A New Method for Musical Onset Detection in Polyphonic Piano Music

GIOVANNI COSTANTINI¹,², 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. The solution proposed consists of an onset detection algorithm based on Short-Time Fourier Transform (STFT) and Non-Negative Matrix Factorization (NMF). This method operates on a frame-by-frame basis and exploits a suitable binary time-frequency representation of the audio signal. 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]. In [2] 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 [3], 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 [4, 5].
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 [6], non-negative matrix factorization (NMF) attracted the interest of many researchers in diverse fields, e.g. text mining [7], document clustering [8], image recognition [9].
NMF was first proposed by Paatero [10] 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 [11], 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 is as follows: given a non-negative n×m matrix D, it is possible to find non-negative matrix factors W and H in order to approximate the original matrix:

\[ D \approx W \cdot H \] (1)

where the n × r matrix W contains the basis vectors and the r × 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

\[ C = \frac{1}{2} \| D - WH \|_F^2 \] (2)

where \( \| \cdot \|_F \) is the Frobenius norm and the cost function is minimized by using update rules, which are given as

\[ W \leftarrow W \cdot (DH^T) / (WHH^T + 10^{-9}) \]
\[ H \leftarrow H \cdot (W^TD) / (W^THW + 10^{-9}) \] (3)
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 new method to detect note onsets. This method is based on Short Time Fourier Transform (STFT) combined with a suitable binary processing and Non-Negative Matrix Factorization (NMF).

Let us consider a discrete time-domain signal \( s(mh) \), whose STFT is given by

\[
S_k(m) = \sum_{n=0}^{N-1} w(n-mh)s(n)e^{-j2\pi k(n-mh)}
\]

where \( N \) is the window size, \( h \) is the hop size, \( m \in \{0, 1, 2,...,M\} \) the hop number, \( k \in 0, 1, ..., N-1 \) 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 \( |S_k(m)| \).

The set of all the \( |S_k(m)| \) can be packed as columns into a non-negative matrix \( D^{L \times M} (f,t) \), where \( M \) are the total number spectra we computed and \( L=N/2 \) is the number of their frequencies. After normalization in the range from 0 to 1 we obtain matrix \( D \in [0,1]^{L \times M} \). Then we perform a binarization of \( D \) giving the binary matrix \( \overline{D} \in \{0,1\}^{L \times M} \):

\[
\overline{D}(l,m) = \begin{cases} 1 & \text{if } D(l,m) > T_1 \\ 0 & \text{if } D(l,m) \leq T_1 \end{cases}
\]

where \( T_1 \) is a threshold. Thresholding is used to avoid the effect of spectral noise and to allow some simple ‘spatial’ operations, as illustrated in the following. The threshold value can be obtained using a suitable validation set of examples, as explained in Section 4. An example of the first two processing steps is shown in Figure 1.

To improve the accuracy of onset detection, we perform two further operations on the binary spectrogram \( \overline{D} \). First, we set to ‘0’ the \( (l, m) \) element of \( \overline{D} \) if the previous adjacent cell \( (l, (m-1)) \), relative to the previous frame, is equal to ‘1’. This operation is adopted to point out only the spectral changes in the time-frequency representation. The result of this processing step is a new binary matrix \( \overline{D} \in \{0,1\}^{L \times M} \):

\[
\overline{D}(l,m) = \begin{cases} 0 & \text{if } \overline{D}(l,m-1) = 1 \\ \overline{D}(l,m) & \text{otherwise} \end{cases}
\]

Then we set to ‘0’ the \( (l, m) \) element of \( \overline{D} \) if both the previous adjacent cell \( (l-1, m) \), relative to the previous frequency bin, and the subsequent adjacent cell \( ((l+1), m) \), relative to the subsequent frequency bin, are equal to ‘0’. This operation is adopted to remove isolated spectral bins in the time-frequency representation. The result of this processing step is a new binary matrix \( \overline{D} \in \{0,1\}^{L \times M} \):

\[
\overline{D}(l,m) = \begin{cases} 0 & \text{if } \overline{D}(l-1,m) = 0 \land \overline{D}(l+1,m) = 0 \\ \overline{D}(l,m) & \text{otherwise} \end{cases}
\]

Operations (6) and (7) are shown in Figure 2.

![Figure 2. Results of operations (6) and (7)](image)

After that, NMF is used to separate the magnitude spectrogram \( \overline{D}^{L \times M} (f,t) \) of each frame \( m \) into a product of non-negative spectrum \( W^{L \times 1} (f) \) and a non-negative time-varying gain \( H^{1 \times M} (t) \), so that

\[
\overline{D}^{L \times M} = W^{L \times 1} \otimes H^{1 \times M}
\]

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

At last, the onset detection function, whose peaks are intended to coincide with the times of note onsets, corresponds to the matrix \( H^{1 \times M} \).

We will demonstrate some results of the onset detection method from real piano polyphonic music of Mozart’s KV
Sonata in B-flat Major, Movement 3, sampled at 8 KHz. We will consider the second and third bar at 120 bpm. Figure 3 shows the Mozart’s Sonata.

![Mozart's Sonata](image)

Figure 3: Mozart's Sonata

We use a STFT with $N=512$, an N-point Hamming window and a hop size $h=256$ that means a 32 millisecond hop between successive frames. The following Figure 4 shows the spectrogram.

![Spectrogram of Mozart's Sonata](image)

Figure 4: Spectrogram of Mozart's Sonata

We apply now the first processing step from which we obtain the normal spectrogram. It is shown in the following figure.

![Normal Spectrogram](image)

NMF separation method is used now to the latter matrix $\hat{D}^{L \times M}$. The peaks of the vector $H^{L \times M}$ we have obtained represent the onset detection function with a 32ms time onset resolution. It is display in Figure 5.

![Onset Detection Function](image)

Figure 5: Onset detection function

The last two processing steps (6) (7) are applied to point up the spectral variance and to remove isolate spectral bin. Results are shown in the following two figures.
4 Audio Dataset and Experimental Results

In this section we report on the simulation results concerning the onset detection methods illustrated in Section 3. The MIDI data used in the experiments were collected from the Classical Piano MIDI Page, http://www.piano-midi.de/. A list of pieces can be found in [13], (p. 8, Table 5). 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 results are summarized by three statistics: the \textit{precision}, \textit{recall} and \textit{F-measure}, which are given by

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{9}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{10}
\]

\[
\text{F-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}
\]

where \(TP\) is the number of correct detections, \(FP\) is the number of false positives and \(FN\) is the number of false negatives. \textit{Precision} represents the percentage of precision in the identification of an example as positive. \textit{Recall} represents the capacity of the onset detector for the identification of the greatest number of positive examples. The threshold \(T_h\) has been computed through maximization of the \textit{F-measure} value on the validation dataset. The results of this test are shown in the following table

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Precision}</td>
<td>98.7%</td>
</tr>
<tr>
<td>\textit{Recall}</td>
<td>98.5%</td>
</tr>
<tr>
<td>\textit{F-measure}</td>
<td>98.6%</td>
</tr>
</tbody>
</table>

4 Conclusion

In this paper, we have developed and tested a musical onset detection method 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 combined with a suitable binary processing and the NMF, is helpful in the determination of note attacks with very modest computational cost and good accuracy.

References:


