Musical Onset Detection by means of Non-Negative Matrix Factorization

GIOVANNI COSTANTINI¹², MASSIMILIANO TODISCO¹, GIOVANNI SAGGIO¹
¹Department of Electronic Engineering
University of Rome “Tor Vergata”
Via del Politecnico, 1 - 00133
ITALY
²Institute of Acoustics “O. M. Corbino”
Via del Fosso del Cavaliere, 100 – 00133 ROMA
ITALY

Abstract: - In this paper, we propose a musical onset detection method, with reference to polyphonic piano music. This method operates on a frame-by-frame basis and exploits a suitable time-frequency representation of the audio signal. The solution proposed consists of an onset detection algorithm based on Short-Time Fourier Transform (STFT) and Non-Negative Matrix Factorization (NMF). To validate this method, we present a collection of experiments using a wide number of musical piano pieces of heterogeneous styles.

Key-Words: - Musical onset, Piano music, Non-Negative Matrix Factorization

1 Introduction
The target of this work dealt with the problem of extracting musical content from audio data, particularly with reference to polyphonic piano music and concentrates on the issue of onset detection.
Onsets are punctual temporal events that correspond to the start of a musical note. These onsets are expected to emphasize the important moments of a melody and the music's beats.
Since the music signal is event-based, segmenting it into individual note events greatly facilitate editing and analysis of audio.
Onset detection has been studied in computer music in order to aid tasks such as automatic music transcription, tempo and beat tracking, audio signals segmentation etc.
Researches on onset detection have been done in areas of beat and tempo tracking such as in [1][2]. In [3] has presented a system that performs onset detection in the complex domain, in contrast with traditional onset detectors that consider only the energy of the signal.
Another approach for detecting onsets of piano notes was proposed in [4], which is based in a gammatone filterbank with output channels centered at piano tones, and a smoothing filter that works together with a peak-picking step.
Other recent approaches of onset detection methods can be found in [5, 6].
The approach for onset detection based on Short-Time Fourier Transform (STFT) and Non-Negative Matrix Factorization (NMF) is proposed in this work.
After the seminal paper of Lee and Seung [7], non-negative matrix factorization (NMF) attracted the interest of many researchers in diverse fields, e.g. text mining [8], document clustering [9], image recognition [10].

NMF was first proposed by Paatero [11] and consists in finding reduced rank non-negative factors to approximate a given non-negative data matrix.
In the framework of music transcription, NMF was investigated by Smaragdis and Brown [12], with the aim of decompose polyphonic music into single notes.
The rest of the paper is organized as follows. An overview of the NMF method is presented in Sec. 2. The onset detection algorithm is then discussed in Sec. 3. Experimental results are conducted in Sec. 4. Concluding remarks are given in Sec. 5.

2 NMF
The problem addressed by NMF [13] is as follows: given a non-negative \( n \times m \) matrix \( D \), it is possible to find non-negative matrix factors \( W \) and \( H \) in order to approximate the original matrix:

\[
D = W \ast H
\]  

where the \( n \times r \) matrix \( W \) contains the basis vectors and the \( r \times m \) matrix \( H \) contains the weights needed to properly approximate the corresponding column of matrix \( D \), as a linear combination of the columns of \( W \). Usually, \( r \) is chosen so that \( (n + m) r < nm \), thus resulting in a compressed version of the original data matrix.
Usually the elements of \( W \) and \( H \) are estimated by minimizing the cost function

\[
C = \frac{1}{2} \| D - W \ast H \|_F^2
\]  

where \( \| \cdot \|_F \) is the Frobenius norm and the cost function is minimized by using update rules, which are given as
\[ W \leftarrow W \cdot (DH^T) \cdot (WHH^T + 10^{-9}) \]
\[ H \leftarrow H \cdot (W^T D) \cdot (W^T WH + 10^{-9}) \]

where \( \cdot \) and \( / \) are element-wise multiplication and division, respectively. \( W \) and \( H \) are initialized with absolute value of random noise, and alternatively updated by rules (3) until the cost function does not significantly change.

3 Onset Detection Algorithm

We now illustrate the method proposed by Wang et al [14], to detect note onsets. This method is based on Short Time Fourier Transform (STFT) and Non-Negative Matrix Factorization (NMF).

The method [14] has been implemented as follows: let us consider a discrete time-domain signal \( s(nh) \), whose STFT is given by

\[ S_k(mh) = \sum_{n=nh}^{mh+N-1} w(n-mh)s(n)e^{-j\Omega_k(n-mh)} \]

(4)

where \( N \) is the window size, \( h \) is the hop size, \( m \in \{0, 1, 2,...,M\} \) the hop number, \( k = 0, 1,...,N - 1 \) is the frequency bin index, \( w(n) \) is a finite-length sliding hanning window and \( n \) is the summation variable.

We obtain a time-frequency representation of the audio signal by means of spectral frames represented by the magnitude spectrum \( S_k(mh) \). The set of all the \( S_k(mh) \) can be packed as columns into a non-negative matrix \( D_{L \times M}^{L \times M}(f,t) \), where \( M \) are the total number spectra we computed and \( L=N/2 \) is the number of their frequencies.

After that, NMF is used to separate the magnitude spectrogram \( D_{L \times M}^{L \times M}(f,t) \) of each frame \( m \) into a product of non-negative spectrum \( W_{L \times R}^{L \times R}(f) \) and a non-negative time-varying gain \( H_{R \times M}^R(t) \), so that

\[ D_{L \times M}^{L \times M} = W_{L \times R}^{L \times R} \cdot H_{R \times M}^R \]

(5)

where \( R \) is the rank of the factorization.

At last, we estimate the onset detection function by summing the rows of \( H \)

\[ h(m) = \sum_{r=1}^{R} H(r,m) \]

(6)

Relative differentiation of \( h(.) \) gives the onset detection function \([10]\)

\[ f(m) = \frac{h(m) - h(m-1)}{h(m)} \]

(7)

whose peaks are intended to coincide with the times of note onsets.

4 Audio Dataset and Experimental Results

In this section we report on the simulation results concerning the onset detection methods illustrated in Section 2. The MIDI data used in the experiments were collected from the Classical Piano MIDI Page,
A list of pieces can be found in Table I.

All the audio files have a sampling rate of 8 kHz. Among the 124 pieces, 24 testing and 13 validation pieces were selected. The first minute from each song was selected for experiments, providing a total of 24 minutes of test audio and 13 minutes of audio for threshold tuning (validation set).

This amounted to 6142 and 3406 note onsets in the test and validation sets, respectively.

The statistical evaluation of the performance of the onset detection methods is summarised by three statistics, i.e. the precision P, the recall R and the F-measure, which are defined as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{7}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{8}
\]

\[
\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}
\]

In the above formulas TP is the number of correct detections, FP is the number of false positives and FN is the number of false negatives. Precision represents the percentage of correct positive predictions in the identification of a note onset. Recall represents the capacity of the onset detector to identify the positive examples. The global variable F-measure is the harmonic mean of Precision and Recall.

A correct positive prediction happens if the system detects the onset within a specified time interval (tolerance) with

<table>
<thead>
<tr>
<th>Composer</th>
<th>Training</th>
<th>Testing</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albéniz</td>
<td>España (Prélude, Malagueña, Sereneta, Zortzico) Suite Española (Granada, Cataluña, Sevilla, Cádiz, Aragon, Castilla)</td>
<td>España (Tango) Suite Española (Cuba)</td>
<td>España (Capricho Catalan)</td>
</tr>
<tr>
<td>Bach</td>
<td>BWV 850</td>
<td>BWV 847</td>
<td>BWV 846</td>
</tr>
<tr>
<td>Balakirew</td>
<td>Islame</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Beethoven</td>
<td>Appassionata (1-3) Moonlight (1,3) Pathétique (1) Waldstein (1-3)</td>
<td>Für Elise Moonlight (2) Pathétique (3)</td>
<td>Pathétique (2)</td>
</tr>
<tr>
<td>Borodin</td>
<td>Petite Suite (In the monastery, Intermezzo, Mazurka, Serenade, Nocturne)</td>
<td>Petite Suite (Mazurka)</td>
<td>Révérie</td>
</tr>
<tr>
<td>Brahms</td>
<td>Fantasia (2,5) Rhapsodie</td>
<td>Fantasia (6)</td>
<td>-</td>
</tr>
<tr>
<td>Burgmueller</td>
<td>The pearls Thunderstorm</td>
<td>The Fountain</td>
<td>-</td>
</tr>
<tr>
<td>Chopin</td>
<td>Opus 7 (1,2) Opus 25 (4) Opus 28 (2, 6, 10, 22) Opus 33 (2,4)</td>
<td>Opus 10 (1) Opus 28 (13)</td>
<td>Opus 28 (3)</td>
</tr>
<tr>
<td>Debussy</td>
<td>Suite Bergamasque (Passepied, Prélude)</td>
<td>Menuet</td>
<td>Clair de Lune</td>
</tr>
<tr>
<td>Granados</td>
<td>Danzas Españolas (Oriental, Zarabanda)</td>
<td>Danzas Españolas (Villanesca)</td>
<td>-</td>
</tr>
<tr>
<td>Grieg</td>
<td>Opus 12 (3) Opus 43 (4) Opus 71 (3)</td>
<td>Opus 65 (Wedding)</td>
<td>Opus 54 (3)</td>
</tr>
<tr>
<td>Haydn</td>
<td>Piano Sonata in G major (1)</td>
<td>Piano Sonata in G major (2)</td>
<td>-</td>
</tr>
<tr>
<td>Liszt</td>
<td>Grandes Études de Paganini (1-5)</td>
<td>Love Dreams (3)</td>
<td>Grandes Études de Paganini (6)</td>
</tr>
<tr>
<td>Mendelssohn</td>
<td>Opus 30 (1) Opus 62 (3,4)</td>
<td>Opus 62 (5)</td>
<td>Opus 53 (5)</td>
</tr>
<tr>
<td>Mozart</td>
<td>KV 330 (1-3) KV 333 (3)</td>
<td>KV 333 (1)</td>
<td>KV 333 (2)</td>
</tr>
<tr>
<td>Mussorgsky</td>
<td>Pictures at an Exhibition (1,3,5-8)</td>
<td>Pictures at an Exhibition (2,4)</td>
<td>-</td>
</tr>
<tr>
<td>Schubert</td>
<td>D 784 (1,2) D 760 (1-3) D 960 (1,3)</td>
<td>D 760 (4)</td>
<td>D 960 (2)</td>
</tr>
<tr>
<td>Schumann</td>
<td>Scenes from Childhood (1-3,5,6)</td>
<td>Scenes from Childhood (4)</td>
<td>Opus 1 (1)</td>
</tr>
<tr>
<td>Tchaikovsky</td>
<td>The Seasons (February, March, April, May, August, September, October, November, December)</td>
<td>The Seasons (January, June)</td>
<td>The Seasons (July)</td>
</tr>
</tbody>
</table>

Table I

http://www.piano-midi.de/. A list of pieces can be found in Table I.

All the audio files have a sampling rate of 8 kHz. Among the 124 pieces, 24 testing and 13 validation pieces were selected. The first minute from each song was selected for experiments, providing a total of 24 minutes of test audio and 13 minutes of audio for threshold tuning (validation set).

This amounted to 6142 and 3406 note onsets in the test and validation sets, respectively.

The statistical evaluation of the performance of the onset detection methods is summarised by three statistics, i.e. the precision P, the recall R and the F-measure, which are defined as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{7}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{8}
\]

\[
\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}
\]

In the above formulas TP is the number of correct detections, FP is the number of false positives and FN is the number of false negatives. Precision represents the percentage of correct positive predictions in the identification of a note onset. Recall represents the capacity of the onset detector to identify the positive examples. The global variable F-measure is the harmonic mean of Precision and Recall.

A correct positive prediction happens if the system detects the onset within a specified time interval (tolerance) with
respect to the ground-truth onset. The tolerance values used in the experiments is 32 ms. The threshold values have been computed through maximization of the F-measure value on the validation dataset. In Tables II we present the results obtained with method [14], using 1024 points FFT, with various window length and hop size. Since the simulation results presented in [14] were obtained using a different sampling rate (22050 Hz), we cannot use the same parameters adopted in [14] (4096 points FFT; window size N = 400; hop size Δ = 200). As concerns the factorization rank R, we tried different values and the best results have been obtained using R = 1 (Tables I-II).

<table>
<thead>
<tr>
<th>Window size</th>
<th>Hop size</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>128</td>
<td>51.0%</td>
<td>59.7%</td>
<td>54.9%</td>
</tr>
<tr>
<td>512</td>
<td>256</td>
<td>96.7%</td>
<td>88.8%</td>
<td>92.5%</td>
</tr>
<tr>
<td>1024</td>
<td>512</td>
<td>82.1%</td>
<td>65.7%</td>
<td>73.6%</td>
</tr>
</tbody>
</table>

4 Conclusion

In this paper, we have implemented and tested a musical onset detection method, proposed by Wang et al., based on non-negative decomposition of a magnitude spectrum matrix.

A wide number of musical pieces of heterogeneous styles we used to validate and test the onset detection method. It has been shown that onset detection algorithm, realized with the aid of STFT and the NMF, is helpful in the determination of note attacks with very modest computational cost and good accuracy.

References: