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.

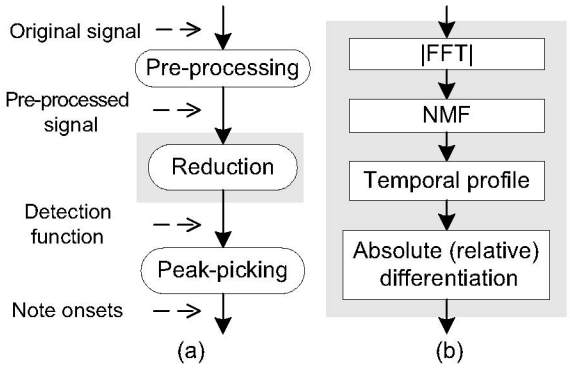

Fig. 1. Diagram of the onset detection: (a) the general scheme, (b) the proposed reduction strategy, i.e., the scheme for deriving the detection functions in this work.

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

1-4244-0657-9/06/$20.00 ©2006 IEEE

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 function, 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 $\mathbf{X} \in \mathbb{R}^{\geq 0, M \times N}$, the goal of NMF is to find nonnegative matrices $\mathbf{W} \in \mathbb{R}^{\geq 0, M \times R}$ and $\mathbf{H} \in \mathbb{R}^{\geq 0, R \times N}$, such that

$$\mathbf{X} \approx \mathbf{WH} \tag{1}$$

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 $\mathbf{W}$ and $\mathbf{H}$, we should minimize the reconstruction error between $\mathbf{X}$ and $\mathbf{WH}$. 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,

$$(\hat{\mathbf{W}}, \hat{\mathbf{H}}) = \arg\min_{\mathbf{W},\mathbf{H}} \|\mathbf{X} - \mathbf{WH}\|_F^2 \tag{2}$$

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

$$(\hat{\mathbf{W}}, \hat{\mathbf{H}}) = \arg\min_{\mathbf{W},\mathbf{H}} \sum_{m=1}^{M} \sum_{n=1}^{N} \mathbf{D}_{mn} \tag{3}$$

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

$$\mathbf{D} = \mathbf{X} \odot \log[\mathbf{X} \oslash (\mathbf{WH})] - \mathbf{X} + \mathbf{WH} \tag{4}$$

where $\odot$ and $\oslash$ 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

$$\mathbf{H} \leftarrow \mathbf{H} \odot (\mathbf{W}^T\mathbf{X}) \oslash (\mathbf{W}^T\mathbf{WH}) \tag{5}$$

$$\mathbf{W} \leftarrow \mathbf{W} \odot (\mathbf{XH}^T) \oslash (\mathbf{WHH}^T) \tag{6}$$

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 $\mathbf{X}$ 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

$$S(f,k) = \sum_{\tau=0}^{T-1} s(k\delta+\tau)w(\tau)e^{-j2\pi f\tau/T} \qquad (7)$$

where $w(\tau)$ denotes a T-point window function, $j=\sqrt{-1}$, δ is the time shift between the adjacent windows, and f is a frequency index, $f = 0, 1, \cdots, T-1$. 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 $K = \lfloor (L-T)/\delta \rfloor$, where $\lfloor\cdot\rfloor$ is an operator taking the maximum integer no greater than its argument[1]. Let $\tilde{S}(f,k)$ be the absolute value of $S(f,k)$, we can then generate the following non-negative matrix by packing $\tilde{S}(f,k)$ together,

$$\mathbf{X} = \begin{pmatrix} \tilde{S}(0,0) & \tilde{S}(0,1) & \cdots & \tilde{S}(0,K-1) \\ \tilde{S}(1,0) & \tilde{S}(1,1) & \cdots & \tilde{S}(1,K-1) \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{S}(T/2,0) & \tilde{S}(T/2,1) & \cdots & \tilde{S}(T/2,K-1) \end{pmatrix} \qquad (8)$$

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

Using the learning rules (5) and (6), $\mathbf{X}$ in (8) can be effectively decomposed into the product of two non-negative matrices, denoted as $\mathbf{W}^o \in \mathbb{R}^{\geq 0,(T/2+1)\times R}$ and $\mathbf{H}^o \in \mathbb{R}^{\geq 0, R\times K}$, i.e., the corresponding local optimum values of $\mathbf{W}$ and $\mathbf{H}$ respectively, which are obtained when the learning algorithm converges. An advantage of exploiting spectral matrix (8) is that both the obtained basis matrices $\mathbf{W}^o$ and $\mathbf{H}^o$ have meaningful interpretation. That is, $\mathbf{H}^o$ is a dimension-reduced matrix which contains the bases of the temporal patterns while $\mathbf{W}^o$ 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 $\mathbf{X}$ [9]. It is worth noting that whether the learned parts reveal 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. However, this issue turns out not to be crucial in our application, as verified in our simulations.

[1] In 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.

4. CONSTRUCTION OF DETECTION FUNCTIONS

By combining all the single *parts* together, we can reconstruct the following time series

$$h^o(k) = \sum_{r=1}^{R} \mathbf{H}^o_{rk} \qquad (9)$$

where $k = 0, \cdots, K-1$. By simulation, we found that, $h^o(k)$ actually contains a good approximation of the overall temporal profile (envelope) of the original signal. (As a result, the column vector $\mathbf{h}^o = [h^o_0, \cdots, h^o_{K-1}]^T \in \mathbb{R}^{\geq 0,K}$ describes the temporal profile of the original signal.). Therefore, $h^o(k)$ in (9), $k = 0, \cdots, K-1$, provides an alternative approach for the construction of a detection function for onset detection. To enhance the sudden changes in the signal to be detected, we take the first-order difference of $h^o(k)$ as a detection function, that is

$$h^o_a(k) = \frac{d}{dk}h^o(k), \quad k = 0, \cdots, K-1 \qquad (10)$$

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 neigbouring samples of $h^o(k)$, 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 function based on the first-order relative difference, that is

$$h^o_r(k) = \frac{\frac{d}{dk}h^o(k)}{h^o(k)}. \qquad (11)$$

Note that, the major difference between $h^o_r(k)$ in (11) and the detection function proposed by Klapuri [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,

$$h^o_b(k) = \frac{\frac{d}{dk}h^o(k)}{\eta + h^o(k)} \qquad (12)$$

where η is a positive constant. By adjusting the constant η, 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 η takes values

approaching to zero, i.e. $\eta \to 0$, in other words, $\eta \ll h^o(k)$, we have $h_b^o(k) \approx h_r^o(k)$. On the other hand, if $\eta \gg h^o(k)$, we have $h_b^o(k) \approx (1/\eta)h_a^o(k)$, which means $h_b^o(k)$ will have the same profile as that of $h_a^o(k)$, with the only difference of a scaling factor. All the above three detection functions are examined in our simulations. In fact, η has practical advantage of preventing the denominator in (11) being zero. Effectively, (12) can also be written as the logarithm,

$$h_b^o(k) = \frac{d}{dk}\log(\eta + h^o(k)) \tag{13}$$

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 196.0Hz, 110.0Hz, and 329.6Hz respectively[2]. The sampling frequency f_s for this

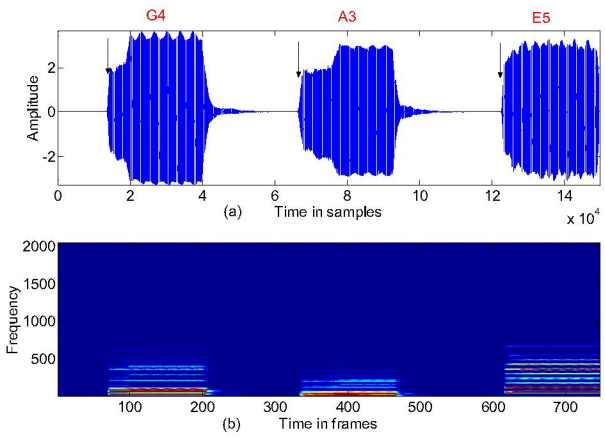

Fig. 2. The waveform of the original audio signal (a) and the generated non-negative magnitude spectrum matrix $\mathbf{X}$ (b). The onset locations are marked manually with arrows.

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 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 δ 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

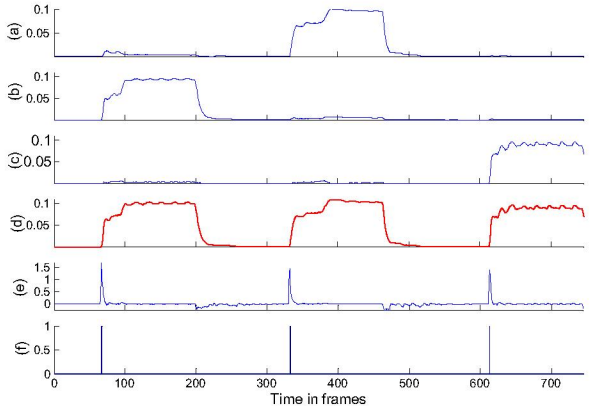

Fig. 3. Detection results of the signal depicted in Fig. 2. (a)-(c) are the visualizations of row vectors of the matrix $\mathbf{H}^o$; (d) denotes the temporal profile of $h^o(k)$, i.e., eqn. (9); (e) visualizes the detection function (13); and (f) represents the final onset locations.

described in Section 3 and 4 (see also Fig. 1 (b)), where the detection function (13) was applied, and the constant η 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 $h_b^o(k)$ 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 $h_b^o(k)$, 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

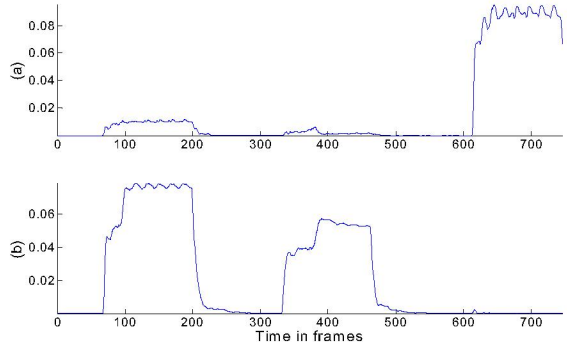

Fig. 4. The visualization of row vectors of the matrix $\mathbf{H}^o$ for rank $R = 2$.

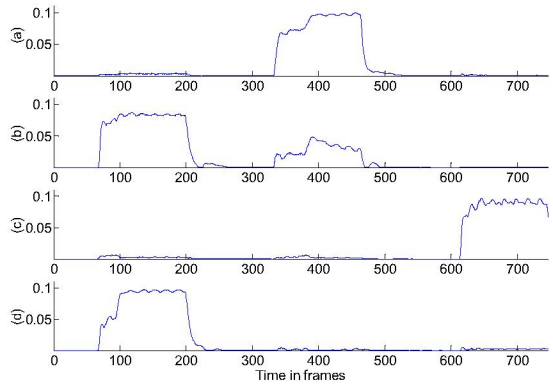

Fig. 5. The visualization of row vectors of the matrix $\mathbf{H}^o$ for rank $R = 4$.

and Fig. 5, are the visualizations of matrix $\mathbf{H}^o$ 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

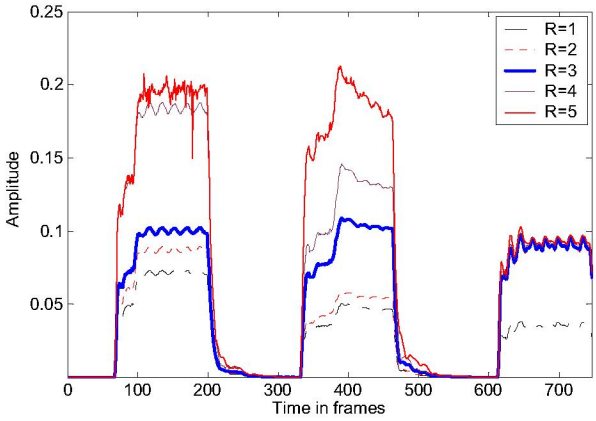

Fig. 6. Temporal profile $h^o(k)$ changes with various R varying from 1 to 5.

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.

$$h^{RMS}(k) = \sqrt{\frac{1}{T}\sum_{\tau=0}^{T-1}(s[k\delta+\tau])^2} \qquad (14)$$

where δ 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 $h_a^{RMS}(k)$, $h_r^{RMS}(k)$, and $h_b^{RMS}(k)$ 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 η 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 $h_r^o(k)$ and $h_r^{RMS}(k)$ are very different. $h_r^o(k)$ tends to be more balanced over the different onsets, while $h_r^{RMS}(k)$ 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 $h^o(k)$, 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

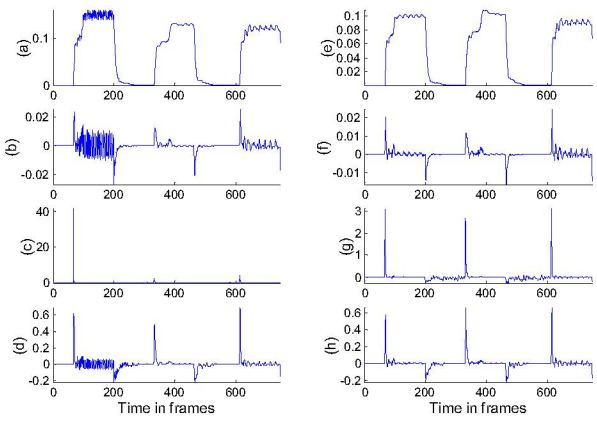

Fig. 7. Comparison between the results of the proposed detection method and that based on RMS, where the plots are (a) $h^{RMS}(k)$, (b) $h_a^{RMS}(k)$, (c) $h_r^{RMS}(k)$, (d) $h_b^{RMS}(k)$, (e) $h^o(k)$, (f) $h_a^o(k)$, (g) $h_r^o(k)$, and (h) $h_b^o(k)$, respectively.

To show the accuracy of the proposed approach, we list in Table 1 the estimated locations of the onsets in Fig. 7 (f)-(h) as compared with the values marked manually (i.e., the true values). From this table, it is observed that the onsets estimated by the difference function have slight delays from the true values, while the relative difference function provides more accurate estimates (i.e., they are closer to the true values). The constant-balanced detection function offers an intermediate performance that may be useful if there is a dramatic unbalance across the amplitude of the various onset peaks in the relative difference function. The maximum estimation error for the relative difference function is less than 5ms, which means the detection accuracy is perfect in this case, as the human auditory system is not capable of detecting gaps in sinusoids under 5ms [12]. Although the difference function appears to be less accurate, considering that the window size and overlap are relatively large (18ms and 9ms) in our experiment, the accuracy of the first-order difference function is also acceptable.

6. CONCLUSIONS

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

Onset Time (s)	G4	A3	E5
Estimated by (10)	0.630	3.016	5.583
Estimated by (11)	0.612	3.007	5.556
Estimated by (13)	0.630	3.016	5.574
Marked Manually	0.614	3.009	5.560

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

7. REFERENCES

[1] J. P. Bello, L. Daudet, S. A. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in musical signals," *IEEE Trans. Speech and Audio Process.*, vol. 13, no. 5, Sept. 2005.

[2] W. A. Schloss, "On the automatic transcription of percussive music - from acoustic signal to high-level analysis," *Ph.D. Dissertation*, Dept. Hearing and Speech, Stanford University, CA, 1985.

[3] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, pp. 3089-3092, 1999.

[4] P. Masri, "Computer modeling of sound for transformation and synthesis of musical signal," *Ph.D. Dissertation*, Univ. of Bristol, Bristol, U.K., 1996.

[5] S. A. Abdallah and M. D. Plumbley, "Probability as metadata: event detection in music using ICA as a conditional density model," *Proc. 4th Int. Symp. Independent Component Analysis and Signal Separation*, pp. 233-238, Nara, Japan, 2003.

[6] J. P. Bello and M. B. Sandler, "Phase based note onset detection for music signals," *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 49-52, 2003.

[7] C. Roads, *The Music Machine - Selected Readings from the Computer Music Journal*, MIT Press, 1989.

[8] P. Paatero, "Least squares formulation of robust non-negative factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 37, pp. 23-35, 1997.

[9] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[10] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing 13* (Proc. NIPS 2000), MIT Press, 2001.

[11] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, no. 5, pp. 1457-1469, 2004.

[12] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, 5th Ed., Academic Press, 2003.

[13] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177-180, 2003.