Máquinas Capazes de Conhecer Automaticamente o Mundo

Tiago Fernandes Tavares[1] **, Jayme Garcia Arnal Barbedo**[2] **, Romis Attux**[3] **, Amauri Lopes**[4]

Departamento de Engenharia de Computação e Automação Industrial (DCA)[1,2,3]
Departamento de Telecomunicações (DECOM) [4]
Faculdade de Engenharia Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Caixa Postal 6101, 13083-970 – Campinas, SP, Brasil

{tavares[1], attux[3]}@dca.fee.unicamp.br, {jgab[2],amauri[4]}@decom.fee.unicamp.br

Abstract – Este trabalho apresenta reflexões sobre o problema de analisar sinais do mundo real e inferir eventos relacionados a eles. Embora possam haver diversas soluções técnicas para esse problema, todas elas, necessariamente, devem abordar questões semelhantes. Tais questões são discutidas neste trabalho.

Keywords – Cognition, Artificial Intelligence, Pattern Recognition, Signal Processing

1. Introdução

Sinais, na natureza, são perturbações de variáveis físicas, sendo geralmente causadas por eventos. Essa relação de causalidade é capaz de prover informações valiosas para um ser humano (quando, por exemplo, ao ouvir um determinado ruído, torna-se possível saber que um carro está se aproximando, ou ainda, ao sentir um certo cheiro, sabe que houve um vazamento de gás). O processo de encontrar, sem interferência humana, eventos que causaram um determinado sinal tem diversas aplicações, como inferir a passagem de uma pessoa por uma porta através da análise dos dados vindos de um sensor laser.

A transcrição automática de música se insere nesse contexto, tratando-se de um problema no qual um sinal de áudio é analisado e os eventos que o causaram – notas musicais – são encontrados. Nesse sentido, o detector de pessoas citado acima poderia ser entendido como um transcritor automático dos dados do sensor laser, buscando encontrar a passagem de pessoas.

Embora aplicações mais simples possam ser voltadas simplesmente à detecção da existência de um certo evento (se o sensor laser está sendo utilizado para segurança, basta saber que uma pessoa passou pelo ponto desejado para que seja preciso soar o alarme), mas pouco aumento na complexidade pode gerar situações mais sensíveis. O sensor laser, se for utilizado para contar o número de pessoas que passou pela porta, deve ser acoplado a um sistema capaz de reconhecer, por exemplo, que existe um intervalo de tempo mínimo que deve existir entre a passagem de duas pessoas, mesmo que um eventual sinal ruidoso indique o contrário.

Assim, pode-se dizer que a tarefa de transcrição automática requer não somente um sensor adequado, mas também um sistema de pós-processamento que busca transformar os dados brutos em conhecimento simbólico. O sistema de pós-processamento é, portanto, um sistema cognitivo.

Neste trabalho, são discutidas questões inerentes ao projeto de qualquer sistema de transcrição, com ênfase no problema da transcrição automática de música. Inicialmente, as funcionalidades necessárias a se realizar um processo de transcrição são discutidas. Após, é realizada uma breve apresentação dos conceitos sobre os quais se baseiam algoritmos cognitivos. Por fim, são apresentadas prospecções para futuros desenvolvimentos na área da transcrição automática de música.

2. Transcritores automáticos

Neste trabalho, são chamados de transcritores automáticos todos os sistemas que utilizam dados brutos do mundo real para inferir informações simbólicas relacionadas a eventos.

Em certos casos, os dados providos pelo sensor analisado são correlacionados ao evento a ser detectado – no caso do sensor laser, trata-se de uma tensão ou corrente que pode ser medida na saída do circuito. Em outros casos, é necessário desenhar algoritmos eficazes para prover, a partir dos dados brutos, um sinal mais correlacionado com o evento a ser detectado. A saída desejada, de forma geral, é um sinal cuja intensidade se modifica quando ocorre um evento e assume níveis estáveis quando o evento não ocorre.

A Figura 1 mostra um caso comumente encontrado na transcrição automática de música. Sobre uma sequência de notas musicais, é executado um certo algoritmo que retorna um sinal chamado de percussividade, que indica o quão forte é a hipótese de encontrar uma nova nota musical naquele

instante de tempo. A natureza desse algoritmo específico não é importante para as discussões que se seguem.

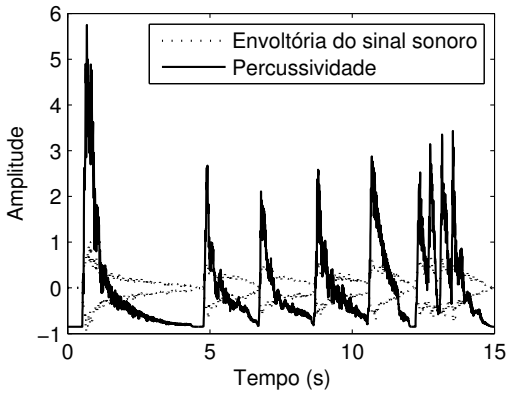

Figura 1. Percussividade em uma sequência de notas musicais.

No sinal de percussividade mostrado, há aspectos que assemelham o processo de detecção de uma nota musical a outros processos de detecção envolvendo sensores com ruído. O sinal de percussividade que, idealmente, deveria apresentar um máximo local para cada nova nota, apresenta, na verdade, diversos máximos locais espúrios. Trata-se de uma consequência dos limites da capacidade do algoritmo de aproximar o comportamento do fenômeno real.

Também no mesmo sinal, é possível verificar que a maior parte dos máximos locais espúrios podem ser descartadas se for levado em consideração o fato de que são precedidos por um máximo local de maior intensidade. É importante perceber que a informação sobre a necessidade de se descartar máximos locais que seguem uma determinada regra não veio da análise do sinal de percussividade, mas de uma determinada expectativa prévia sobre o comportamento dos eventos que geraram esse sinal.

Assim, o processo de cognição se iniciou com a captação do som de algumas notas musicais, prosseguiu com o cálculo de uma função que se aproxima parcialmente da característica da percussividade (e, portanto, da força da hipótese de se encontrar uma nova nota se iniciando naquele instante) e então passou para uma etapa de interpretação que toma por base algum modelo previamente conhecido. A Figura 2 mostra um diagrama de blocos no qual estão presentes todas essas etapas desse processo cognitivo.

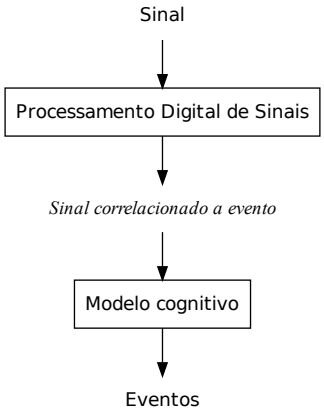

Figura 2. Diagrama de blocos de um sistema de transcrição.

A presença de duas fontes de informação gera um problema adicional, que envolve decidir quando se deve confiar no sinal observado e quando se deve confiar na expectativa prévia quanto ao comportamento dos eventos.

A seguir, serão discutidos modelos cognitivos utilizados no problema de transcrição de música.

3. Sistemas cognitivos

Os sistemas mais recentes para transcrição automática de música modelam a música como uma cadeia probabilística na qual cada nota depende somente da anterior [3, 11, 12, 13], na forma de uma cadeia de Markov [9].

Uma cadeia de Markov, como mostrada na Figura 3, é definida como um conjunto de estados, sendo que para cada iteração o sistema passa a assumir que está num determinado estado, escolhido por uma probabilidade condicional que depende do estado assumido anteriormente.

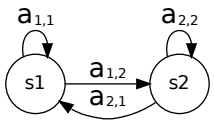

Figura 3. Uma cadeia de Markov. Na figura, $a_{i,j}$ é a probabilidade de se passar para o estado s_j dado que se estava no estado s_i.

Há dois tipos de entidades envolvidos no uso de cadeias de Markov: os estados s_i, que se relacionam aos eventos que se busca conhecer, e o sinal de entrada, $x[n]$, que é aquele obtido da captação de sensores ou o resultado da aplicação de algoritmos de

processamento digital de sinais. Cada estado da cadeia de Markov se relaciona com um certo comportamento de um sinal de forma probabilística. Assim, temos, para cada estado, uma probabilidade conhecida para valores assumidos por $x[n]$. Através dessa relação, é possível tomar uma série de observações – o sinal de entrada $x[n]$ – e inferir qual é a sequência de estados – e, portanto, de eventos – que mais provavelmente ocorreu.

Cadeias de Markov, no problema da transcrição, trazem benefícios significativos. Trata-se de um algoritmo cujos parâmetros são obtidos por aprendizagem supervisionada, e, portanto, não é necessário conhecê-los previamente. Nesse processo de aprendizagem, a medida do quão confiável é a observação e quão confiável é a expectativa prévia é ajustada automaticamente. Além disso, é importante perceber como a cadeia de estados modela a forma pela qual o sinal é *gerado*, mas é utilizada de forma a permitir que os estados – inicialmente desconhecidos – sejam *inferidos*.

Os sistemas mais recentes de transcrição automática utilizando cadeias de Markov [11, 12, 13] apresentam índice de acertos por volta de 60%.

Processos de Markov, porém, ignoram aspectos específicos do sinal a ser transcrito. No caso da música, é marcante que todos os aspectos rítmicos sejam ignorados pela cadeia de Markov.

Em peças musicais, em especial na música popular ocidental contemporânea, são marcantes os aspectos rítmicos. Através do posicionamento eficaz de estruturas – notas, timbres, etc. – em uma peça, um compositor é capaz de criar diferentes formas de estética musical [14]. Análises sobre o processo de composição mostram que, mesmo que a música seja criada através de processos individuais a cada compositor [2], há determinadas formas de organização musical significativamente mais comuns que outros [2, 5]. Experimentos psicoacústicos mostram que é marcante e mensurável o fato de que seres humanos, ao ouvir músicas, criam expectativas sobre eventos musicais futuros [4].

O conhecimento sobre o funcionamento das estruturas musicais só veio a ser utilizado explicitamente na transcrição de música em 2009 por Mauch [6], que fez uso da repetição de seções da música (verso, refrão, etc.) para melhorar o desempenho de seu transcritor de acordes. O trabalho de Mauch [6] mostra que o uso de conhecimento prévio quanto às regras de estruturação da música tende a melhorar o desempenho de sistemas cognitivos ligados à transcrição.

Incorporar maiores informações sobre o funcionamento da música implica, também, num aumento da relevância dada ao sistema para sua expectativa. Pode-se esperar que o desempenho do transcritor seja melhorado tanto quanto for melhor a correlação entre o sistema cognitivo utilizado e o fenômeno observado.

É importante perceber que assim como uma cadeia de Markov aparece como um processo de geração de sinais mas é utilizada como um processo de cognição de sinais, nada impede que qualquer algoritmo que gera sinais seja utilizado de forma semelhante para analisa-los.

Isso significa que, ao se desenvolver sistemas capazes de gerar sinais que se assemelham a um determinado fenômeno tomando por base relações entre símbolos, simultaneamente se está desenvolvendo uma parcela significativa de um sistema cognitivo que interpretará sinais dessa mesma natureza e os converterá em símbolos. No contexto da transcrição automática de música, isso significa que é importante estudar sistemas que geram sequências de símbolos musicais, em especial aqueles que geram continuações para sequências musicais incompletas.

Alguns sistemas de geração de sequências musicais são especialmente relevantes. Uma modificação de um sistema de predição universal de texto foi proposta por Assayag *et al.* [1] para prever sequência musicais baseando-se numa representação textual de partituras. Um modelo mais complexo, proposto por Maxwell [7], explicitamente modela estruturas hierárquicas de uma música buscando prever continuações com consistência de estilo. Em um trabalho de doutorado, Paiement [8] observa que algoritmos construídos especialmente para uso na predição de séries de símbolos musicais tendem a apresentar melhor desempenho que algoritmos genéricos como redes neurais [10] ou modelos de Markov [2].

A conversão de sistemas de geração de símbolos para sistemas de inferência depende da aplicação da mesma ideia utilizada no caso das cadeias de Markov [9] – essencialmente, a aplicação do Teorema de Bayes para o sistema de geração construído. Conceitualmente, alguns problemas serão certamente abordados.

Inicialmente, a semelhança entre os símbolos gerados pelo sistema de geração de símbolos e o fenômeno real deve ser avaliada – espera-se que sistemas que geram símbolos de forma mais próxima ao fenômeno real sejam capazes de prover uma expectativa mais realista sobre o prosseguimento de cadeias. Além disso, devem ser avaliados quais os sensores e algoritmos de processamento digital de sinais capazes de obter informações mais relevantes ao problema avaliado. Por fim, a união dos dois sistemas deve necessariamente (mesmo que implicitamente) passar por uma etapa em que se determina uma medida do quão se deve confiar em cada uma das fontes de informação – as observações e as expectativas. Um balanço adequado de ambas as informações permite ao sistema realizar inferências de melhor qualidade.

4. Conclusão

Este texto apresentou conceitos que devem estar presentes em todo sistema cognitivo que busque encontrar símbolos relacionados a sinais reais. Embora tenha como foco principal a transcrição automática de música, os conceitos gerais abordados são imediatamente aplicáveis a diversos campos de atuação.

Referências

[1] Gérard Assayag, Shlomo Dubnov, and Olivier Delerue. Guessing the composer's mind: Applying universal prediction to musical style. In *Proceedings ICMC 99*, Beijing, China, 1999.

[2] David Cope. *Techniques of the Contemporary Composer*. Schirmer Thomson Learning, 1997.

[3] Goto, M. A robust predominant-f0 estimation method for real-time detection of melody and bass lines in cd recordings. In *Proc. 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II757–II760, Istanbul, Turkey, June 2000.

[4] David Huron. *Sweet Expectation: Music and the Psychology of Expectation*. MIT Press, 2006.

[5] N. C. Maddage. Automatic structure detection for popular music. *IEEE Multimedia*, 13(1):65–77, January 2006.

[6] Mauch, M., Noland, K., and Dixon, S. Using musical structure to enhance automatic chord transcription. In *Proc. 10th International Conference on Music Information Retrieval*, pages 231–236, Kobe, Japan, October 2009.

[7] James B. Maxwell, Philippe Pasquier, and Arne Eigenfeldt. Hierarchical sequential memory for music: A cognitively-inspired approach to generative music. 2010.

[8] Jean-François Paiement. Probabilistic models for music. Master's thesis, École Polytechnique Fédérale De Lausanne, 2008.

[9] Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. of the IEEE*, 77(2):257–286, February 1989.

[10] Axel Robel. Neural networks for modeling time series of musical instruments. In *In Proc. of the International Computer Music Conference, ICMC*, pages 424–428, 1995.

[11] Ryynänen, M. and Klapuri, A. Transcription of the singing melody in polyphonic music. In *Proc. 7th International Conference on Music Information Retrieval*, pages 222–227, Victoria, BC, Canada, October 2006.

[12] Ryynänen, M. and Klapuri, A. Automatic bass line transcription from streaming polyphonic audio. In *Proc. 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1437–1440, Honolulu, Hawai'i, USA, April 2007.

[13] Ryynänen, M. and Klapuri, A. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008.

[14] Schoenberg, A. *Fundamentals of Musical Composition*. Faber and Faber Limited, 1 edition, 1970.