Machine learning for note onset detection.

Alexandre Lacoste & Douglas Eck
Outline

- What is note onset detection and why is it useful?
- Small review of the field
- The details of the incredible algorithm
- Results of the contest
- Results of the custom dataset
What are note onsets?

- Percussive instruments are modeled as shown (right).
- Basic definition:
  Note onset is the time where the slope is the highest, during the attack time.
More general definition

- What happens if we have sounds that are not percussive? (pitch changing, singing, vibrato ...)
- Then we define onsets as being unpredictable events.
- If, with information near in the past, we can’t predict the future, then a new event just arrived.
- This is the definition used to label the onsets.
Onset detection is not trivial

- In other words, percussive note onsets in monophonic songs is trivial.

- But if you want to make it work for complex polyphonic with singing, it is another story.
What can we do with a good note onset detector?

Not directly useful, but it is present in many music algorithms.

- Music transcription (from wave to midi)
- Music editing (Song segmentation)
- Tempo tracking (with onset, finding the tempos is much easier)
- Musical fingerprinting (the onset trace can serve as a robust id for fingerprinting)
Scheirer’s Psycho-acoustical experiment

- Scheirer showed that only the envelope of a few frequency band was important for the rhythmical information.
- By modulating the envelopes with a noise source, the song can be rebuilt and almost no rhythmical aspect is lost.
Most onset detection algorithms use Scheirer’s model and use a filter to find positive slopes. For example: $I_i - I_{i-1}$

Then, they use a peak-picking algorithm to find the onset position.

This method is fast simple and works fine for monophonic percussive songs.

But it got very poor results on complex polyphonic with singing.

And it is very sensitive to parameter adjustment.
The information is mainly local in time

- Why not apply a simple feed-forward neural network directly on all the inputs of the window.
- And just ask if there is an onset at this position.
- Finally, we repeat this for every time step.
The algorithm can be split in 3 main steps

- Get the spectrogram of the song
- Convolve a feed-forward neural network across the spectrogram
- Find the onset location
Many different time-frequency representation might be useful for this task. Let’s explore some of them.

1. Short-time Fourier transform (STFT)
2. Constant-Q transform
3. Phase plane of STFT
Short-time Fourier Transform

- $STFT(t, \omega) = \int_{-\infty}^{\infty} x(\tau) w^*(\tau - t) e^{-j\omega \tau} d\tau$
- The yellow curve represents the onset time
The constant-Q transform has a logarithmic frequency scale which provides:

- a much better frequency resolution for lower frequency.
- a better time resolution for high frequency.
Can we do something with the phase plane?

- The phase plane, without any manipulation, doesn’t seem to contain any information.
Phase Acceleration

- Bello and Sandler [1] have found a way to use phase information for onset detection.
  - They take the principal argument of the phase acceleration.
  - \[ \alpha_{k,n} = \text{princarg}\left[\varphi_{k,n} - 2\varphi_{k,(n-1)} + \varphi_{k,(n-2)}\right] \]

Patterns not evident enough!
Instead, if we simply take the difference along the frequency axis, we get interesting patterns.

\[ \omega_{k,n} = \text{princarg}\left[\varphi_{k,n} - \varphi_{(k-1),n}\right] \]

Results show performance equivalent to the magnitude plane, using only the phase.
Feed Forward Neural Network

- Remember, the algorithm is simply the FNN convolved across time and frequency.
- The target is a mixture of thin Gaussians that represents the expectation of having an onset for time $t$. 
Net Inputs

- For a decent spectrogram resolution
  - Time: 200 bins / s
  - Frequency: 200 bins
- And a window width of 50 ms
- We have 2000 input variables
- This is too many !!!
- We randomly sample 200 variables inside the window.
  - Uniform distribution across frequency
  - Gaussian distribution across time (more variables near the center)
Net Structure and Training

- Two hidden layers
  - 20 units in the first layer
  - 15 units in the second layer
  - 1 output neuron
- Learning algorithm: Polak-Ribiere version of conjugate gradient
- K-fold cross-validation for performance estimation
Most peaks are really sharp and there is very low background noise.

Some peaks are smaller but still can be detected.

The precision is also very good.
The neural networks only emphasize the onsets.

We now have to find the location of each onset.

We simply apply a threshold.
- positive crossing is the beginning
- Negative crossing is the end
- Location is the center of mass

The value of the threshold is learned by exhaustive search.
To maximize the performance, we want to find the maximum number of onsets (Recall)

But we also want to minimize the number of spurious onsets (Precision)

The F-measure offers an equilibrium between the two.

\[ R = \frac{n_{cd}}{N_{\text{target}}} \]
\[ P = \frac{n_{cd}}{N_{\text{found}}} \]
\[ F = \frac{2PR}{P + R} \]
MIREX 2005 Results

- No other participants used machine learning.
- With a simple FNN, we have a huge performance boost.
- We also have the best equilibrium between precision and recall.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant</th>
<th>Avg F-measure</th>
<th>Avg Precision</th>
<th>Avg Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lacoste &amp; Eck (MULTI-NET)</td>
<td>80.07%</td>
<td>79.27%</td>
<td>83.70%</td>
</tr>
<tr>
<td>2</td>
<td>Lacoste &amp; Eck (SINGLE-NET)</td>
<td>78.35%</td>
<td>77.69%</td>
<td>83.27%</td>
</tr>
<tr>
<td>3</td>
<td>Ricard, J.</td>
<td>74.80%</td>
<td>81.36%</td>
<td>73.70%</td>
</tr>
<tr>
<td>4</td>
<td>Brossier, P.</td>
<td>74.72%</td>
<td>74.07%</td>
<td>81.95%</td>
</tr>
<tr>
<td>5</td>
<td>Rbel, A. (2)</td>
<td>74.64%</td>
<td>83.93%</td>
<td>71.00%</td>
</tr>
<tr>
<td>6</td>
<td>Collins, N.</td>
<td>72.10%</td>
<td>87.96%</td>
<td>68.26%</td>
</tr>
<tr>
<td>7</td>
<td>Rbel, A. (1)</td>
<td>69.57%</td>
<td>79.16%</td>
<td>68.60%</td>
</tr>
<tr>
<td>8</td>
<td>Pertusa, Klapuri, &amp; Iesta</td>
<td>58.92%</td>
<td>60.01%</td>
<td>61.62%</td>
</tr>
<tr>
<td>9</td>
<td>West, K.</td>
<td>48.77%</td>
<td>48.50%</td>
<td>56.29%</td>
</tr>
</tbody>
</table>
For better tests, we built a custom dataset.

It is composed only of complex polyphonic songs with singing.

There is in total 60 segments of 10 seconds.

The onsets were all hand-labeled, using a graphical user interface.
## Results for Different Spectrograms

<table>
<thead>
<tr>
<th>Plane</th>
<th>Spectral window size</th>
<th>F-meas train</th>
<th>F-meas valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>STFT log mag</td>
<td>10 ms</td>
<td>86 ± 2</td>
<td>86 ± 5</td>
</tr>
<tr>
<td>STFT log mag</td>
<td>30 ms</td>
<td>86 ± 1</td>
<td>86 ± 5</td>
</tr>
<tr>
<td>STFT log mag</td>
<td>100 ms</td>
<td>84 ± 2</td>
<td>83 ± 8</td>
</tr>
<tr>
<td>C-Q log mag</td>
<td>10 ms</td>
<td>86 ± 2</td>
<td>86 ± 5</td>
</tr>
<tr>
<td>C-Q log mag</td>
<td>30 ms</td>
<td>87 ± 2</td>
<td>87 ± 5</td>
</tr>
<tr>
<td>C-Q log mag</td>
<td>100 ms</td>
<td>84 ± 2</td>
<td>84 ± 6</td>
</tr>
<tr>
<td>STFT ph accel</td>
<td>10 ms</td>
<td>49 ± 2</td>
<td>49 ± 4</td>
</tr>
<tr>
<td>STFT ph accel</td>
<td>30 ms</td>
<td>47 ± 1</td>
<td>47 ± 5</td>
</tr>
<tr>
<td>STFT ph accel</td>
<td>100 ms</td>
<td>49 ± 4</td>
<td>47 ± 6</td>
</tr>
<tr>
<td>STFT ph freq-diff</td>
<td>10 ms</td>
<td>62 ± 2</td>
<td>61 ± 6</td>
</tr>
<tr>
<td>STFT ph freq-diff</td>
<td>30 ms</td>
<td>80 ± 1</td>
<td>79 ± 4</td>
</tr>
<tr>
<td>STFT ph freq-diff</td>
<td>100 ms</td>
<td>74 ± 2</td>
<td>73 ± 6</td>
</tr>
<tr>
<td>Noise</td>
<td>—</td>
<td>40 ± 2</td>
<td>40 ± 6</td>
</tr>
</tbody>
</table>
Combining Phase and Magnitude Does Not Help.

<table>
<thead>
<tr>
<th>planes</th>
<th>Nb input variables</th>
<th>Hamming window Size</th>
<th>F-meas train</th>
<th>F-meas valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>STFT log mag + ph freq-diff</td>
<td>100</td>
<td>30 ms</td>
<td>85 ± 2</td>
<td>84 ± 5</td>
</tr>
<tr>
<td>STFT log mag + ph freq-diff</td>
<td>100</td>
<td>50 ms</td>
<td>85 ± 1</td>
<td>84 ± 7</td>
</tr>
<tr>
<td>STFT log mag + ph freq-diff</td>
<td>100</td>
<td>100 ms</td>
<td>80 ± 2</td>
<td>79 ± 8</td>
</tr>
<tr>
<td>STFT log mag + ph freq-diff</td>
<td>200</td>
<td>30 ms</td>
<td>86 ± 2</td>
<td>86 ± 5</td>
</tr>
<tr>
<td>STFT log mag + ph freq-diff</td>
<td>200</td>
<td>50 ms</td>
<td>86 ± 2</td>
<td>85 ± 6</td>
</tr>
<tr>
<td>STFT log mag + ph freq-diff</td>
<td>200</td>
<td>100 ms</td>
<td>84 ± 2</td>
<td>84 ± 7</td>
</tr>
</tbody>
</table>
## Deceptively simple

<table>
<thead>
<tr>
<th>1st layer</th>
<th>2nd layer</th>
<th>F-meas Valid</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>30</td>
<td>87±5</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
<td>87±4</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>87±5</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>86±4</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>86±3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>85±5</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>83±4</td>
</tr>
</tbody>
</table>

- Complex network structure does not help
- Very simple structure still gets good performance
- Only one neuron can get most of the performance
Applying machine learning for the onset detection problem is simple and very efficient.

This provides an algorithm that is accurate and robust to a wide variety of songs.

It is not sensitive to hyper-parameter adjustment.
Onset labeling GUI
Results for Different Spectrograms

- Phase acceleration (Bello and Sandlers) is slightly better than noise.
- Phase frequency difference is almost as good as magnitude plane but highly depends on the spectral window width.
- Constant-Q and STFT give the best results, provided the spectral window width is small enough.