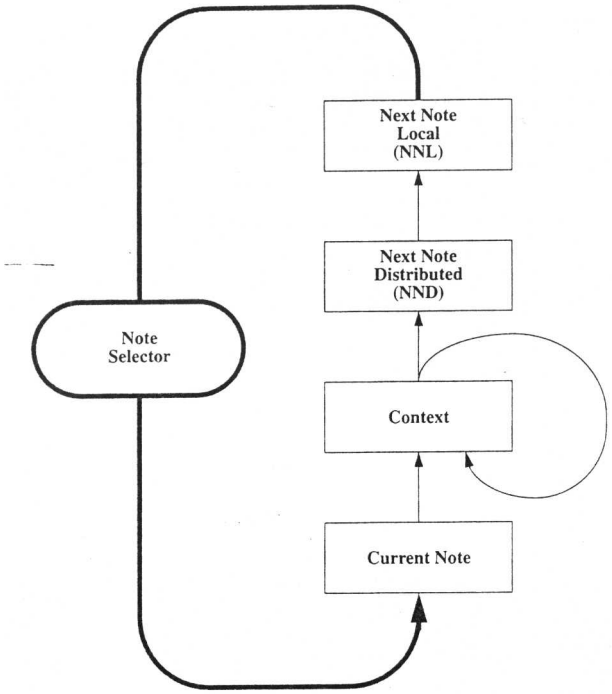

Fig. 4.23. Arquitetura de CONCERT. (In: MOZER, 1994, p.232)

Como já vimos anteriormente, em geral o algoritmo mais empregado para o treinamento de RNAs supervisionadas é o *back-propagation*, mas ele só se aplica à redes *feedfoward*. Tendo em vista que CONCERT apresenta conexões *feedback* foi utilizada uma variação do *back-propagation* adequada a este tipo de rede conhecida como *back-propagation through time* (BPTT). O BPTT transforma uma rede recorrente numa rede *feedfoward* equivalente, na qual adiciona-se uma nova camada a cada passo de tempo[7] (HAYKIN, 1994, p.520). Dessa forma, como numa rede *feedfoward*, pode-se alterar os pesos das conexões visando a redução do erro da saída da rede perante a saída idealizada.

Estando a rede treinada adequadamente, ela pode prever quais serão as notas mais prováveis de acordo com o contexto de *n* notas anteriores. Mas além deste poder de previsão a rede pode operar para a geração de música originais, que refletirão estatisticamente os procedimentos estilísticos (sintáticos) do conjunto de peças utilizado para o treinamento. Para tanto, entra em cena um módulo externo chamado de seletor de notas, apresentado na Fig. 35. A função deste módulo é retrogradar uma nota, dentro do conjunto de notas previstas pelo espaço probabilístico de saída, para a camada de entrada da rede. Dessa forma, a rede apresenta um procedimento de

[7] Esse processo não ocorre em tempo real.

geração nota-a-nota que resultará em uma peça musical original, mas dentro de um estilo específico.

Quanto aos aspectos da forma de representação empregada em CONCERT, podemos dizer que Mozer (1994) tentou superar muitos dos problemas apontados por Barucha (1991). Vamos tratar primeiro da representação da altura. Vários aspetos apontados como problemáticos quando se utiliza apenas um tipo representação, como a falta de equivalência entre notas separadas oitavas com o *pitch-height*, ou a similaridade entre notas dentro da relação do ciclo de quintas que não aparece em representação por *pitch-class*, falta de invariância sobre transposição, são possivelmente superados pela forma de representação apresentada por Mozer (1994, p. 236), baseada na teoria de Shepard (1987). O sistema representacional, então, constitui-se num espaço 5-dimensional que combina *pitch-height* (PH), *pitch-class* (*chroma circle*, CC) e ciclo de quintas (*circle of fifths*, CF), chamado PHCCCF. As cinco dimensões são: um ponto sobre a escala de PH, uma coordenada (x, y) sobre o CC e outra coordenada (x, y) sobre o CF.

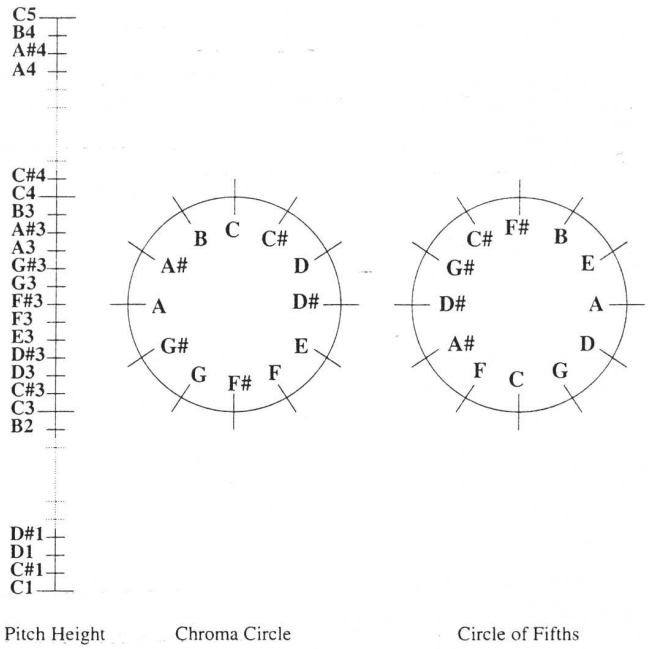

Fig. 4.24. Representação da altura. (In: MOZER, 1994, p.237)

A importância relativa dos várias representações envolvidas pode ser determinada por um escalonamento entre os diâmetros dos círculos relativos um ao outro e à escala PH. Para as simulações realizadas por Mozer (1994, p.238) o

diâmetro dos dois círculos foi igualado num valor equivalente a distância de uma oitava na escala de PH.

Entretanto, para a geração dos vetores de *input* da RNA, foi empregada uma tradução destes valores 5-dimensionais, que necessitariam de 5 unidades de entrada, numa representação sobre 11 unidades de entrada. A justificativa oferecida relaciona-se a influência que mesmo pequenas variações na atividade de uma unidade (ruídos) podem embutir no sistema, e que as funções de ativação sigmóides não preservaria a igual distribuição das notas no PH e nos CC e CF (MOZER, 1994, p.238). Dessa forma, cada coordenada dos círculos é representado por seis unidades booleanas (com valores –1 e +1). Como cada unidade pode ter apenas dois valores de ativação, elas são menos sujeitas a variações causadas por ruído. Porém, não é possível uma forma semelhante para a representação do PH, a não ser com a utilização de 49 unidades distintas, uma para cada nota entre C1 e C5. Assim, uma única unidade é responsável pela codificação da informação de PH respondendo para um range de valores discretos entre –9.789 a +9.789. Como a escala de PH serve apenas para a codificação da oitava (registro), e relações entre notas dentro de uma oitava são abarcadas pelas outras dimensões., uma representação extremamente precisa neste caso não é necessária, a influência do ruído aqui não exerceria uma alteração significativa na representação como um todo (PHCCCF) (MOZER, 1994, pp.238-239).

Tone	Representation					
C	-1	-1	-1	-1	-1	-1
C#	-1	-1	-1	-1	-1	+1
D	-1	-1	-1	-1	+1	+1
D#	-1	-1	-1	+1	+1	+1
E	-1	-1	+1	+1	+1	+1
F	-1	+1	+1	+1	+1	+1
F#	+1	+1	+1	+1	+1	+1
G	+1	+1	+1	+1	+1	-1
G#	+1	+1	+1	+1	-1	-1
A	+1	+1	+1	-1	-1	-1
A#	+1	+1	-1	-1	-1	-1
B	+1	-1	-1	-1	-1	-1

Pitch	PH	CC						CF					
C1	-9.798	-1	-1	-1	-1	-1	-1	-1	-1	-1	+1	+1	+1
F#1	-7.349	+1	+1	+1	+1	+1	+1	+1	+1	+1	-1	-1	-1
G2	-2.041	+1	+1	+1	+1	+1	-1	-1	-1	-1	-1	+1	+1
C3	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	+1	+1	+1
D#3	1.225	-1	-1	-1	+1	+1	+1	+1	+1	+1	+1	+1	+1
E3	1.633	-1	-1	+1	+1	+1	+1	+1	-1	-1	-1	-1	-1
A4	8.573	+1	+1	+1	-1	-1	-1	-1	-1	-1	-1	-1	-1
C5	9.798	-1	-1	-1	-1	-1	-1	-1	-1	-1	+1	+1	+1
rest	0	+1	-1	+1	-1	+1	-1	+1	-1	+1	-1	+1	-1

Tabela 2. (a) Representação dos tons em CC, e (b) representação PHCCCF de algumas alturas. (Adaptado de MOZER, 1994, pp.238-239)

Até agora foi mostrado como a representação PHCCCF é estabelecida para cada nota na camada de entrada, mas precisamos considerar a representação de várias notas ao mesmo tempo, visto que as camadas NND e NNL da rede apresentam não uma única nota, mas o conjunto de notas com maior probabilidade de ocorrência. Nesse sentido, é estabelecido um ponto no espaço de PHCCCF que é simultaneamente o mais próximo das notas envolvidas (o vetor médio). Mozer (1994, p.240) ainda afirma que apesar das vantagens oferecidas pela representação PHCCCF, ela ainda não é totalmente plausível em termos psicológicos pois relaciona apenas poucas notas em isolamento do restante de uma peça, enquanto que "*Listerners of music do not process individual notes in isolation; notes appear in a musical context which in suggests a musical key which in turn contributes to an interpretation of the note*." Contudo, a representação PHCCCF já apresentam avanços perante os problemas tradicionais da representação da informação musical em ambiente computacional.

A representação da duração é análoga a representação PHCCCF da altura, também empregando um espaço 5-dimensional. Esta representação está baseada na divisão de cada tempo em 12 partes, e o valor de cada nota é relacionado com as dimensões de *duration-height*, o círculo de 1/3 de tempo e o círculo de 1/4 de tempo. Estes dois círculos juntos à subdivisão em 12 partes da duração permitem a representação da subdivisão ternária e binária para cada tipo de figura musical (colcheia [6/12], semicolcheia [3/12], tercina de colcheia [4/12], tercina de semicolcheia [2/12]e assim por diante).

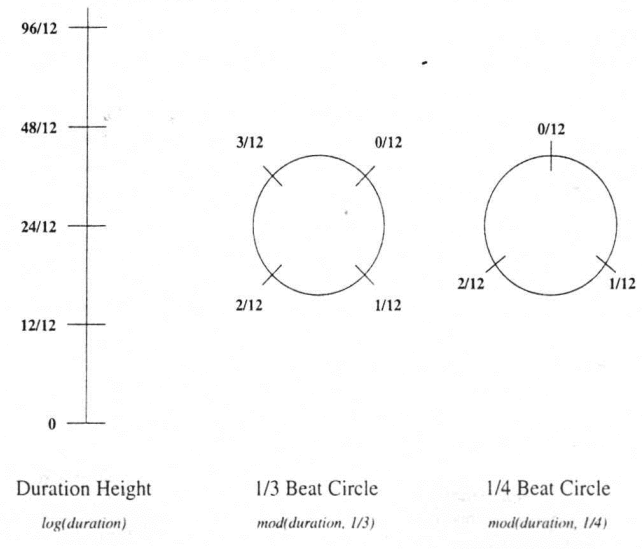

Fig. 4.25. A representação da duração em CONCERT. (In: MOZER, 1994, p.241)

O valor sobre a escala de *duration-height* é proporcional ao valor logaritmo da duração, baseado na lei geral da psicofísica que relaciona a intensidade de um estímulo com a sensação percebida logaritmicamente (MOZER, 1994, p.240). O ponto sobre o círculo 1/*n* é a duração subtraída do maior múltiplo integral de 1/*n*. Esta representação possibilita que durações próximas ocupem pontos próximos nos círculos 1/n, por exemplo, colcheias e semínimas ocupam o mesmo lugar no círculo 1/4, ou tercinas de colcheia e de semínima ocupam o mesmo ponto do círculo 1/3.

Diferentemente da representação PHCCCF, a representação 5-dimensional da duração pode ser representada diretamente em cinco unidades da camada de entrada, pois estão menos sujeitas à ruídos. Mozer (1994, p.241) afirma que por estarem menos povoados os círculos da duração pequenas variações na ativação das unidades não irão modificar a representação final.

Quanto a representação dos acordes do acompanhamento harmônico, Mozer (1994, p.242) utilizou apenas tríades e tétrades em posição fundamental codificados em *pitch-class*. Cada nota de um acorde foi considerada como a sobreposição da altura fundamental com seus quatro primeiros harmônicos. Ou, nas palavras de Mozer (1994, p.242): "(…) *a pitch was represented by activating the [units in pitch-class for the] fundamental and the first four harmonics. The representation of a chord consisted of the superimposition of the representation of the component pitchs*." O range de alturas coberto por CONCERT vai de C1 até C5, resultando em 49 diferentes *pitch-classes*, mas para os acordes, Mozer (1994, p.242) reduziu a representação para apenas uma oitava, requerendo, dessa forma, apenas 12 unidades. A nota fundamental foi codificada com um valor de ativação de 1.0, enquanto que os parciais com 0.5 para o primeiro, 0.25 para o segundo, e assim por diante. Estes valores num alcance de 0 a 1.0 foram re-escalonados para um limite de –1 a 1 para adequarem-se aos valores de ativação das unidades da rede. Ainda, para levar em consideração a similaridade entre acordes próximos de acordo com ciclo de quintas, "(…) *an addicional element was added to force these chords close together. The element had value +1.5 for these three chords [tonic, subdominant and dominat] (as well as C7, F7 and G7) [in C major key], -1.5 for all other chords*" (MOZER, 1994, p.243).

Agora, após a descrição da arquitetura e das representações de CONCERT, resta verificar o desempenho desta RNA para a previsão de estruturas musicais e para a composição musical. Inicialmente foram projetados cinco testes básicos para avaliar

CONCERT: estender uma escala de dó maior; aprender a estrutura de escalas diatônicas em vários tons; aprender a estrutura de seqüências diatônicas randômicas; aprender estruturas de seqüências randômicas interpoladas; e, aprender padrões de frases AABA.

Para o primeiro, CONCERT foi treinado com uma escala de dó maior sobre três oitavas para prever, a cada nota da escala, qual seria a próxima nota. Em 10 testes, utilizando 15 neurôdos na camada de contexto, CONCERT levou cerca de 30 apresentações da escala de treinamento para aprender a estrutura. Em seguida, a RNA foi apresentada a mesma escala em três oitavas para estender a estrutura uma oitava a mais. Os resultados mostraram que nas 10 tentativas ele atingiu os objetivos, exceto em 4 delas que a última nota foi transposta uma oitava abaixo.

O teste número dois, aprender escalas diatônicas maiores em vários tons, foi igualmente satisfatório. Sobre o alcance de alturas possíveis (C1 até C5) existem 37 escalas; o conjunto de treinamento englobou 28 delas selecionadas randomicamente, e o conjunto de treinamento as 9 restantes. CONCERT, com 20 unidades de contexto, chegou a um erro razoável após um treinamento de 55 apresentações. O percentual de acerto para o conjunto de teste foi de 98.4%. Em termos de comparação, uma tabela de transição para esta previsão atingiu apenas 26.6% de acerto. Segundo Mozer (1994, p.245) seria necessária uma tabela de 3ª-ordem para se alcançar bons resultados, porém muito mais exigente em termos computacionais do que CONCERT. Ainda, Mozer (1994, p.245) testou CONCERT utilizando representação por *pitch-class* em vez de PHCCCF, e a taxa de acerto caiu para 54.4%.

O terceiro teste aplica-se sobre seqüências randômicas construídas sobre a escala de dó maior, pela seguinte regra: "*the first pitch was selected at random, and then sucessive pitchs were either one step up or down the C-major scale from the previous pitch, the direction chosen at random*" (MOZER, 1994, p.245). CONCERT, com 15 unidades de contexto, foi treinada 50 vezes sobre um conjunto de 100 destas seqüências. Após o treinamento, outras 100 seqüências foram utilizadas para o teste. A porcentagem de acerto foi de 99.95%. Esse valor foi obtido pela análise do espaço probabilístico da saída da rede; se entre as possibilidades previstas estivessem as duas notas separadas por grau conjunto da nota atual um acerto era computado a avaliação.

Seqüências randômicas interpoladas foram os objetos do quarto teste. Cada seqüência de 10 notas tem a seguinte forma: a_1, b_1, a_2, b_2, …, a_5, b_5, onde a_i e b_i são notas randômicas e a_{i+1} é um tom acima ou abaixo de ai na escala de dó maior.

CONCERT, com 25 unidades de contexto, foi teve 50 épocas de treinamento sobre um conjunto de 200 seqüências, e depois foi testado sobre outras 100. A avaliação foi sobre a relação entre as notas a_i, da mesma forma que as b_i, mas não considerou a relação entre a_i e b_i, pois esta é impossível de ser prevista. Em 94.8% das notas foram previstas acertadamente.

O último desses testes básicos foi sobre estruturas musicais compostas de frases AABA. Cada seqüência foi composta por frases de 5 notas cada, em ordem ascendente cromática, com a primeira nota selecionada randomicamente. Este exemplo avalia a capacidade da rede tratar de aspectos de baixo-nível como de nível mais alto; as frases propriamente ditas possuem uma estrutura melódica clara, enquanto que existe uma relação de nível mais alto sobre as frases que obedece a estrutura AABA. Para este teste utilizou-se 35 unidades de contexto e um conjunto de treinamento de 200 seqüências repetidas 300 vezes, e mais outras 100 seqüências para o teste. Pela estrutura das frases, algumas notas podem ser inferidas mais facilmente pela estrutura local (da segunda a quinta, por exemplo), enquanto que outras requerem considerações sobre níveis mais altos. Dessa forma, a análise foi separada em dois quesitos: notas dependentes apenas estruturas locais e notas dependentes de estruturas de um nível superior. Para o primeiro caso, CONCERT obteve sucesso em 97.3% das notas; para o segundo aspecto a taxa de acerto foi de apenas 58.4%.

O último teste demonstrou a falta de capacidade da rede em lidar com os aspectos de alta-ordem sempre presentes em estruturas musicas. Se o intuito da rede é manipular e estatisticamente e gerar obras representativas dentro de um estilo musical este tipo de dificuldade deve ser superada. Mozer (1994, p.247) diz:

> "(…) *back-propagation is not sufficiently powerful, especially for contingencies that span long temporal intervals and that involve high-order statistics (...).*
>
> *This poses s serious limitation on the use of back-propagation to induce musical structure in a note-by-note prediction paradigm because important structure can be found at long time-scales as well as short. (...) Within a phrase, local structure can probably be captured by a transition table (...). (...)Across phrases, however, a more global view of the organization is necessary.*"

Mozer (1994, p. 248) desenvolve um procedimento alternativo para transpor esta dificuldade, alterando a função da ativação das unidades da camada de contexto.

O objetivo é gerar uma representação reduzida das seqüências de notas tornando os aspectos globais mais explícitos. Com um procedimento como tal a detecção de características globais torna-se uma tarefa de detecção de características locais, o que a rede já demonstrou capacidade de realizar. A função das unidades de contexto foi definida como:

$$c_i(n) = s\left[\sum_j w_{ij}\, x_j(n) + \sum_j v_{ij}\, c_j(n-1)\right] \qquad (1)$$

onde $c_i(n)$ é a atividade da unidade i no instante n, $x_j(n)$ é a atividade da unidade j de entrada no instante n, w_{ij} é o peso da unidade j da camada de entrada para a unidade i da camada de contexto, e v_{ij} é o peso da unidade j para a unidade i ambas da camada de contexto (conexões recorrentes). Para adequar-se as unidades de contexto para este novo intuito, onde elas devem poder operar com diferentes constantes temporais, a regra de ativação torna-se:

$$c_i(n) = \tau_i c_i(n-1) + (1-\tau_i)\; s\left[\sum_j w_{ij}\, x_j(n) + \sum_j v_{ij}\, c_j(n-1)\right] \qquad (2)$$

onde cada unidade de contexto i tem uma constante temporal τ_i associada, que varia de 0 até 1 determinando a responsividade da unidade, a taxa pela qual a sua atividade é alterada pela camada de entrada (MOZER, 1994, p.248).

> "*With τ_i=0, the activation rule reduces to equation (1) and the unit can sharply change its response based on a new input. With large τ_i, the unit is sluggish, holding on to much of its previous value and thereby averaging the response to the input over time. At the extreme of τ_i=1, the second term drops out and the unit's activity becomes fixed.*" (MOZER, 1994, pp.248-249).

Após esta alteração da função de ativação das unidades de contexto um novo teste foi realizado sobre estruturas musicas contendo frases ordenadas em AABA. CONCERT, com 30 unidades de contexto com $\tau = 0$ e 5 unidades com $\tau = 0.8$, foi submetida ao mesmo exame, mantendo-se o conjunto de treinamento e teste e o número de épocas de treinamento. Quanto aos aspectos locais, que antes atingiram 97.3% de acerto, a taxa para a rede alterada foi de 98.7%; enquanto que para os aspectos globais, que anteriormente tinha atingido 58.4%, os resultados apontaram 75.6% de acerto. Mozer (1994, p. 250) comenta que "*Overall, modest improvements in performance are observed, yet the global structure is never learned as well as the*

local, and it is clear that CONCERT's capabilities are no match for those of people in this simple domain."

Mas o teste mais interessante e que pode realmente avaliar as habilidades de CONCERT, na perspectiva de um teste de Turing musical, envolve o reconhecimento e a criação de peças musicais num determinado estilo. Neste sentido, foram estipulados três novos testes para a rede: composição no estilo de J. S. Bach; composição de melodias folclóricas européias; e composição de valsas. Para todos os testes foram empregadas unidades de contexto com a função de ativação modificada, conforme descrito anteriormente (MOZER, 1994, p.251).

Para o primeiro destes testes foi criado um conjunto de treinamento com 10 peças simples de Bach (como minuetos, marchas, uma *musette* e um pequeno prelúdio). Cada peça do conjunto de treinamento foi transposta para dó maior ou lá menor[8], e teve o seu final demarcado pelo acréscimo de 3 pausas. Buscando uma maior estabilidade da rede, foram adicionadas duas unidades à sua camada de entrada, um apontando quando a peça está em modo maior ou menor, e outra apontando se a métrica é binária ou ternária. A arquitetura de CONCERT foi elaborada com 40 unidades de contexto, sendo 35 delas com $\tau = 0$ e 5 com $\tau = 0.8$. A fase de treinamento consistiu em 3000 épocas e permitiu uma taxa de acerto de previsão de 95% para alturas e durações. Após o treinamento, CONCERT passou a operar no modo de composição sendo apresentado à apenas as 4 primeiras notas de uma das peças do conjunto de treinamento, e em seguida utilizando a seleção de notas e durações do espaço probabilístico da saída da rede para a alimentação da camada de entrada; novamente o fim de uma composição foi determinado pela inclusão de 3 pausas consecutivas.

[8] Esta fato não deixa de ser surpreendente pois no início de seu artigo Mozer (1994, p.237) afirma que "*One desiderable property of the overall PHCCCF representation is that distances between pitchs are invariant over transposition*."

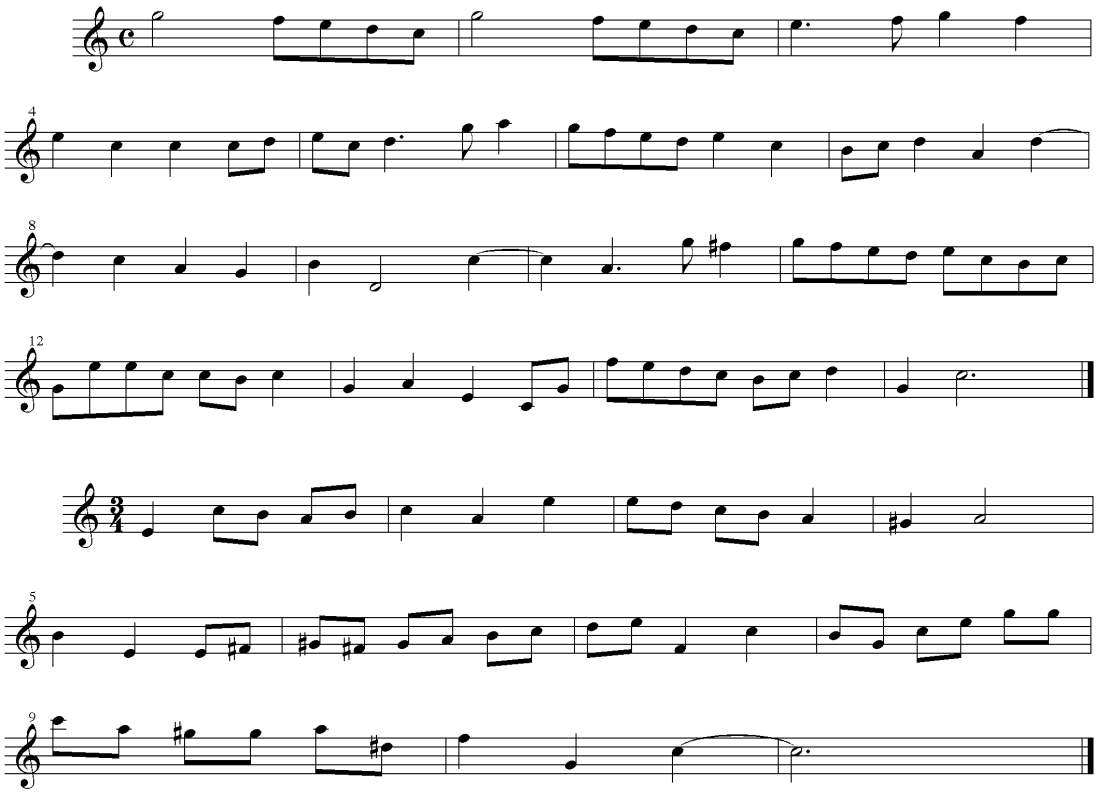

Fig. 4.26. Dois exemplos composicionais de CONCERT baseados no estilo de J. S. Bach. (adaptado de MOZER, 1994, p.252)

Paralelamente, um experimento semelhante foi conduzido com tabelas de transição de 3ª ordem, gerando a mesma quantidade de peças que CONCERT gerou e sobre o mesmo conjunto de treinamento. No intuito de realizar o que chamamos de teste de Turing musical, foi selecionado um grupo de 12 pessoas sem formação musical para avaliar qualitativamente as composições tanto de CONCERT quanto da tabela de transição. Mozer (1994, p.252) comenta:

> "*Two representative examples of each technique were played. Order of presentation was counterbalanced across listeners. All twelve [listeners] chose CONCERT, some with ambivalence, others with a strong preference. Listeners commented that the CONCERT compositions were more coherent and had a more consistent beat. The final cadences of the CONCERT composition were also noted as superior, no doubt because at this point in the piece CONCERT did actually consider more than third-order structure.*"

Um segundo teste utilizou 25 melodias folclóricas européias como conjunto de treinamento. Todas as peças são mais curtas do que as de Bach utilizadas anteriormente, apresentavam compasso quaternário e também foram transpostas para

dó maior. O período de treinamento foi de 2000 épocas, com a rede configurada com 50 unidades de contexto, 45 delas com $\tau = 0$ e 5 com $\tau = 0.8$. As previsões foram corretas em 93% das alturas e 90% das durações. "*Compositions using the trained network sounded reasonable, occasionally having more appeal than the training examples*" (MOZER, 1994, p.253).

O último teste envolveu a utilização de acordes formando um acompanhamento harmônico, dentro do estilo de valsas. Como a taxa de mudança dos acordes normalmente é menor do que a das notas e durações, tal teste pode verificar se a rede é capaz de tratar aspectos de nível mais alto em situações envolvendo músicas reais. "(…) *in order to learn the structure of the chord progression, it would be necessary to span longer periods of time, and, hence, it would be necessary for CONCERT to extract higher-order structure from the pieces*" (MOZER, 1994, p.254). Os resultados obtidos foram semelhantes aos anteriores, quanto às taxas de acerto e quanto aos problemas apresentados. Mozer (1994, p.254) comenta que "*There was little evidence in the compositions that significant global structure was learned*".

Algumas críticas que apontamos quanto aos sistemas de IA aplicadas à geração automática de música, como falta de coerência motívica, se fazem novamente presente, como corrobora Mozer (1994, p.254):

> "*One critic described CONCERT's creations as "compositions that only their mother could love". To summarize more delicately, few listeners would be fooled into the believing that the pieces had been composed by humans. While the local contours made sense, the pieces were not musically coherent, lacking thematic structure and having minimal phrase structure and rhythmic organization. (...) the experiments reported here show no case for optimism in practice, despite the use of state-of-art connectionism architectures and training algorithms, and attempts to encourage CONCERT to learn global structure*."

4.3 Análise de sistemas conexionistas aplicados à percepção e composição musicais.

Nesta última seção deste capítulo iremos discutir, em termos gerais, a aplicação de sistemas conexionistas à música. As descrições apresentadas anteriormente de algumas RNAs nos servirão de apoio para alguns dos pontos que consideramos críticos e merecedores de comentários. Mas antes de prosseguirmos nesta direção, achamos interessante verificarmos até que ponto os problemas que relacionamos à Inteligência Artificial (OLIVEIRA e ZAMPRONHA, 2002) foram superados pela abordagem conexionista.

4.3.1 Sistemas composicionais consolidados e não consolidados.

Verificamos, anteriormente, no segundo capítulo, que os programas da IA normalmente são aplicados à geração de obras musicais dentro de sistemas composicionais consolidados. Cabe perguntar, se no caso das RNAs, esta limitação não se observa. Pelas descrições aqui apresentadas, que vale ressaltar que acreditamos serem significativas dentro dos programas de pesquisa do conexionismo aplicado à música, podemos verificar que certamente não existe um direcionamento voltado a música contemporânea ou não-tonal em geral. Em grande parte das implementações o intuito do pesquisador é a investigação da percepção e geração de obras concebidas dentro do paradigma tradicional, consolidado, da música tonal ocidental. Todas os aspectos que levantamos no final do primeiro capítulo sobre sistemas musicais consolidados e não consolidados continuam válidos dentro do paradigma conexionista. São aspectos que não são resolvidos pela simples mudança de metodologia ou de pressupostos epistemológicos. São aspectos não resolvidos se se muda de uma abordagem *top-down* para uma *bottom-up*. Por que o estudo de tais aspectos envolve fatores além daqueles alcançados nas modelagens computacionais até hoje propostas, sejam pela IA ou pelas RNAs.

Para um estudo das diferenças entre sistemas musicais consolidados e não consolidados, ou da transformação de sistemas não consolidados em consolidados,

caso isso seja possível, precisamos considerar aspectos históricos e sociais. Mas tal consideração é muito distante da atual capacidade das modelagens computacionais, se é que em alguma época será possível. Estes fatores podem ser, e normalmente são, abarcados pela musicologia tradicional, enquanto que aspectos cognitivos voltados à percepção e geração de música são abarcados pela musicologia cognitiva. O que nos leva a defender que uma abordagem multi-disciplinar[9] em musicologia é um caminho promissor para a explicação da atividade musical como um todo, seja sobre sistemas consolidados ou não. O que precisa ficar claro, para o nosso argumento, é que a musicologia tradicional precisa admitir que a musicologia cognitiva é fundamental para um estudo mais completo da atividade musical. Não é suficiente abordar aspectos sociais, antropológicos, históricos e culturais, sem a consideração de estados cognitivos da atividade musical. Da mesma forma, não é pertinente defender o estudo de aspectos cognitivos relacionados à música sem a relação destes com fatores histórico-sociais. Por isso a defesa de uma musicologia multi-disciplinar, na qual uma abordagem pode analisar fatores histórico-sociais pelo viés típico das ciências humanas enquanto a outra pode, pelo método científico tradicional, experimental, explicar os aspectos cognitivos envolvidos em atividades musicais.

Mas, pudemos verificar que algumas das implementações apresentadas neste capítulo relacionam-se a explicação, mesmo que parcial, da consolidação de sistemas musicais, e da influência a consolidação de um sistema tem sobre a atividade cognitiva em música. Especificamente, destacamos o modelo de Leman (1991) e de Katz (1994). Marc Leman (1991) busca explicar o surgimento de funções tonais num sistema auto-organizado, que podemos relacionar à ontogênese de uma semântica tonal em humanos, tendo-se em conta a enorme diferença entre os últimos e o primeiro. Adotando-se o modelo de Leman, podemos estabelecer o processo de internalizarão de funções tonais, gerando a semântica tonal, pela exposição do sistema a um *corpus* de acordes típicos da música ocidental. O mapa topográfico resultante aponta em sua superfície aspectos funcionais da organização tonal que surgem pela reação auto-organizada do sistema aos estímulos. Já o modelo de Katz (1994), em um dos experimentos realizados, investiga o efeito da familiaridade com um sistema musical sobre a apresentação de músicas dentro e fora de tal sistema. Em geral, existe um aumento de ativação quando a RNA é exposta a um estilo de música com o qual já

[9] Estamos utilizando o termo multi-disciplinar, em detrimento de inter-disciplinar, pois acreditamos que as áreas envolvidas não possuem o mesmo método de investigação.

foi treinada. Da mesma forma, quando exposta à um estilo não familiar o aumento de ativação, se existente, é significativamente inferior. Este experimento pode explicar a preferência de ouvintes para obras concebidas dentro de um sistema com o qual já esteja familiarizado, entendendo-se a rede de Katz (1994) como um modelo (abstrato) explicativo e não uma simulação de um ouvinte.

Contudo, cabe ainda ressaltar explicitamente um dos possíveis motivos pela preferência pelos sistemas musicais consolidados nas implementações conexionistas. Sendo um sistema consolidado, existe um conjunto de regras codificadas e bem definidas aplicadas a materiais também bem definidos. Em outras palavras, a música de um sistema consolidado é explicada e realizada por regras definidas *a priore*. Sendo, portanto, o sistema musical formalmente claro e preciso, torna-se mais fácil a avaliação do comportamento e dos resultados apresentados por uma RNA aplicada à música. Se uma rede for concebida para operar sobre um sistema não consolidado a análise de seus *outputs* será bastante imprecisa pela falta de um conjunto claro de regras e elementos do sistema musical que comprovem a eficiência da RNA, e, conseqüentemente, sua validade como modelo explicativo será descartada. No máximo ela poderá ter um interesse enquanto um sistema gerativo autômato de músicas concebidas dentro de um sistema musical não consolidado, mas sem um comprometimento direto com a explicação da atividade composicional sobre tais sistemas.

4.3.2 Regras imperativas e restritivas.

Outro problema que devemos analisar quanto a sua superação pelas RNAs são as regras imperativas e restritivas utilizadas para a geração de obras musicais. Quando se tratando de modelagens e simulações com máquinas de Turing, verificamos que, diferentemente dos procedimentos tradicionais de composição e análise músicas, existe larga utilização de regras imperativas na elaboração de procedimentos algoritmos para a composição automática de música. No entanto, nas RNAs verificamos, em geral, a utilização de regras restritivas. Existe a decorrência deste tipo de regra pela fase de aprendizagem ou treinamento da rede, que consiste em expô-la a um conjunto de vetores de entrada que representam obras musicais ou aspectos de

obras musicais. Por este procedimento, a rede gera um mapa probabilístico que reflete as regras musicais presentes nas obras do conjunto de treinamento, mas, em geral, sem uma determinação nomológica entre entrada e saída. Em vez de termos um conjunto de regras imperativas estabelecidas por um programador ad hoc, temos a extração de um conjunto de regras restritivas estabelecidas na configuração de pesos da RNA, estabelecida pela exposição a um conjunto de treinamento. Como Mozer (1994) aponta, tais RNAs são semelhantes à cadeias de Markov, ou tabelas de transição mas apresentam características que vão além destas tabelas implementadas em algoritmos da IA. As propriedades típicas dos sistemas conexionistas permitem um maior e mais flexível poder computacional sobre tabelas de transição no caso das RNAs. Estendendo a típica aplicação de RNAs a sistemas musicais consolidados, talvez seja interessante verificar se RNAs treinadas sobre sistemas não consolidados podem estabelecer estatisticamente um conjunto de regras condizentes com músicas desses sistemas. Tendo-se em conta o problema aponta logo acima sobre o alcance explicativo sobre sistemas não consolidados. De qualquer forma, este parece ser um ponto ainda por ser investigado por implementações conexionistas na área da música. Mas voltando a questão do início do parágrafo, podemos afirmar que o problema da utilização quase que exclusiva de regras imperativas na IA, foi superado na abordagem conexionista.

4.3.3 Criação de regras.

A afirmação acima nos leva a outro problema comparativo entre a IA e as RNAs: a criação de regras. Recordamos que um dos pontos elencados no caso de sistemas de IA aplicados à composição foi o fato de que em geral o sistema não tem a capacidade de gerar suas próprias regras composicionais, o que pelo menos intuitivamente podemos dizer que um ser humano pode fazer quando compondo música. Na verdade, nem o compositor humano gera suas regras quando envolvido com composição de música sobre sistemas consolidados, o que leva a afirmação mais radical de que o sistema compõe pelo compositor, o sistema musical restringe as ações do compositor (ZAMPRONHA, 2000). Neste caso não existe muita diferença entre um compositor humano e um mecânico. Ambos seguem regras ad hoc e *a*

priore. Muitas vezes, como afirmamos no caso de Cage no capítulo segundo, compor está mais relacionado à criação das próprias regras do que a utilização de regras predefinidas, portanto temos um sistema não consolidado. Nesse sentido, cada obra pode apresentar seu próprio conjunto de elementos musicais e suas próprias regras de relacionar organizacionalmente tais elementos. Nas palavras mais poéticas de Rosenboom (1997b, p.36):

> "*A composer's license includes the opportunity to construct entire universes. It may be useful to consider some fundamental steps in constructing a compositional method. These may be unique for each individual and may apply to single works or bodies of work.*"

Os passos destacados acima por Rosenboom são: a escolha do universo próprio da obra, determinar como tal universo é ordenado, determinar as escalas de medida e comparação entre os elementos de tal universo, estabelecer os níveis de significação de cada elemento e conjuntos de elementos, e estabelecer uma pragmática composicional de acordo com certas premissas filosóficas escolhidas pelo compositor (ROSEMBOOM, 1997b, p.37). De fato, Rosenboom (1997a e 1997b) aprofunda nossa idéia de que compor pode ser mais do que seguir regras *a priore*; pode ser a criação das próprias regras, e indo ainda mais além, a criação de universos inteiros.

Sendo assim, conseguem as RNAs sustentar este tipo de postura enquanto sistemas compositores autômatos? No caso da IA, já vimos que não é possível uma resposta positiva para essa pergunta. No caso das RNAs, a resposta também será negativa na nossa opinião. Como já pudemos observar, as RNAs utilizam regras que são extraídas pela apresentação do conjunto de treinamento, de forma que sua atuação gerativa será baseada em tais regras, e poderá até extrapolar tais regras, mas não utilizar regras diferentes. O problema de geração de regras composicionais não é superado por redes neurais. A criação de regras composicionais, ou mesmo de universos musicais parafraseando Rosenboom (1997a e 1997b), parece envolver processos cognitivos altamente complexos que as atuais modelagens conexionistas não conseguem atingir. A alta complexidade envolve, além de parâmetros musicais que podem ser formalizáveis, aspectos históricos, sociais, ambientais (perceptuais) e estéticos, para citarmos apenas alguns, que muito dificilmente podem ser inseridos

num sistema conexionista devido, inclusive, aos limites representacionais sobre os quais operam as RNAs.

4.3.4 Relações arbitrárias.

Um outro problema apontado no segundo capítulo, especificamente o das relações arbitrárias entre parâmetros do algoritmo e externos, pode ser parcialmente superado em RNAs compositoras. No caso da IA, vimos (na seção 1.5.4) que o programador pode estabelecer uma relação arbitrária entre variáveis do programa e valores gerados externamente para quebrar a homogeneidade da geração, e talvez garantir um direcionamento formal da obra gerada pelo sistema. No caso da arbitrariedade em RNAs, podemos analisar um outro aspecto. Pode existir uma relação arbitrária entre a entrada e a saída de uma rede se ela for supervisionada. O programador determinará a associação entre entradas e saídas, o que por fim resulta na determinação do tipo de classificação e geração de valores por uma RNA. Contudo, nas RNAs não supervisionadas tal determinação arbitrária e externa não existe. O algoritmo, como o de Kohonen, por exemplo, possui regras predeterminadas sobre como se comportar para a geração de uma saída da rede para cada estímulo, mas a determinação final dependerá do estado atual da rede, dos estímulos já apresentados e do estímulo atual a ser classificado. O que em parte supera a programação totalmente predeterminada do caso da IA, ou mesmo da determinação entre entrada e saída nas RNAs supervisionadas. No caso de empregar-se valores gerados externamente numa composição autômata, uma rede não supervisionada pode, por si só, estabelecer uma relação entre estes valores e os parâmetros composicionais que emprega para a geração de uma obra; mas, de qualquer forma, nada garante o interesse de obra, da mesma forma que nada garante, em qualquer tipo associação de valores composicionais à valores não-musicais, um resultado de interesse musical e/ou estético..

4.3.5 Limitações representacionais do fenômeno acústico bruto.

Podemos agora, ainda discutir alguns pontos críticos que não foram postulados para a IA, mas que se colocam sobre a perspectiva conexionista. A primeira delas, e talvez uma das mais sérias críticas, relaciona-se ao tipo de representação dos dados sobre os quais um RNA opera. A dimensão deste problema representacional fica evidente na discussão apresentada por Bharucha (1991), que descrevemos anteriormente (seção 2.4.2). Em termo gerais, achamos pertinente discutirmos aqui a adequação dos tipos de representação normalmente utilizadas em RNAs perante o fenômeno sonoro acústico. Como pudemos perceber pelas descrições apresentadas, normalmente as entradas da rede são elaboradas sobre categorias abstratas, como *pitch-class*, estabelecidas sobre pressupostos psicológicos, como que numa perspectiva *top-down*. Utilizamos o termo *top-down* porque se parte de um pressuposto geral sobre o funcionamento psicológico da percepção auditiva que é implementado num sistema computacional; a forma da representação não emerge da interação dos componentes do sistema perante um determinado estímulo do ambiente, mas é determinada de antemão sobre pressupostos psicológicos. Em alguns casos, como na RNA de Sano e Jenkins (1991) existe uma maior pertinência em termos de características anatômicas e fisiológicas do sistema auditivo onde parte-se do sinal acústico e chega-se a uma representação em *pitch-class*. Em outros casos, abandona-se uma representação mais próxima do sinal acústico e emprega-se diretamente representações em *pitch-class*. O modelo de Leman (1991) é um exemplo disto. Porém, até que ponto o comportamento temporal dinâmico do espectro sonoro é importante para a percepção auditiva? Se pensarmos que sua importância é mínima, podemos dizer que não existem grandes problemas com as formas abstratas de representação de altura. Por outro lado, se considerarmos que a importância do comportamento temporal do som é capital para a percepção auditiva, e que além de categorias de altura, o timbre deve ser considerado nas modelagens da percepção, verificamos que existem problemas significativos com relação às formas representacionais empregadas por RNAs para a representação do sinal acústico.

A falta de consideração dos aspectos dinâmicos do fenômeno sonoro fica evidente pela ausência de modelos conexionistas relacionados à percepção de

timbre[10]. Fica também evidente por afirmações dos próprios pesquisadores que corroboram nossa opinião. Leman, por exemplo, afirma (1991, p.106-108):

> "*Our ultimate goal is to start from raw acoustic data (music "as it sounds", that is, as it appears as vibration in the air), but at this stage of our research we adopted a more modest approach over which strict control could be more easily exercised. (...) We are all aware that [the representational] restrictions impose severe limits on the scope of our model, but we believe that they can be overcome in the very near future.*" (aspas do autor)

Porém o futuro muito próximo que trará a superação destas limitações ainda não chegou. A consideração destes aspectos dinâmicos do som ainda espera por um maior poder computacional suficiente para lidar com a alta dimensionalidade do fenômeno acústico bruto. Fica ainda a pergunta: até que ponto podemos postular modelos explicativos que se apóiam em pressupostos psicológicos definidos *a priore* e implementados em redes de capacidade modesta que descartam o fenômeno acústico bruto e dinâmico? Fica, da mesma forma, a resposta típica de que em tempos vindouros poderemos considerar o fenômeno acústico bruto com sua alta dimensionalidade com um maior poder computacional, ou mesmo com novos paradigmas computacionais. Talvez aí teremos sistemas realmente *bottom-up*, e não sistemas mistos como os atuais, com comportamentos explicados numa perspectiva *bottom-up*, mas apoiados em pressupostos representacionais de dados à maneira *top-down*.

Esta questão da perspectiva *bottom-up* contaminada pela perspectiva *top-down* fica evidente na modelagem de Katz (1994). Especificamente, Katz (1994) afirma que sua investigação por um modelo conexionista é sobre o afeto musical. Basicamente, ele analisa a performance de sua RNA buscando encontrar indícios do afeto musical pela ativação das camadas superiores da rede. Mas é controverso afirmar que esta relação é válida. Parece haver poucos indícios que garantam esta relação, se é que algo de fato garanta, entre o comportamento da rede medido em termos da ativação média das camadas e o afeto musical, que possui um caráter principalmente estético. Katz parte de um pressuposto tipicamente top-down entendendo que existem princípios psicológicos de alta-ordem que instanciam o afeto musical, buscando um

[10] Pelo menos na bibliografia por nós consultada, que acreditamos ser abrangente o suficiente para sustentar tal afirmação.

modelo bottom-up que o explique. Parece existir, então, uma questão de interpretação sobre a análise do comportamento da rede em termos de associa-lo ao afeto musical. Mas, trata-se de uma interpretação realizada antes da própria simulação; justamente pelo contrário, a simulação foi concebida para corroborar a idéia já estabelecida da existência de afeto musical. A análise dos resultados já estava contaminada por pressupostos psicológicos estabelecidos. Tal fato aparece mais ou menos evidentemente em outras modelagens conexionistas, mas na proposta de Katz (1994) ele é bastante claro. Acreditamos que numa perspectiva realmente *bottom-up* deva-se explicar os fenômenos perceptuais ou composicionais apresentados por uma RNA em termos dos elementos de baixo nível que compõem o sistema e que, por sua vez, instanciam um comportamento global, como no modelo de Large e Kolen (1994). Para postular-se a existência de um modelo explicativo válido, deve-se buscar por uma relação mais objetiva e nomologicamente determinada entre níveis locais e globais para, dessa forma, haver uma garantia, mesmo que mínima, entre aspectos anatômico-fisiológicos e psicológicos envolvidos com a atividade musical.

4.3.6 Considerações finais deste capítulo.

Neste segundo capítulo descrevemos, em primeiro lugar duas redes que modelam a percepção rítmica; uma delas envolvida com a questão clássica de quantização do tempo musical, e outra que apresenta um modelo de ressonância para a percepção métrica. Em seguida, passamos a investigar as propostas conexionistas dirigidas à percepção de altura e de tonalidade. Em primeiro lugar tratamos dos aspectos relacionados à forma de representar os dados musicais e sonoros, para prosseguirmos descrevendo três implementações. A primeira RNA procura estabelecer um modelo para a percepção de altura; uma outra a percepção melódica, que envolve altura e duração; e a terceira RNA investiga a ontogênese de funções tonais. Após todas estas redes que buscam modelar aspectos de caráter mais perceptual, analisamos uma proposta conexionista voltada à composição musical. Por fim, destacamos alguns pontos críticos com relação à aplicação de RNAs à musica. Pudemos verificar que muitos dos aspectos problemáticos apontados no segundo capítulo, no caso da IA, continuam presentes e válidos no caso das RNAs. Ainda,

discutimos alguns pontos relacionados à forma de representar os dados musicais em sistemas conexionistas e a distância existente entre o fenômeno acústico bruto e a representação deste em RNAs.

Podemos afirmar, que ainda relacionado ao problema da representação do fenômeno acústico bruto, um aspecto de vital importância a ser considerado pelas RNAs é a proposta de uma abordagem conexionista ecologicamente orientada. O próprio Leman (1991) aponta nesta direção sua proposta conexionista. Por proposta conexionista ecológica estamos entendendo um sistema artificial que não só seja capaz de lidar com os estímulos de alta dimensão provenientes do ambiente (acústico), mas que possa guiar a sua ação, seu comportamento pela interação com tais estímulos. Trataremos em mais detalhes a perspectiva ecológica da percepção auditiva no próximo capítulo. De forma semelhante, acreditamos que é preciso estabelecer modelos conexionistas capazes de lidar de aspectos dinâmicos relacionados à percepção do timbre. Para tanto é necessária uma forma de representar o fenômeno acústico o mais próximo possível de seu estado bruto, ou da forma como ele se apresenta para nós, parafraseando Leman (1991). Com o objetivo de abordar tanto aspectos ecológicos, sobre a interação entre sistema e meio-ambiente, quantos aspectos dinâmicos dos estímulos as redes booleanas randômicas (RBN, ou *Random Boolean Networks*) podem mostrar-se como uma interessante ferramenta visto responderem comportamentalmente a estímulos provenientes do ambiente, e, ainda, apresentarem comportamentos diferentes (indo do caos à ordem) para cada tipo de estímulo, de acordo com seu grau de caoticidade[11]. Hipotetisamos, em Lima et al. (2003), que tais redes são adequadas tanto para a modelagem da percepção timbrística quanto para a geração de novos timbres:

> "Zampronha (2001b) observa que uma das possibilidades em seu escalonamento de periodicidade é a de que a percepção auditiva opere sobre uma análise comportamental do fenômeno acústico e não sobre análises como a transformada de Fourier, muito aplicada à análise computacional de objetos sonoros. Assim, a arquitetura RBN surge como uma robusta ferramenta para testar tal hipótese, sem falar da possibilidade de podermos vislumbrar a aplicação destas redes neurais para a geração de novos timbres para a aplicação em música contemporânea, especialmente a eletroacústica (Zampronha, 2001a)."

[11] Em Lima, Oliveira e Broli (2003) apresentamos uma proposta ainda bastante inicial sobre redes randômicas aplicadas à música.

Porém, além das RBNs, ainda são dignas de menção propostas computacionais também bastante atuais que têm sido relacionadas à composição musical. Especificamente, estamos falando de autômatos celulares. Autômatos celulares são unidades computacionais interativas que simulam comportamentalmente um indivíduo (ou agente) de uma população, ou uma célula de um tecido, ou outros fenômenos biológicos e sociais. Usualmente são empregados em área como *A-Life* (*Artificial Life*, ou vida artificial) para estudar, por meio de simulação, fenômenos biológicos de alta complexidade. Recentemente, tem-se observado aplicações de algoritmos genéticos, nome que designa algoritmos construídos sobre autômatos celulares, à áreas fora do escopo das ciências biológicas. No caso da relação entre *A-Life* e composição musical, Miranda e Todd (2003) vêem três possibilidades: execução de comportamento extra-musical; abordagem inspirada em algoritmo genético; e abordagem cultural.

A primeira possibilidade ocorre pela associação entre o comportamento da simulação e parâmetros sonoros ou musicais.

> "*These agents are not musical in the sense that they are not designed with any musical task in mind. Rather, some sort of "sonification" or "musification" to their behavior patterns is applied in order to see (or hear) what emerges. (...) Because cellular automata are commonly used to study the creation of complexity and dynamic patterns, their behavior can produce interesting musical patterns as well when sonified.*" (MIRANDA e TODD, 2003, p.60) (aspas dos autores)

Este tipo de abordagem pode ter um interesse musical dependendo da associação estabelecida no processo de sonorização ou musicalização dos padrões dinâmicos e complexos emergentes da simulação. Obviamente, caímos aqui no mesmo ponto que apontamos anteriormente (seções 1.5.4 e 2.5.4) com respeito às relações arbitrárias.

A segunda abordagem, inspirada em algoritmos genéticos, possui uma relação mais direta com a produção musical, no sentido em que cada agente da simulação produz sua própria música. Existem fatores determinantes na vida de cada agente, como sua capacidade reprodutiva, suas restrições corporais e comportamentais etc., que são ajustadas de acordo com a adequação da música por ele gerada. A adequação musical pode ser avaliada por um agente externo humano ou artificial, como uma RNA, sobre a musica gerada pela simulação como um todo, ou a produção musical

individual de cada agente da simulação (MIRANDA e TODD, 2003, p.61). Existe a necessidade deste avaliador externo porque os agentes da simulação não ouvem uns aos outros, ao invés disso, a música que eles geram é determinada pela sua constituição genética. O avaliador determina, pelo resultado sonoro, quais agentes sobreviverão e reproduzirão e quais não o farão.

> "*When human critics are used, these evolutionary systems can produce pleasing and sometimes surprising music, but usually after many tiresome generations of feedback. Fixed artificial critics take the human out of the loop, but have had little musical success so far. What would happen if we unfix the critics and/or replace the human critic by other agents in the artificial world? This is one of the central ideas of the cultural approach, where individuals become both producers and receivers of music.*" (MIRANDA e TODD, 2003, p.61)

A afirmação acima já esboça a idéia da terceira abordagem, a cultural. Neste caso, existem dois tipos de agentes em interação social numa simulação: agentes produtores e agentes críticos. Ou, pode haver agentes que hora são produtores e hora críticos. Todd e Werner (1999) propõem uma simulação onde agentes do gênero masculino produzem melodias e agentes do gênero feminino avaliam tais melodias para determinar qual agente será seu parceiro no processo de reprodução. A avaliação das melodias é feita por tabelas de transição, de forma que as melodias mais surpreendentes[12] são o fator determinante na escola do parceiro. A evolução do sistema de agentes apresenta uma manutenção do repertório assim como uma contínua transformação deste.

Em Miranda (2002) encontramos uma simulação semelhante, chamada de modelo mimético. Nesta simulação, cada agente é produtor (por síntese sonora) e imitador (por análise sonora) de expressões vocais. Cada um deles apresenta parâmetros relacionados à suas habilidades motoras, perceptuais e cognitivas, além de um instinto básico de imitar sons de outros agentes. Após um certo número de gerações a sociedade simulada apresenta um número de expressões sonoras compartilhadas por seus agentes, mas ainda com a possibilidade de surgimento ocasional de novas expressões.

[12] Melodias consideradas surpreendentes são aquelas que geram expectativas (tabelas de probabilidade não uniformemente distribuída) e quebram tais expectativas (sucessão de estados com pouca probabilidade). Dessa forma, melodias randômicas não são consideradas surpreendentes pois o grau de expectativa é muito baixo (a distribuição propabilistica é unifirme) (TODD e WERNER, 1999).

Estes exemplos, ainda que brevemente descritos, esclarecem o tipo de proposta atualmente em voga na área da computação musical. Um dos aspectos mais interessantes é possibilidade de geração de complexidade e desenvolvimento dinâmico pela simulação de interações sociais entre agentes. A alcance enquanto modelo explicativo de aspectos socio-culturais em música, assim como da onto e filogênese de sistemas musicais (incluindo aspectos sintáticos e semânticos), ainda precisa ser efetivamente comprovada, mas o caminho nos parece, atualmente, promissor. As simulações da A-Life aplicadas à música introduzem um maior grau de complexidade e dinamismo nos processos autômatos de composição. A A-life abre, assim, um novo paradigma da composição musical por computador, seja o sistema supervisionado ou não por um compositor humano, onde fatores ambientais, sociais e estéticos podem ser considerados metodologicamente de maneira mais adequada. Tal maior adequação metodológica é importante no sentido que, acreditamos, os fatores acima citados exercem um papel decisivo e indispensável na atividade musical, e na sua explicação.