Detecção automática do início de notas usando o aprendizado de máquina

MODALIDADE: COMUNICAÇÃO

SUBÁREA: SONOLOGIA

Willy Garabini Cornelissen
CEGeME, Escola de Música, UFMG

Maurício Alves Loureiro
CEGeME, Escola de Música, UFMG

Resumo: A detecção do instante de início de notas (*note onset*) e sua percepção vem sendo estudada há muito tempo. A detecção de (*note onset*) de eventos é amplamente aplicada em diversos campos: eletrocardiogramas, dados sismográficos, resultados de mercado de ações e muitas tarefas de MIR (Transcrição Automática de Música, Detecção de Batidas, etc). Recentemente, várias abordagens baseadas em aprendizado de máquina tem apresentado potencial de otimizar e até superar métodos mais tradicionais de detecção. Este trabalho faz uma revisão dos métodos de detecção e compara com a nova abordagem oferecida pelo aprendizado de maquina.

Palavras-chave: Onsets. Pesquisa de Informações Musicais. Processamento Digital de Som. Aprendizado de Máquina.

Automatic note onset detection using machine learning

Abstract: The detection of the note start time (*note onset*) and its perception has been studied for a long time. The detection of *onsets* is widely applied in many fields: electrocardiograms, seismographic data, stock market results, and many MIR (Automatic Music Transcription, Beating Detection, etc.) tasks. Recently, several approaches based on machine learning have presented the potential to optimize and even surpass more traditional methods of detection. This work reviews methods of detection and compares with the new approach offered by machine learning.

Keywords: Onsets. Music Information Retrieval. Digital Signal Processing, Machine Learning.

1. Introdução

A capacidade de ouvir, entender e reconhecer sons é vital em muitos aspectos da nossa vida. Entre os sentidos humanos, na audição, mais do que em qualquer outro sentido, a percepção temporal de estímulos sonoros é muito importante e transmite muita informação semântica. O tempo é uma dimensão muito importante na audição, já que quase todos os sons flutuam com o tempo (MOORE, 2012). Além disso, para sons que transmitem informações, como fala e música, grande parte da informação parece ser carregada nas próprias mudanças, e não em partes dos sons que são relativamente estáveis. Dada a importância fundamental das mudanças temporais na audição, não é surpreendente que a maioria dos sistemas auditivos seja "rápida", pelo menos em comparação com outros sistemas sensoriais, como, por exemplo, a percepção de mudanças de frequência (YOST; FAY, 2012). Daí nasce a necessidade de detectar com precisão instantes temporais em que começam eventos sonoros.

2. Objetivos

Este documento investiga as etapas para a detecção de início de notas musicais *note onsets* e os vários métodos propostos para resolver o problema da sua detecção. Além disso, aponta a contribuição que o aprendizado de máquina trouxe para este domínio, apresentando perspectivas de superar métodos tradicionais de detecção em precisão e cobertura.

Na Seção 3, a definição da tarefa de detecção de *onsets* será apresentada e, na Seção 4, uma breve linha do tempo de sua história será desenhada. Na Seção 5, as etapas do processo de detecção serão analisadas e expostas junto com os resultados na seção 6. Finalmente, na seção 7, as conclusões finais são tecidas.

3. Definição de *onset*

Não há uma definição formal para o *onset* de eventos sonoros. Várias definições foram tentadas com foco no instante inicial de um evento (BELLO et al., 2005) e na percepção humana de *onsets* (VOS; RASCH, 1981). Ainda não existe uma solução final e uma definição formal, embora ambas abordagens não impliquem em forte disparidade, em termos perceptivos. Uma definição recente e ampla foi dada por Jan Schluter e Sebastian Bock, na qual um *onset* é definido pelo instante de tempo em que um evento relevante ocorre em um sinal de áudio (SCHLÜTER; BÖCK, 2014). Do ponto de vista musical, detectar um inícios de eventos sonoros é identificar o ponto de início de eventos relevantes, como, por exemplo, de notas musicais.

Em seu livro *Machine Audition*, no Capítulo 12, Zhou e Reiss (WANG, 2010) definem o sinal de áudio como uma sucessão de eventos acústicos discretos. Os autores se referem a *Music Onset Detection*, como a detecção do instante de início de um evento discreto, como uma nota musical. A detecção de *onsets* é essencial para o processamento de sinais musicais, visando uma ampla gama de aplicações, tais como transcrição automática de música, controle de batidas, identificação de andamento e pesquisa de informações musicais.

Bello et al. (BELLO et al., 2005; DIXON, 2006) postula que uma questão central é fazer uma distinção clara entre os conceitos relacionados a transientes, *onsets* e ataques, (Figura 3), cada demandando abordagens distintas, dependendo da aplicação.

- Ataque: é o intervalo de tempo durante o qual o envelope de amplitude aumenta.
- Transiente: não é fácil descrever com precisão, mas são os intervalos curtos durante os quais o sinal evolui rapidamente de alguma forma não trivial ou relativamente imprevisível.

- *Onset*: definido pelo único instante escolhido para marcar o transiente temporariamente prolongado. Na maioria dos casos, coincidirá com o início do transiente ou com o instante mais antecipado em que o transiente pode ser detectado com confiabilidade.

Voos e Rach (VOS; RASCH, 1981) definiram o início perceptual de um som musical como um instante temporal no qual o estímulo é percebido pela primeira vez. O início físico, no entanto, pode ser definido como o instante em que a geração do estímulo foi iniciada. Geralmente, o início perceptivo é atrasado em relação ao início físico. O intervalo de tempo entre o início físico e o perceptivo é resultante, entre outras coisas, pelo fato de que a maioria dos estímulos musicais e de fala não se iniciam em níveis próximos a seu máximo, mas começam com amplitudes que crescem gradualmente. No início do estímulo físico, o nível de amplitude é frequentemente muito baixo para atrair a atenção consciente do ouvinte. Neste trabalho seguiremos a definição de *onset* proposta por Bello: (BELLO et al., 2005): instante inicial de um evento sonoro.

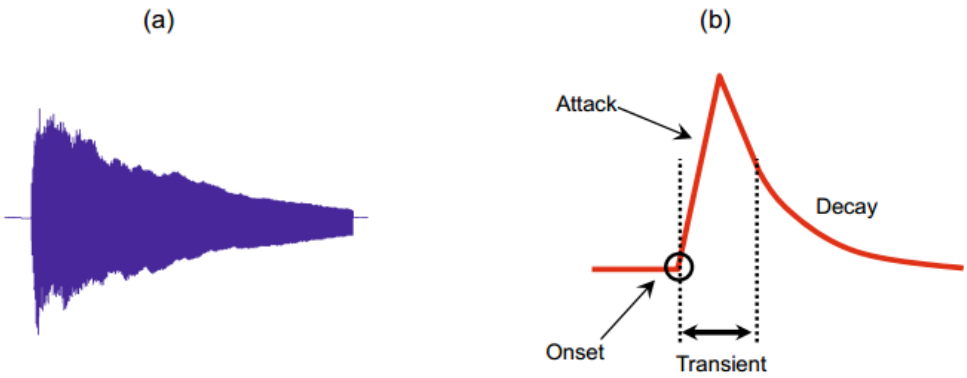

Figura 1: Uma nota tocada no piano (a) e seu envelope de amplitude indicando as regiões de *ataque, transiente, onset* e *decaimento* (b) adaptado de (MÜLLER, 2015, p. 305)

.

4. Histórico da detecção de *onsets*

A percepção de eventos relevantes em sinais de áudio vem sendo estudada, desde os anos oitenta, identificados como *onsets* perceptuais (VOS; RASCH, 1981). Neste período, ferramentas no domínio do tempo eram mais frequentemente usadas para detectar as variações no sinal ao longo do tempo.

Nos anos noventa, um trabalho importante de Masri (MASRI, 1996) faz uso da linha espectral do sinal como um novo caminho para detecção de *onset*. O conteúdo espectral em altas frequências e alterações bruscas na distribuição espectral foram utilizados para definir os instantes de início de eventos. Em trabalho posterior, Klapuri (KLAPURI, 1999)

persiste no uso do envelope da forma de onda, mas agregado com separações de bandas de frequência e conhecimento psicoacústico.

A partir da emergência do paradigma de aprendizado de máquina, surge, em 2002, o perceptron [1] de multicamadas (*Multi Layer Perceptron - MLP*) com filtros de detecção, que foi aplicado na detecção de instantes de início de notas em um estudo de Marolt e colegas (MAROLT; KAVCIC; PRIVOSNIK, 2002) com obras executadas no piano.

Outras arquiteturas foram construídas usando o várias abordagens matemáticas e probabilísticas: *Support Vector Machines* (DAVY; GODSILL, 2002), *Short-Time Fourier Transform* (DUXBURY; SANDLER; DAVIES, 2002), *Hidden Markov Model* (ABDALLAH; PLUMBLEY, 2003), *Complex Domain* (DUXBURY et al., 2003), *Phase Based* (BELLO; SANDLER, 2003), *Multi Function Detection* (LEVEAU; DAUDET, 2004).

Em 2005, um artigo seminal de Juan Pablo Bello apresenta um tutorial de detecção de *onsets* (BELLO et al., 2005), sinalizando que a real importância da detecção de *note onsets* surgia. Ao mesmo tempo, com os primeiros esforços da Sociedade Internacional para a Recuperação de Informações Musicais - ISMIR (*International Society Musical Information Retrieval*) [2] este tipo de pesquisa musical se proliferou. Desta comunidade de pesquisadores surgiu o MIREX (*Music Information Research Evaluation eXchange*) [3].

A proposição de diferentes abordagens continuou em ritmo acelerado: *Adaptive Whitening* (STOWELL; PLUMBLEY, 2007), *Neural Networks* (LACOSTE; ECK, 2007), *Pitch Information* (ZHOU; MATTAVELLI; ZOIA, 2008), *Multiple-feature Fusion* (TOH; ZHANG; WANG, 2008), *Psycho-Acoustic Filters* (THOSHKAHNA; RAMAKRISHNAN, 2008), *Mel-Frequency Cepstral Coefficients* (ZHANG; W ANG, 2 009), *Multi-Dimension Detection* (HOLZAPFEL et al., 2010).

Desde então, a busca por melhores Funções de Detecção de *Onset* (*ODF - Onset Detection Function*) estava acontecendo principalmente no paradigma de *aprendizado de máquina (machine learning*), buscando a melhor arquitetura e customização para otimizar a tarefa de detecção. Muitas abordagens foram propostas e competem entre si em competições promovidas pelo MIREX, visando alcançar o estado da arte na detecção do *onset*: *BiLSTM-NN* (EYBEN et al., 2010), *Transient Detection with STFT* (THOSHKAHNA; NSABIMANA;

on Retrieval Systems Evaluation Laboratory, sediado na University of Illinois at Urbana-Champaign (UIUC)

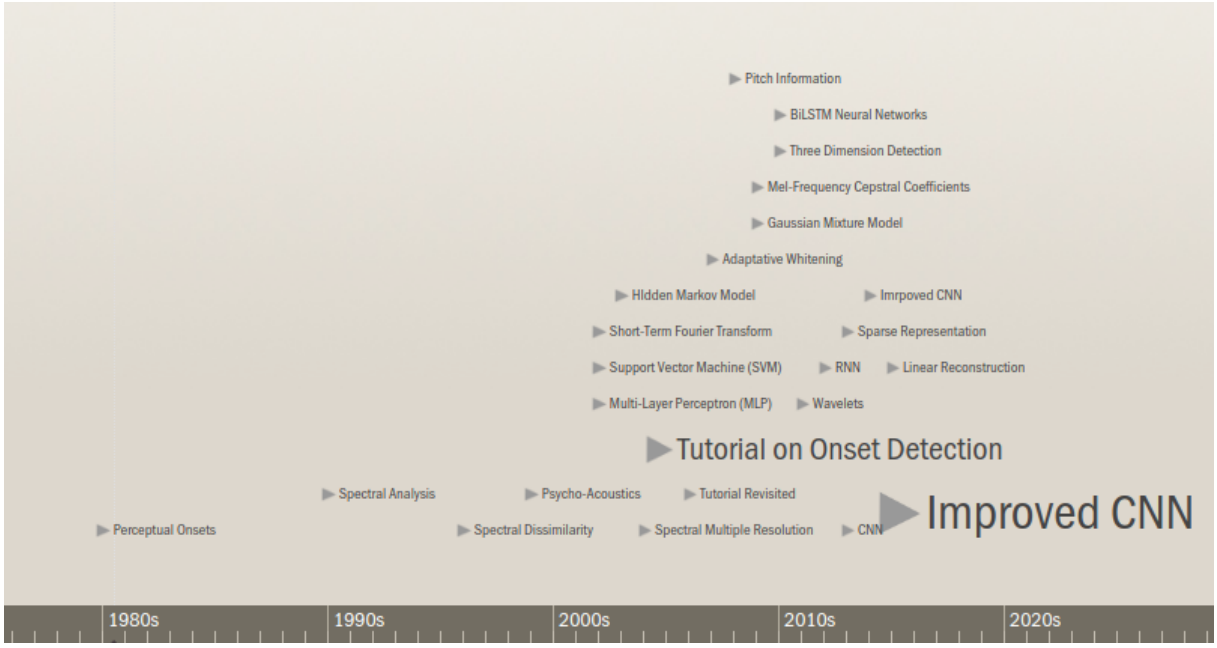

Figura 2: Linha do tempo dos métodos mencionados de detecção de *onset*

KALPATHI, 2011), *Wavelets and Adaptative Filters* (GABRIELLI; PIAZZA; SQUARTINI, 2011), *RNN* (BÖCK et al., 2012), *Segmentation* (GLOVER; LAZZARINI; TIMONEY, 2012), *CNN* (SCHLÜTER; BÖCK, 2013) and (SCHLÜTER; BÖCK, 2014). A figura 2 exibe a linha do tempo dos métodos acima mencionados.

5. Metodologia de aprendizado de máquina na detecção de *onsets*

A aplicação de aprendizado de máquina na detecção de *onsets* envolve processos que podemos resumir de uma maneira geral em três etapas, *anotação, extração de características* e *treinamento do modelo*, as quais detalhamos sucintamente a seguir.

Anotação: Esta etapa envolve a detecção dos inícios de notas manualmente por especialistas, por exemplo, músicos profissionais de forma a garantir melhor indicação possível dos inícios das notas musicais, os quais serão estabelecidos como valores verdadeiros (*ground truth*) a ser utilizado para treinar os modelos de *machine learning*

Extração de características: Esta etapa envolve o processamento do sinal analisado, a partir de experiências de laboratório, que buscam por candidatos a instantes de *onset* a partir de observações de variações no sinal processado. Diferentes técnicas de processamento apresentam suas especificidades e podem funcionar melhor em algumas situações e não em outras. A observação destas características é que permitirão dizer onde estão localizados os *onsets*. Recentemente, essas características tem siso usadas na construção de modelos de detecção automática.

Treinamento do modelo: A ação de treinar um modelo para detectar *onsets* em um sinal de áudio equivale a encontrar uma equação que gere o resultado desejado (*onset* ou *não-onset*) a partir do processamento das informações de entrada(características do audio). Este modelo é construído a partir de um algoritmo que vai refinando os coeficientes desta equação iterativamente. Ou seja, a partir de informações construídas na fase de anotações, é gerado um modelo que vai detectar a presença de onsets em cada um dos frames analisados. A maneira de treinar o modelo é que define as várias arquiteturas(ANN, SVM, CNN, RNN e outras) de aprendizado de máquina que podem ser utilizadas.

Este processo de detecção de *onsets* usando modelos computacionais é similar aos experimentos de Bioinformática chamados *in silico*. A expresão *in silico* foi usada publicamente pela primeira vez em 1989 no workshop "Autômatos Celulares: Teoria e Aplicações", em Los Alamos, EUA(TISSEAU, 2008). É uma expressão usada no âmbito da simulação computacional e áreas correlatas para indicar algo ocorrido através de uma simulação computacional.

6. Resultados

Para comparar os métodos de detecção de *onsets* e confirmar a convergência das diferentes abordagens de aprendizado de máquina, usamos um conjunto de dados disponibilizado pelo MIREX [4] para a a avaliação de cada método de detecção. O conjunto de dados do MIREX é composto pot 17 trilhas musicais, com 10 exemplos de músicas monofônicas e 7 exemplos polifônicos (mais de uma voz).

A métrica padronizada pelo MIREX é a medida denominada *f-measure*, também chamada de *f-score* ou *f1-measure*. É definida como a média harmônica ponderada da *precisão (precision)* e da *revocação (recall)* do teste. A precisão, também chamada de valor preditivo positivo, é a proporção de resultados positivos que são verdadeiramente positivos. Revocação, também chamado de sensibilidade, é a capacidade de um teste para identificar corretamente os resultados positivos para obter a taxa positiva verdadeira. O valor de *f-measure* alcança o valor 1 quando as medidas de *precision* e *recall* são máximas e zero quando este valores são mínimos. O valor de *f-measure* é estimado pela equação:

*F = 2 * (precision x recall / (precision + recall)* [5].

Este conjunto de dados foi submetido a dez métodos de detecção já citados neste documento. Para fazer a comparação entre os métodos, foram colhidas três métricas que são

padronizadas pelo MIREX: *precision, recall* e *f-measure*.

As figura 3, 4 e 5 mostram a porcentagem atingida em cada métrica, por cada um dos 10 métodos avaliados, *Envelope, Energy, HFC, Phase Deviation, Spectral Difference, Complex Domain, Super Flux, Bi-LSTM, RNN e CNN*.

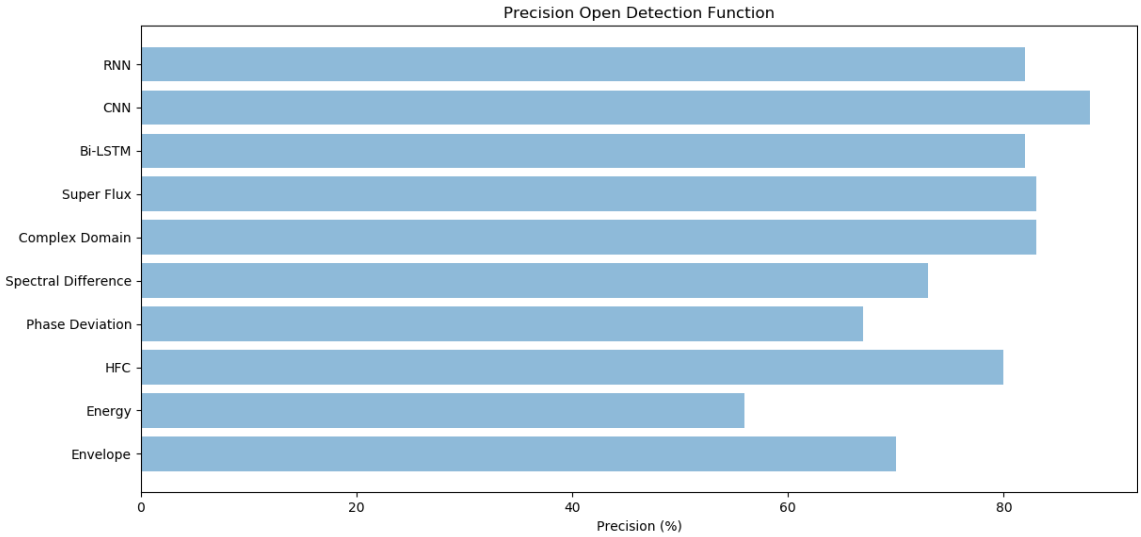

Figura 3: Porcetagem atingida por cada um dos 10 métodos avaliados, *Envelope, Energy, HFC, Phase Deviation, Spectral Difference, Complex Domain, Super Flux, Bi-LSTM, RNN e CNN*, da métrica *precision*.

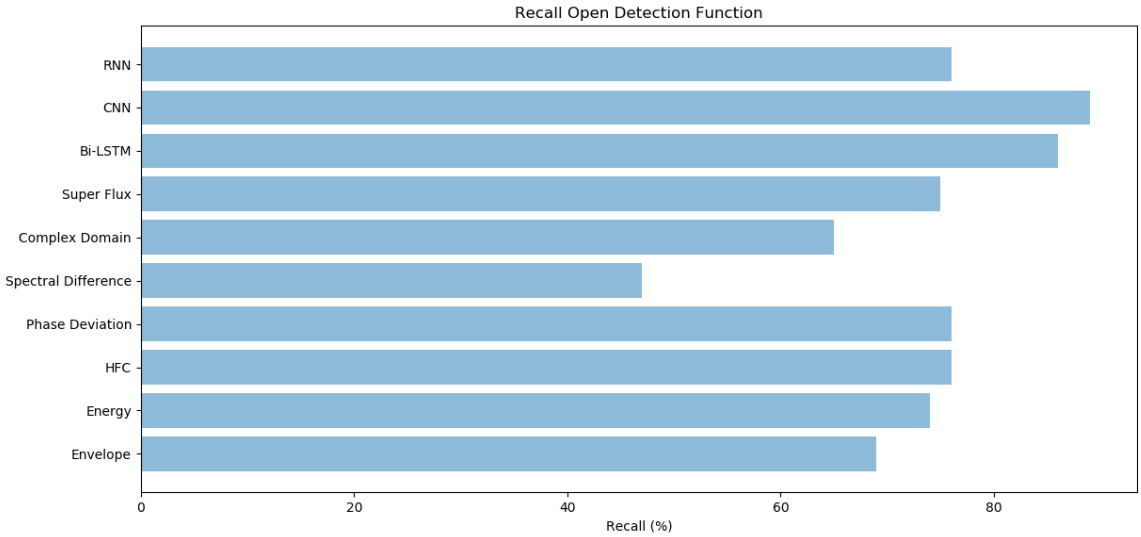

Figura 4: Porcetagem atingida por cada um dos 10 métodos avaliados, na métrica *recall*.

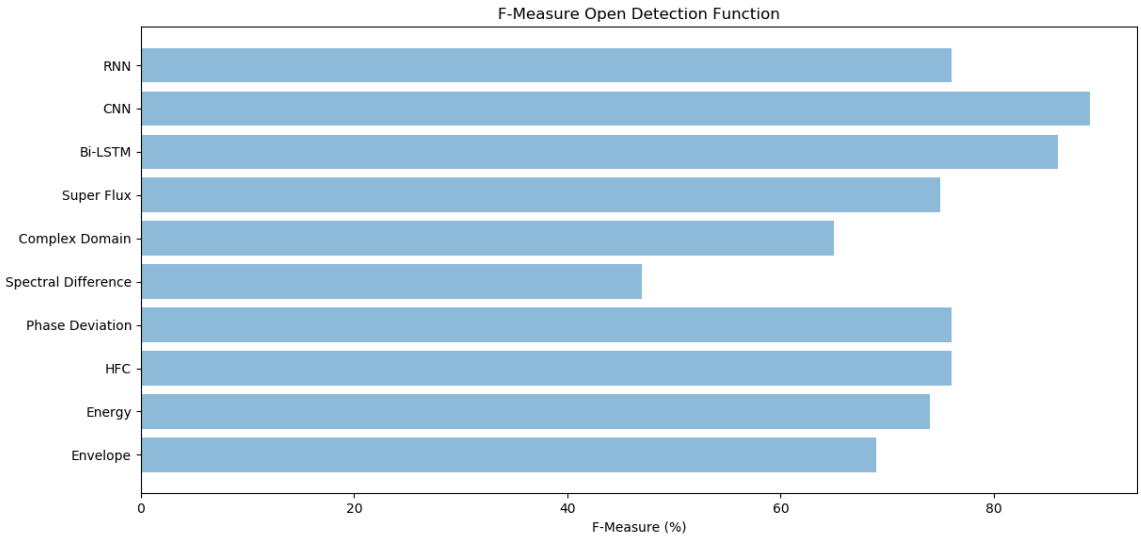

Figura 5: Porcetagem atingida por cada um dos 10 métodos avaliados, na métrica *F-Measure*.

7. Discussão

Nota-se uma clara vantagem das técnicas *Bi-LSTM, RNN* e *CNN*. Considerado o estado da arte atual na detecção de *onsets*, o método *CNN* obteve 0.88 de precisão, 0.89 de revocação e 0.88 de *f-measure*. No contexto de inteligência artificial e aprendizagem de máquina, uma rede neural convolucional (CNN do inglês Convolutional Neural network ou ConvNet) é uma classe de rede neural artificial do tipo feed-forward, que vem sendo aplicada com sucesso no processamento e análise de imagens digitais. Originalmente aplicadas na área de visão computacional (computer vision), as CNNs mantiveram sua eficácia quando foram aplicadas as outros cenários, como é o caso da detecção de *onsets*. Esta eficácia pode ser explicada pela utilização da imagem dos espectrogramas para detectar os instantes dos *onsets*.

8. Conclusões

Nos últimos vinte anos, com o aperfeiçoamento dos computadores e os avanços nas técnicas de aprendizado de máquina, o campo chamado MIR (Music Information Retrieval) se desenvolveu massivamente. Este importante domínio da pesquisa musical trouxe resultados impressionantes, que conseguiram resolver problemas que pareciam insolúveis, quando aplicados a sinais musicais de maior complexidade (CELLA, 2017).

Este trabalho mostrou como o aprendizado de máquina integrou-se bem ao processo de de detecção de *onsets* e vem mostrando contribuições importantes para a otimização de métodos mais tradicionais. Por se tratar de uma área emergente e dinâmica, onde novas soluções são apresentadas frequentemente, a expectativa é que possam surgir aperfeiçoamentos e arquiteturas com potencial de melhorar ainda mais a tarefa de detecção.

Referências:

ABDALLAH, Samer A; PLUMBLEY, Mark D. Probability as metadata: event detection in music using ica as a conditional density model. In: CITESEER. *Proc. 4th Int. Symp. Independent Component Analysis and Signal Separation (ICA2003)*. [S.l.], 2003. p. 233–238.

BELLO, Juan Pablo et al. A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing*, IEEE, v. 13, n. 5, p. 1035–1047, 2005.

BELLO, Juan Pablo; SANDLER, Mark. Phase-based note onset detection for music signals. In: IEEE. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*. [S.l.], 2003. v. 5, p. V–441.

BÖCK, Sebastian et al. Online real-time onset detection with recurrent neural networks. In: *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12), York, UK*. [S.l.: s.n.], 2012.

CELLA, Carmine-Emanuele. Machine listening intelligence. *arXiv preprint arXiv:1706.09557*, 2017.

DAVY, Manuel; GODSILL, Simon. Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation. In: IEEE. *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*. [S.l.], 2002. v. 2, p. II–1313.

DIXON, Simon. Onset detection revisited. In: CITESEER. *Proceedings of the 9th International Conference on Digital Audio Effects*. [S.l.], 2006. v. 120, p. 133–137.

DUXBURY, Chris et al. Complex domain onset detection for musical signals. In: QUEEN MARY UNIVERSITY LONDON. *Proc. Digital Audio Effects Workshop (DAFx)*. [S.l.], 2003. v. 1, p. 6–9.

DUXBURY, Chris; SANDLER, Mark; DAVIES, Mike. A hybrid approach to musical note onset detection. In: *Proc. Digital Audio Effects Conf.(DAFX,'02)*. [S.l.: s.n.], 2002. p. 33–38.

EYBEN, Florian et al. Universal onset detection with bidirectional long-short term memory neural networks. In: *Proc. 11th Intern. Soc. for Music Information Retrieval Conference, ISMIR, Utrecht, The Netherlands*. [S.l.: s.n.], 2010. p. 589–594.

GABRIELLI, Leonardo; PIAZZA, Francesco; SQUARTINI, Stefano. Adaptive linear prediction filtering in dwt domain for real-time musical onset detection. *EURASIP Journal on Advances in Signal Processing*, Springer, v. 2011, n. 1, p. 650204, 2011.

GLOVER, John; LAZZARINI, Victor; TIMONEY, Joseph. Real-time segmentation of the temporal evolution of musical sounds. In: ASA. *Proceedings of Meetings on Acoustics 163ASA*. [S.l.], 2012. v. 15, n. 1, p. 035002.

HOLZAPFEL, André et al. Three dimensions of pitched instrument onset detection. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 18, n. 6, p. 1517–1527, 2010.

KLAPURI, Anssi. Sound onset detection by applying psychoacoustic knowledge. In: IEEE. *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*. [S.l.], 1999. v. 6, p. 3089–3092.

LACOSTE, Alexandre; ECK, Douglas. A supervised classification algorithm for note onset detection. *EURASIP Journal on Applied Signal Processing*, Hindawi Publishing Corp., v. 2007, n. 1, p. 153–153, 2007.

LEVEAU, Pierre; DAUDET, Laurent. Methodology and tools for the evaluation of automatic onset detection algorithms in music. In: CITESEER. *In Proc. Int. Symp. on Music Information Retrieval*. [S.l.], 2004.

MAROLT, Matija; KAVCIC, Alenka; PRIVOSNIK, Marko. Neural networks for note onset detection in piano music. In: *Proceedings of the 2002 International Computer Music Conference*. [S.l.: s.n.], 2002.

MASRI, Paul. *Computer modelling of sound for transformation and synthesis of musical signals.* Tese (Doutorado) — University of Bristol, 1996.

MOORE, Brian CJ. *An introduction to the psychology of hearing*. [S.l.]: Brill, 2012.

MÜLLER, Meinard. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. [S.l.]: Springer, 2015.

SCHLÜTER, Jan; BÖCK, Sebastian. Musical onset detection with convolutional neural networks. In: *6th International Workshop on Machine Learning and Music (MML), Prague, Czech Republic*. [S.l.: s.n.], 2013.

SCHLÜTER, Jan; BÖCK, Sebastian. Improved musical onset detection with convolutional neural networks. In: *ICASSP*. [S.l.: s.n.], 2014. p. 6979–6983.

STOWELL, Dan; PLUMBLEY, Mark. Adaptive whitening for improved real-time audio onset detection. In: *Proceedings of the 2007 International Computer Music Conference, ICMC 2007*. [S.l.: s.n.], 2007. p. 312–319.

THOSHKAHNA, Balaji; NSABIMANA, Francois Xavier; KALPATHI, Ramakrishnan R. A transient detection algorithm for audio using iterative analysis of stft. In: *ISMIR*. [S.l.: s.n.], 2011. p. 203–208.

THOSHKAHNA, Balaji; RAMAKRISHNAN, KR. A psychoacoustics based sound onset detection algorithm for polyphonic audio. In: IEEE. *Signal Processing, 2008. ICSP 2008. 9th International Conference on*. [S.l.], 2008. p. 1424–1427.

TISSEAU, Jacques. In vivo, in vitro, in silico, in virtuo. In: *1st Workshop on SMA in Biology at meso or macroscopic scales, Paris*. [S.l.: s.n.], 2008.

TOH, Chee-Chuan; ZHANG, Bingjun; WANG, Ye. Multiple-feature fusion based onset detection for solo singing voice. In: *ISMIR*. [S.l.: s.n.], 2008. p. 515–520.

VOS, Joos; RASCH, Rudolf. The perceptual onset of musical tones. *Perception & psychophysics*, Springer, v. 29, n. 4, p. 323–335, 1981.

WANG, Wenwu. *Machine Audition: Principles, Algorithms and Systems: Principles, Algorithms and Systems*. [S.l.]: IGI Global, 2010.

YOST, William A; FAY, Richard R. *Human psychophysics*. [S.l.]: Springer Science & Business Media, 2012. v. 3.

ZHANG, Bingjun; WANG, Ye. Automatic music transcription using audio-visual fusion for violin practice in home environment. TRA7/09, 2009.

ZHOU, Ruohua; MATTAVELLI, Marco; ZOIA, Giorgio. Music onset detection based on resonator time frequency image. *IEEE Transactions on Audio, Speech, and Language Processing*, IEEE, v. 16, n. 8, p. 1685–1695, 2008.

Notas

[1]O perceptron é um tipo de rede neural artificial inventada em 1957 por Frank Rosenblatt no Cornell Aeronautical Laboratory. Ele pode ser visto como o tipo mais simples de rede neural feedforward: um classificador linear.

[2]A ISMIR é um fórum internacional para pesquisas sobre a organização de dados relacionados à música https://www.ismir.net/

[3]O MIREX é um framework comunitário para a avaliação padronizada de tarefas de MIR (Musica Information Retrieval), coordenado e administrado pelo Laboratório IMIRSEL (International Music Informati

[4]https://www.music-ir.org/mirex/wiki/MIREX-HOME

[5]https://deepai.org/machine-learning-glossary-and-terms/f-score