
Piano Music Transcription With Fast Convolutional Sparse Coding

Andrea Cogliati*, Zhiyao Duan*, Brendt Wohlberg**

<andrea.cogliati@rochester.edu> * University of Rochester, ** Los Alamos National Labs

Abstract
- Automatic music transcription (AMT) is the process of converting an acoustic musical signal into a symbolic musical representation
- Most existing algorithms for AMT operate in the frequency domain, which introduces the well-known time/frequency resolution trade-off of the Short Time Fourier Transform and its variants
- We propose a time-domain transcription algorithm based on an efficient convolutional sparse coding algorithm in an instrument and environment specific scenario
- The proposed method outperforms a current state-of-the-art AMT method by over 26% in F-measure, achieving a median F-measure of 93.6%

Motivations
- Piano notes are characterized by significant and consistent temporal evolutions in magnitude and phase
- Phase contains important information for grouping partials belonging to the same note
- The best performing single-pitch estimation method works in the time-domain (YIN algorithm)
- High accuracy
- Short training period

Training
- All the individual notes of a piano played at the same dynamic (mf) and for the same duration (1s) are sampled

Transcription
- CBPDN on the audio signal
- Sparse peak picking

Denoising (CBPDN)
- Convolutional Basis Pursuit
- Activation coefficients (i.e., note onsets)
- Sparsity regularization
- Dictionary elements
- Signal to be transcribed

Results

<table>
<thead>
<tr>
<th></th>
<th>Benetos</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure on 30 pieces from the ENSTDkCI collection of MAPS re-rendered with synthetic piano</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.65</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Average results on ENSTDkCI collection of MAPS (no re-rendering)
- Precision: 85.9%
- Recall: 75.8%
- F-measure: 79.7%

Future work
- Estimate note lengths/dynamic with multiple templates per note
- Dictionary adaptation to different instruments

Future Work

Table 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample rate</td>
<td>11,025 Hz</td>
</tr>
<tr>
<td>Initial Number of iterations</td>
<td>500</td>
</tr>
<tr>
<td>Number of iterations</td>
<td>50</td>
</tr>
<tr>
<td>Positive values indicate late estimated onsets.</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4 shows the distribution of the onset difference of the estimated notes from the ground truth, calculated for the true positives:

The figure shows that the results of the frequency-domain method are significantly better than Benetos's method, which only considers the magnitude of the signal. The proposed method shows a dramatic improvement over the frequency domain method in terms of F-measure and other metrics.

Fig. 5 compares the average F-measure achieved by the two methods along the different octaves of a piano keyboard (the first octave is from A0 to B1, the second one from C2 to B2 and so on).

Table 2

<table>
<thead>
<tr>
<th>Octave</th>
<th>Benetos</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>0.58</td>
<td>0.80</td>
</tr>
</tbody>
</table>

Although the comparison with Benetos's method is fair, in the future work, we aim to generalize the results to different pianos and environments.

Fig. 6 shows the distribution of the notes in the ground truth per octave, which are generally about 8 per octave, as illustrated in the following experiments. There is a positive value indicating that the estimated onset is later than the true onset, and vice versa, as demonstrated by the results of the proposed method. It should be noted that, except for templates for the same note played at different dynamics, the results are consistent and significant.

Table 3

<table>
<thead>
<tr>
<th>Note</th>
<th>Benetos</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>0.63</td>
<td>0.72</td>
</tr>
</tbody>
</table>

However, the results are not as good as expected, especially for the proposed method, which only considers temporal evolutions and phase information. In the future work, we aim to improve the results by incorporating more features and information from the signal.
Piano Music Transcription With Fast Convolutional Sparse Coding
Andrea Cogliati*, Zhiyao Duan*, Brendt Wohlberg**
* University of Rochester, ** Los Alamos National Labs

---

**Training**

All the individual notes of a piano played at the same dynamic (mf) and for the same duration (1s) are sampled

**Transcription**

- CBPDN on the audio signal
- Sparse peak picking
- Binarization

---

**Convolutional Basis Pursuit Denoising (CBPDN)**

Activation coefficients (i.e., note onsets) vs. dictionary elements

- The objective function is minimized through the Alternating Direction Method of Multipliers (ADMM)

---

**Parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample rate</td>
<td>11.025 Hz</td>
</tr>
<tr>
<td>λ</td>
<td>0.05</td>
</tr>
<tr>
<td>Number of iterations</td>
<td>3000</td>
</tr>
<tr>
<td>Initial ρ</td>
<td>100A</td>
</tr>
<tr>
<td>Peak picking threshold</td>
<td>10%</td>
</tr>
<tr>
<td>Peak picking sparsity window</td>
<td>50 ms</td>
</tr>
</tbody>
</table>

---

**F-measure per octave**

Fig. 4 shows the distribution of the onset difference of the estimated onsets from the ground truth for true positives.

---

**References**

- Andrea Cogliati and Zhiyao Duan, “Piano music transcription modeling note temporal evolution,” IEEE International Conference of Acoustics, Speech and Signal processing, 2015