Non-negative matrix factorization for note onset detection of audio signals

This item was submitted to Loughborough University's Institutional Repository by the/an author.


Additional Information:

- This is a conference paper [© IEEE]. It is also available at: http://ieeexplore.ieee.org/ Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Metadata Record: https://dspace.lboro.ac.uk/2134/5563

Version: Published

Publisher: © IEEE

Please cite the published version.
This item was submitted to Loughborough’s Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/
NON-NEGATIVE MATRIX FACTORIZATION FOR NOTE ONSET DETECTION OF AUDIO SIGNALS

Wenwu Wang†, Yuhui Luo‡, Jonathon A. Chambers¶, and Saeid Sanei†

† Tao Group Limited, Reading, RG6 1AZ, U.K.
Email: wenwu.wang@ieee.org
‡ Samsung Electronics Research Institute, Staines, TW18 4QE, U.K.
Email: effy.yuhui.luo@gmail.com
¶ Cardiff University, Cardiff, CF24 3AA, U.K.
Emails: [chambersj, saneis]@cf.ac.uk

ABSTRACT

A novel approach using non-negative matrix factorization (NMF) for onset detection of musical notes from audio signals is presented. Unlike most commonly used conventional approaches, the proposed method exploits a new detection function constructed from the linear temporal bases that are obtained from a non-negative matrix decomposition of musical spectra. Both first-order difference and psychoacoustically motivated relative difference functions of the temporal profile are considered. As the approach works directly on input data, no prior knowledge or statistical information is thereby required. A practical issue of the choice of the factorization rank is also examined experimentally. Numerical examples are provided to show the performance of the proposed method.

1. INTRODUCTION

The aim of onset detection is to locate the starting point of a noticeable change in intensity, pitch or timbre of sound. It plays an important role in a number of music applications, such as automatic transcription, content delivery, synthesis, indexing, retrieval and low bit-rate audio coding. Due to several major difficulties, e.g., identifying changes in different notes with wide range of temporal dynamics, distinguishing vibrato from changes in timbre, detecting fast passages of musical audio, and extracting onsets generated by different instruments, onset detection remains an open problem demanding further research effort.

A variety of approaches have been proposed in the literature, with most of them sharing an approximately common procedure, as depicted in Fig.1(a). A musical audio track may be initially pre-processed to remove the undesired noises and fluctuations. Then, a so-called detection function is formed from the enhanced signal, such that the occurrence of a note is made more distinguishable as compared with the steady-state of note transition. Finally, the locations of onsets are determined by a peak-picking algorithm [1]. Undoubtedly, the detection function is of great importance to the overall performance of an onset detection algorithm. For the onsets to be easily detected, a good detection function should reveal sharp peaks at the locations of those onsets, which would effectively facilitate the subsequent peak-picking process.

Although similar concepts relevant to human perception have been used in most existing approaches to detect onset changes, they are essentially very distinctive as regards to the various information of signals being employed in the construction of detection functions. These include the intensity change based methods using temporal features, e.g. [2] [3]; the timbre change based methods using spectral features, e.g. [4]; model based detection methods using statistical properties, e.g. [5], and methods based on phase and

![Fig. 1. Diagram of the onset detection: (a) the general scheme, (b) the proposed reduction strategy, i.e., the scheme for deriving the detection functions in this work.](image-url)
pitch information of signals, e.g. [6] [7], among many others (see e.g. [1] for a recent review and more references therein).

In this paper, we propose a novel approach for onset detection. This approach is essentially based on the representation of audio content of the musical passages by a linear basis transform, and the construction of the detection function from the bases learned by non-negative decomposition of the musical spectra. The overall detection scheme is shown in Fig. 1(b). In this scheme, musical magnitude (or power) spectra of the input data are firstly generated using a discrete Fourier transform (DFT). Then, the non-negative matrix factorization (NMF) algorithm is applied to find the crucial features in the spectral data. With the transformed data, the individual temporal bases are exploited to reconstruct an overall temporal feature function of the original signal. The detection function is thereby derived by taking the first-order difference (or relative difference) of the feature function whose sudden bursts are converted into narrower peaks for easier detection.

The proposed approach has several promising properties. First of all, the proposed technique is a data-driven approach, no prior information is needed, as otherwise required for many knowledge based approaches. Secondly, the algorithm works directly on the original data, hence it avoids a frequently used pre-processing stage in some state-of-the-art approaches. Additionally, thanks to the temporal features obtained implicitly from the NMF decomposition, explicit computation of the signal envelope or energy feature, which is required for many existing intensity based detection approaches, is no longer necessary. Moreover, the NMF based temporal feature is more robust for both first-order difference and relative difference as compared with direct envelope detection based approaches (more will be demonstrated in the subsequent simulation section).

The remainder of this paper is organized as follows. The concept of NMF and the algorithm used in this work are briefly reviewed in Section 2. The method for generating the non-negative spectral matrix from the input data is presented in Section 3, where the method of how to apply the NMF learning algorithm is also included. The proposed detection functions based on respectively the first-order difference, the relative difference, and a constant-balanced relative difference, are described in Section 4. Section 5 is dedicated to the experimental verification of the proposed approach. Finally, the paper is summarized in Section 6.

2. NMF

NMF is an emerging technique for data analysis that was proposed recently [8] [9]. Given an $M \times N$ non-negative matrix $X \in \mathbb{R}^{\geq 0, M \times N}$, the goal of NMF is to find nonnegative matrices $W \in \mathbb{R}^{\geq 0, M \times R}$ and $H \in \mathbb{R}^{\geq 0, R \times N}$, such that

$$X \approx WH$$

(1)

where $R$ is the rank of the factorization, generally chosen to be smaller than $M$ (or $N$), or akin to $(M + N)R < MN$, which results in the extraction of some latent features whilst reducing some redundancies in the original data. To find the optimal choice of matrices $W$ and $H$, we should minimize the reconstruction error between $X$ and $WH$. Several error functions have been proposed for this purpose [8]-[11]. For instance, an appropriate choice is to use the criterion based on the Euclidean distance,

$$\begin{align*}
(W, H) &= \arg \min_{W,H} \| X - WH \|^2_F \\
&= (\hat{W}, \hat{H})
\end{align*}$$

(2)

where $\hat{W}$ and $\hat{H}$ are the estimated optimal values of $W$ and $H$, and $\| \cdot \|_F$ denotes the Frobenius norm. Alternatively, we can also minimize the error function based on the extended Kullback-Leibler divergence,

$$\begin{align*}
(W, H) &= \arg \min_{W,H} \sum_{m=1}^{M} \sum_{n=1}^{N} D_{mn} \\
&= (\hat{W}, \hat{H})
\end{align*}$$

(3)

where $D_{mn}$ is the $mn$-th element of the matrix $D$ which is given by

$$D = X \odot \log(X \odot (WH)) - X + WH$$

(4)

where $\odot$ and $\odot$ denote the Hadamard (element-wise) product and division respectively, i.e., each entry of the resultant matrix is a product and division of the corresponding entries from two individual matrices respectively. Although gradient decent and conjugate gradient approaches can both be applied to minimize these cost functions, we are particularly interested in the multiplicative rules developed by Lee and Seung [9] [10]. In compact form, the multiplicative update rules for minimizing criterion (2) can be re-written as

$$\begin{align*}
H &\leftarrow H \odot \left( W^T X \right) \odot \left( W^T WH \right) \\
W &\leftarrow W \odot \left( XH^T \right) \odot \left( WWH^T \right)
\end{align*}$$

(5)

(6)

where $\cdot^T$ is the matrix transpose operator, and $\leftarrow$ denotes iterative evaluation. Comparatively, these rules are easy to implement and also have good convergence performance. Additionally, a step size parameter which is normally required for gradient algorithms, is not necessary in these rules. The iteration of these update rules is guaranteed to converge to a locally optimal matrix factorization [10]. The rules (5) and (6) are used in the following analysis of our work.

3. NON-NEGATIVE DECOMPOSITION OF MUSICAL SPECTRA

For the NMF algorithm to be applied, we should first prepare a non-negative matrix that contains an appropriate representation of the original data to be analyzed. Unlike the
image data analyzed in [9], musical audio data can not be
directly used as they contain negative-valued samples. In
our problem, the non-negative matrix $X$ is generated as the
magnitude spectra of the input data, similar to [13]. We
denote the original audio signal as $s(t)$, where $t$ is the time
instant. Using a $T$-point windowed DFT, a time-domain
signal $s(t)$ can be converted into a frequency-domain time-
series signal as

$$S(f, k) = \sum_{\tau=0}^{T-1} s(k\delta + \tau) w(\tau) e^{-j2\pi f\tau/T}$$

(7)

where $w(\tau)$ denotes a $T$-point window function, $j = \sqrt{-1}$,
$\delta$ is the time shift between the adjacent windows, and $f$
is a frequency index, $f = 0, 1, \ldots, T - 1$. Clearly, the
time index $k$ in $S(f, k)$ is generally not a one-to-one mapping
to the time index $t$ in $s(t)$. If the whole signal has, for instance,
$L$ samples, then the maximum value of $k$, i.e. $K$, is given
as $K = [(L - T)/\delta]$, where $[\cdot]$ is an operator taking
the maximum integer no greater than its argument1. Let $\hat{S}(f, k)$
be the absolute value of $S(f, k)$, we can then generate the
following non-negative matrix by packing $\hat{S}(f, k)$ together,

$$X = \begin{pmatrix}
\hat{S}(0, 0) & \hat{S}(0, 1) & \cdots & \hat{S}(0, K - 1) \\
\hat{S}(1, 0) & \hat{S}(1, 1) & \cdots & \hat{S}(1, K - 1) \\
\vdots & \vdots & \ddots & \vdots \\
\hat{S}(T/2, 0) & \hat{S}(T/2, 1) & \cdots & \hat{S}(T/2, K - 1)
\end{pmatrix}$$

(8)

where only half frequency bins (from 0 to $T/2 + 1$) are used
since the magnitude spectra are symmetrical along the fre-
quency axis, and the dimension of $X$, i.e. $M \times N$, then
becomes $(T/2 + 1) \times K$. This non-negative matrix contain-
ing the magnitude spectra of the input signal will be used
for decomposition.

Using the learning rules (5) and (6), $X$ in (8) can be
effectively decomposed into the product of two non-negative
matrices, denoted as $W^o \in \mathbb{R}_{\geq 0}^{(T/2+1) \times R}$ and $H^o \in
\mathbb{R}_{\geq 0}^{R \times K}$, i.e., the corresponding local optimum values
of $W$ and $H$ respectively, which are obtained when the learn-
ing algorithm converges. An advantage of exploiting spec-
tral matrix (8) is that both the obtained basis matrices $W^o$
and $H^o$ have meaningful interpretation. That is, $H^o$ is a
dimension-reduced matrix which contains the bases of the
temporal patterns while $W^o$ contains the frequency patterns
of the original data. For musical audio, these patterns can
be interpreted as the time-frequency features of individual
notes as the NMF learns a parts-based representation of $X$
[9]. It is worth noting that whether the learned parts re-
veal the true (very often latent) patterns of the input data
depends on the choice of $R$, for which, there has been no
generic guidance for different application scenarios. How-
ever, this issue turns out not to be crucial in our application,
as verified in our simulations.

4. CONSTRUCTION OF DETECTION FUNCTIONS

By combining all the single parts together, we can recon-
struct the following time series

$$h^o(k) = \sum_{r=1}^{R} H^o_{r,k}$$

(9)

where $k = 0, \cdots, K - 1$. By simulation, we found that,
$h^o(k)$ actually contains a good approximation of the over-
all temporal profile (envelope) of the original signal. (As a
result, the column vector $h^o = [h^o_0, \cdots, h^o_{K-1}]^T \in \mathbb{R}_{\geq 0}^{K}$
describes the temporal profile of the original signal.) Therefore,
hence it is able to reveal sudden intensity changes in the signal.
However, there exists psychoacoustic evidence showing that a human’s hearing is generally
more sensitive to the relative than to the absolute intensity
changes [12]. Therefore, we can also use a detection func-
tion based on the first-order relative difference, that is

$$h^r(k) = \frac{d}{dk} h^o(k), \quad k = 0, \cdots, K - 1$$

(10)

where $\frac{d}{dk}$ is a difference operator for discrete series (taken
from its continuous counterpart derivative). This function
takes the absolute difference between the neighbouring
samples of $h^o(k)$, hence it is able to reveal sudden intensity
changes in the signal. Note that, the major difference between
$h^r(k)$ in (11) and the detection function proposed by Klupuri [3]
lies in the different strategies taken for the construction of the temporal
profile. In [3], it is formed directly from the energy or amplitude envelope of the original signal.

To consider a trade-off between the performance by the above two functions, we also introduce a constant-balanced
detection function,

$$h^b(k) = \frac{\frac{d}{dk} h^o(k)}{\eta + h^o(k)}$$

(12)

where $\eta$ is a positive constant. By adjusting the constant
$\eta$, we can obtain the desirable performance in the interim
that may be achieved by (10) and (11) independently. To see
this, we consider two extreme cases. If $\eta$ takes values

---

1In practice, zero-padding may be required to allow the remaining $p$
($0 \leq p < \delta$) samples in the end of the signal to be covered by the analysis
window.
approaching to zero, i.e. \( \eta \to 0 \), in other words, \( \eta \ll \omega^o(k) \),
we have \( \hat{h}^o(k) \approx \hat{h}^o(k) \). On the other hand, if \( \eta \gg \omega^o(k) \),
we have \( \hat{h}^o(k) \approx (1/\eta)\hat{h}^o(k) \), which means \( \hat{h}^o(k) \) will have
the same profile as that of \( \hat{h}^o(k) \), with the only difference of a scaling factor. All the above three detection functions
are examined in our simulations. In fact, \( \eta \) has practical advantage of preventing the denominator in (11) being zero.
Effectively, (12) can also be written as the logarithm,
\[
\hat{h}^o(k) = \frac{d}{dk} \log(\eta + \hat{h}^o(k))
\]
which is a psychoacoustic-relevant expression implying that
a human’s hearing ability is not perceptually equal over the
intensity changes in the sound signal.

5. NUMERICAL EXPERIMENTS

5.1. Detection example for a percussive audio signal

To illustrate the detection method described above, we first
apply the proposed approach to the onset detection of a simple
audio signal which was played by a violin and contains three consecutive music notes G4, A3 and E5 (see
Fig. 2 (a)), whose note numbers are 55, 45 and 64 respectively, and whose frequencies are 196.01Hz, 110.00Hz, and
329.6Hz respectively\(^2\). The sampling frequency \( f_s \) for this

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig2.png}
\caption{The waveform of the original audio signal (a) and the generated non-negative magnitude spectrum matrix \( \mathbf{X} \) (b). The onset locations are marked manually with arrows.}
\end{figure}

signal is 22050Hz. The whole signal has \( L = 149800 \) samples with an approximate length of 6794ms. This signal is transformed into the frequency domain by the procedure described in Section 3, where the frame length \( T \) of the Fast

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig3.png}
\caption{Detection results of the signal depicted in Fig. 2. (a)-(c) are the visualizations of row vectors of the matrix \( \mathbf{H}^o \); (d) denotes the temporal profile of \( h^o(k) \), i.e., eqn. (9); (e) visualizes the detection function (13); and (f) represents the final onset locations.}
\end{figure}

Fourier transform (FFT) is set to 4096 samples, i.e., the frequency resolution is approximately 5.4Hz. The signal is segmented by a Hamming window with the window size being set to 400 samples (approximately 18ms), and the time shift \( \delta \) to 200 samples (approximately 9ms), that is, a half-window overlap between the neighboring frames is used. Note that, the choice of the window size is slightly different from that in (7), for which the window size is identical to FFT frame length \( T \). The small size of the signal segments is chosen to guarantee a sufficient time resolution, and each segment is then zero-padded to have the same size as \( T \) for FFT operation. The factorization rank \( R \) is set to 3, i.e., exactly the same as the total number of the notes in the signal. The matrices \( \mathbf{W} \) and \( \mathbf{H} \) were initialized as absolute values of two random matrices. The NMF algorithm was running 100 iterations. In fact, the algorithms only took 11 iterations to converge to a local minimum in this experiment. The generated non-negative magnitude spectrum matrix \( \mathbf{X} \) is visualized in Fig. 2(b). Fig. 3 demonstrates the process
described in Section 3 and 4 (see also Fig. 1 (b)), where the detection function (13) was applied, and the constant \( \eta \) is set to 0.01. From Fig. 3 (a)-(c), it is clear that the NMF algorithm has learned the parts of the original signal, and these three parts represent the individual notes in this case. By summing these three parts using Eqn. (9), the overall temporal profile \( h^o(k) \) of the original signal is reconstructed, as shown in Fig. 3 (d). After applying Eqn. (13) to this profile, the detection function \( \hat{h}^o(k) \) reveals apparent peaks on the locations where the notes start to attack, see Fig. 3 (e). The onset locations can thereby be easily determined by thresholding the local maxima of \( \hat{h}^o(k) \), see Fig. 3 (f), which are 630ms, 3016ms and 5574ms respectively.

\(^2\)The MIDI specification only defines note number 60 as “Middle C”, and all other notes are relative. The absolute octave number designations can be arbitrarily assigned. Here, we define “Middle C” as C5.
5.2. On choice of factorization rank $R$

The rank $R$ was chosen to be 3 in the above experiment, as we know exactly how many latent parts are contained in this case. In many practical situations, however, the number of hidden parts are not known a priori. Either a greater or a smaller value of $R$ than the real number of the latent parts in the signal to be learned may be used for the factorization. Unfortunately, there is no generic guidance on how to choose optimally the rank $R$. Here, we show experimentally the effect of $R$ on the performance of our detection method. We use the same experimental set-up for the parameters as above, except for $R$, which we change from 1 to 5. Fig. 4 and Fig. 5, are the visualizations of matrix $H^o$ with $R$ equal to 2 and 4 respectively. Fig. 4 (b) indicates that the total parts have not been fully separated, as there are two parts bound together in one row. Fig. 5 shows that although all parts have been separated as shown in (a) (c) and (d), there is an extra row that may contain the weighted components of all latent parts. Fortunately, these side effects are not crucial in our application. Fig. 6 plots $h^o(k)$ changing with various

![Graph](image1)

**Fig. 4.** The visualization of row vectors of the matrix $H^o$ for rank $R = 2$.

![Graph](image2)

**Fig. 5.** The visualization of row vectors of the matrix $H^o$ for rank $R = 4$.

and $R$. We can see clearly that the profiles are very similar for different $R$ and only differ from their amplitude, especially the change points of the intensity remain the same for different $R$. This implies that various $R$ still gives the same detection result.

5.3. Comparisons with RMS approach

In this section, we compare the proposed approach with the approach based on the direct detection of the signal envelope using the root-mean-square (RMS), i.e.

$$h^{RMS}(k) = \sqrt{\frac{1}{T} \sum_{\tau=0}^{T-1} [s(k \delta + \tau)]^2}$$

(14)

where $\delta$ is the time shift, $k$ denotes the frame index, and $T$ is the frame length. For simplicity, the detection functions derived from (14), corresponding to those described by Eqn. (10), (11) and (13) respectively in Section 4, are denoted as $h^{RMS}_a(k)$, $h^{RMS}_o(k)$, and $h^{RMS}_b(k)$ respectively, which are obtained simply by replacing $h^o(k)$ with $h^{RMS}(k)$. To make an appropriate comparison, the parameters are set to be identical for both approaches, as in Section 5.1. In practical implementation, Eqn. (11) is approximated by Eqn. (13) through setting $\eta$ to be $10^{-22}$ (a trivial value approximating zero). Fig. 7 shows the results. From this figure, we can see that, surprisingly, although the temporal profiles look similar for both RMS and NMF approaches, the derived detection functions are relatively different, especially the behaviors of $h^o(k)$ and $h^{RMS}(k)$ are very different. $h^o(k)$ tends to be more balanced over the different onsets, while $h^{RMS}(k)$ is seriously unbalanced which would make the final step "peak-picking" depicted in Fig. 1 (a) much more difficult, an optimal threshold is not easy to be accurately predefined as the subsequent onset peaks may easily fall down to the similar levels of noise components. Additionally, by comparing Fig. 7 (a) and (e), it appears that $h^o(k)$ is
less sensitive to the window size selection as both methods are using the same window size. This is a good property for \( h^0(k) \), as compared with \( h^{RMS}(k) \), as we find from Fig. 7 (b) and (f) that the fluctuations in (b) may be too large to apply global thresholding for peak-picking. The similar properties have also been found for other signals, such as the signals played by piano and guitar (the results are omitted here). Note that, the analysis of the constant-balanced detection function described in Section 4 is also confirmed presented. Feasible detection functions constructed from the non-negative basis learned from the factorization of the magnitude spectrum have been proposed. As the approach is a data-driven technique, no statistical knowledge or prior information is required. The proposed technique has also been compared with the RMS envelope based approach and shown its advantages. Practical selection of the factorization rank is also examined numerically. The provided detection examples have demonstrated the good performance of the proposed technique for onset detection.

6. CONCLUSIONS

A new approach for note onset detection of musical audio by using non-negative matrix decomposition has been

<table>
<thead>
<tr>
<th>Onset Time (s)</th>
<th>G4</th>
<th>A3</th>
<th>E5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated by (10)</td>
<td>0.630</td>
<td>3.016</td>
<td>5.583</td>
</tr>
<tr>
<td>Estimated by (11)</td>
<td>0.612</td>
<td>3.007</td>
<td>5.556</td>
</tr>
<tr>
<td>Estimated by (13)</td>
<td>0.630</td>
<td>3.016</td>
<td>5.574</td>
</tr>
<tr>
<td>Marked Manually</td>
<td>0.614</td>
<td>3.009</td>
<td>5.560</td>
</tr>
</tbody>
</table>

Table 1. Onset detection results by the proposed approach as compared with the true values marked manually.

7. REFERENCES