Automatic onset detection using convolutional neural networks

Willy Garabini Cornelissen[1] **, Mauricio Alves Loureiro**[1]

[1]CEGeME – Escola de Musica/UFMG
Av. Antonio Carlos, 6627, Pampulha - Belo Horizonte - MG

willy@ufmg.br, mauricio@musica.ufmg.br

Abstract. *A very significant task for music research is to estimate instants when meaningful events begin (onset) and when they end (offset). Onset detection is widely applied in many fields: electrocardiograms, seismographic data, stock market results and many Music Information Research(MIR) tasks, such as Automatic Music Transcription, Rhythm Detection, Speech Recognition, etc. Automatic Onset Detection(AOD) received, recently, a huge contribution coming from Artificial Intelligence (AI) methods, mainly Machine Learning and Deep Learning. In this work, the use of Convolutional Neural Networks (CNN) is explored by adapting its original architecture in order to apply the approach to automatic onset detection on audio musical signals. We used a CNN network for onset detection on a very general dataset, well acknowledged by the MIR community, and examined the accuracy of the method by comparison to ground truth data published by the dataset. The results are promising and outperform another methods of musical onset detection.*

1 Introduction

The extraction of onset times from a spectrogram is equivalent of detecting edges on an image. Oriented edges in images can be found by convolution with small filter kernels even of random values. This lead to the idea of training a Convolutional Neural Network (CNN) to find onsets in spectrogram excerpts. Convolutional learning in Music Information Research has been applied before for genre and artist classification [[1], [2]]. Their application on onset detection, a comparably low-level task, achieve promising results. [[3],[4]]

2 Onset definition

Flights and Rach [5] defined the perceptual beginning of a musical sound as a time instant in which the stimulus is perceived for the first time. The physical onset, however, can be defined as the instant at which the generation of the stimulus was initiated. Usually, the perceptive onset is delayed in relation to physical onset. The time interval between the physical and the perceptual initiation results, among other things, from the fact that most musical and speech stimuli do not begin at levels near their maximum, but begin with gradually increasing amplitudes. At the beginning of the physical stimulus, the amplitude level is often too low to attract the conscious attention of the listener. In this work we will follow the definition of onset proposed by Bello: initial instant of a sound event [6].

3 Convolutional Neural Networks

Convolution is the process of adding each element of the image to its local neighbors, weighted by the kernel. The kernel, or convolutional matrix, is multiplied to the original matrix, resulting in a single value, as shown in the example. This operation gives to CNN a high accuracy on image recognition tasks, although the computational cost is high and need a lot of training data.

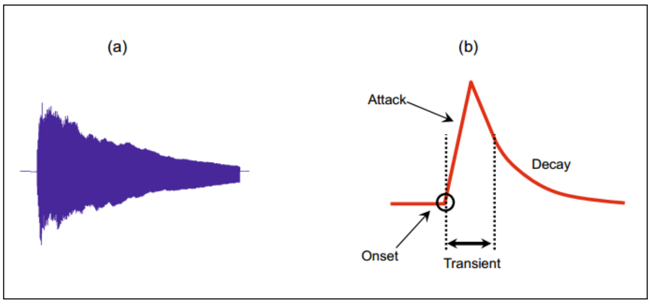

Figure 1: A note played on the piano (a) and its amplitude envelope indicating the regions of attack, transient, onset and decay (b) adapted from [7, p. 305]

.

3.1 Computer Vision and Machine Listening

CNNs are great for computer vision task, bu to apply it in spectrograms for machine listening, some challenges must be overcomed:

- **Sound objects are transparent:** visual objects and sound events in a image behaves differently. The problem is that discrete sound events do not separate into layers on a spectrogram: Instead, they all sum together into a distinct whole. Visual objects are ”individualized”, and sound events in a spectrogram cannot be assumed to belong to a single sound, as the ”magnitude of of that frequency could have been produced by any number of accumulated sounds or even by the complex interactions between sound waves such as phase cancellation. This makes it difficult to separate simultaneous sounds in spectrogram representations.” [8]
- **Meaning of axes:** one of the big advantages of a CNN is that they are built on the assumption that features of an image carry the same meaning regardless of their location. But dealing with spectrograms, the two dimensions represent fundamentally different units, one being strength of frequency and the other being time. Moving a sound event horizontally is just shift in time, but to move it vertically causes a notable change on its nature. Therefore, the spatial invariance that

2D CNNs provide might not perform as well for this form of data.

- **Sounds are not local:** the frequencies represented in a spectrogram are not locally grouped. Instead of this, they move together according to a common relationship(the fundamental frequency).
- **Sound is a temporal event:** in a visual scenario, objects persist on time and can be re-scanned. This it not true true for sound events. This is why it makes sense to refer to these phenomena as sound events rather than sound objects.

4 Methodology

4.1 Data

Sebastian Bock, the author of the model called state of the art (SOTA) in onset detection [4], prepared a dataset that we used as graund-truth to illustrate onset detection using CNN. The dataset contains 321 audio excerpts taken from various sources. 87 tracks were taken from the dataset used in [9], 23 from [6], and 92 from [10].

4.2 Detection Method

The method use two convolutional and pooling layers to perform local processing, a single fully-connected hidden layer and a single output unit.

4.3 Evaluation Metric

For the comparison of detected onset with the ground-truth, if the detected instant falls within a tolerance time-window around that instant, it is considered as a true positive (TP). If not, there is a false negative (FN). The detections outside all the tolerance windows are counted as false positives (FP). Doubled onsets (two detections for one ground-truth onset) and merged onsets (one detection for two ground-truth onsets) will be taken into account in the evaluation. Doubled onsets are a subset of the FP onsets, and merged onsets a subset of FN onsets. **Precision, Recall** and **F-measure** were used to evaluate the performance of the detection.

5 Results

Figure 2 illustrate the power of the state of the art algorithm. The histogram of this figure shows that most outcomes of f-measure lies between 0.8 and 0.9 with median 0.9 and a and a concentration of files that achieve a value between 0.75 and 0.95 for all these metrics.

6 Conclusion

This work showed how machine learning with convolutional neural networks was well integrated in the process of detecting onsets and has been showing important contributions to the optimization of more traditional methods. It was verified that this method performs very well in a large and generic dataset, confirming the state of the art achieved by CNN on AOD.

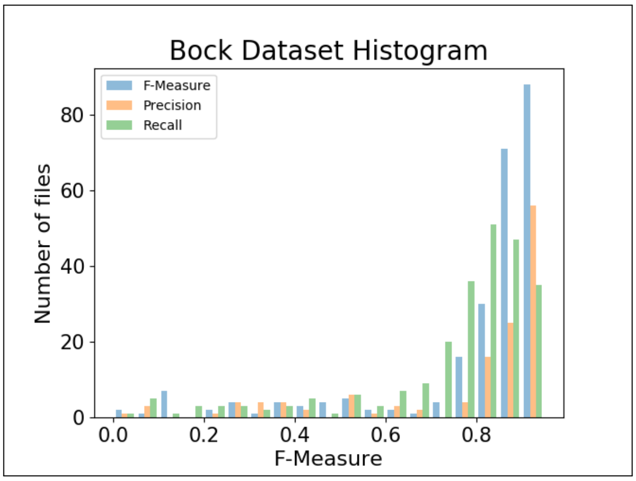

Figure 2: Bock Dataset Histogram with CNN Onset Detection. Blue for *f-measure*, orange for *precision*— and green for *recall*.

References

[1] Tom LH Li, Antoni B Chan, and A Chun. Automatic musical pattern feature extraction using convolutional neural network. In *Proc. Int. Conf. Data Mining and Applications*, volume 161. Citeseer, 2010.

[2] Sander Dieleman, Philémon Brakel, and Benjamin Schrauwen. Audio-based music classification with a pretrained convolutional network. In *12th International Society for Music Information Retrieval Conference (ISMIR-2011)*, pages 669–674. University of Miami, 2011.

[3] Jan Schlüter and Sebastian Böck. Musical onset detection with convolutional neural networks. In *6th international workshop on machine learning and music (MML), Prague, Czech Republic*, 2013.

[4] Jan Schlüter and Sebastian Böck. Improved musical onset detection with convolutional neural networks. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 6979–6983. IEEE, 2014.

[5] Joos Vos and Rudolf Rasch. The perceptual onset of musical tones. *Perception & psychophysics*, 29(4):323–335, 1981.

[6] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on speech and audio processing*, 13(5):1035–1047, 2005.

[7] Meinard Müller. *Fundamentals of music processing: Audio, analysis, algorithms, applications*. Springer, 2015.

[8] What's wrong with spectrograms and cnns for audio processing? `https://t.co/qequ0e3l18`. (Accessed on 05/14/2019).

[9] Florian Eyben, Sebastian Böck, Björn Schuller, and Alex Graves. 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*, pages 589–594, 2010.

[10] André Holzapfel, Yannis Stylianou, Ali C Gedik, and Barış Bozkurt. Three dimensions of pitched instrument onset detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1517–1527, 2010.