Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2008, Article ID 231367, 15 pages
doi:10.1155/2008/231367

Research Article

Note Onset Detection via Nonnegative Factorization of Magnitude Spectrum

Wenwu Wang,[1] Yuhui Luo,[2,3] Jonathon A. Chambers,[4] and Saeid Sanei[5]

[1] *Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, United Kingdom*
[2] *Samsung Electronics Research Institute, Communication House, Staines, TW18 4QE, United Kingdom*
[3] *Winton Capital Management Ltd., London, W8 6LS, United Kingdom*
[4] *Advanced Signal Processing Research Group, Department of Electronic and Electrical Engineering, Loughborough University, Loughborough, Leics LE11 3TU, United Kingdom*
[5] *Centre of Digital Signal Processing, Cardiff University, Cardiff, CF24 3AA, United Kingdom*

Correspondence should be addressed to Wenwu Wang, w.wang@surrey.ac.uk

Received 6 November 2007; Revised 20 February 2008; Accepted 6 May 2008

Recommended by Sergios Theodoridis

A novel approach for onset detection of musical notes from audio signals is presented. In contrast to most commonly used conventional approaches, the proposed method features new detection functions constructed from the linear temporal bases that are obtained from the decomposition of musical spectra using nonnegative matrix factorization (NMF). Three forms of detection function, namely, first-order difference function, psychoacoustically motivated relative difference function, and constant-balanced relative difference function, are considered. As the approach works directly on input data, no prior knowledge or statistical information is therefore required. Practical issues, including the choice of the factorization rank and detection robustness to instruments, are also examined experimentally. Due to the scalability issue with the generated nonnegative matrix, the proposed method is only applied to relatively short, single instrument (or voice) recordings. Numerical examples are provided to show the good performance of the proposed method, including comparisons between the three detection functions.

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, editing, information retrieval, classification, music fingerprinting, and low bit-rate audio coding [1, 2]. For example, robust detection of note onsets, note durations, pitch frequencies, and melodies becomes a common requirement in a pitch to MIDI converter which is an important component of many commercial music consoles and audio signal processing software. A significant portion of music information retrieval research has focused upon the problem of note onset detection from audio signals, which forms a basis of many algorithms for automatic beat tracking [3], rhythm description [4], and temporal segmentation of audio [5]. A recent study reveals that onset detection can also provide useful cues for sound localization in spatial audio [6]. Although onset detection is conceptually simple, it is a challenging task in audio engineering when performing robust automatic detection using computers. This is due to several major difficulties, that is, 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. Consequently, onset detection remains an open problem and demands further research effort.

A variety of approaches has been proposed in the literature, with most of them sharing an approximately common procedure, as depicted in Figure 1(a). A musical audio track may be initially preprocessed 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].

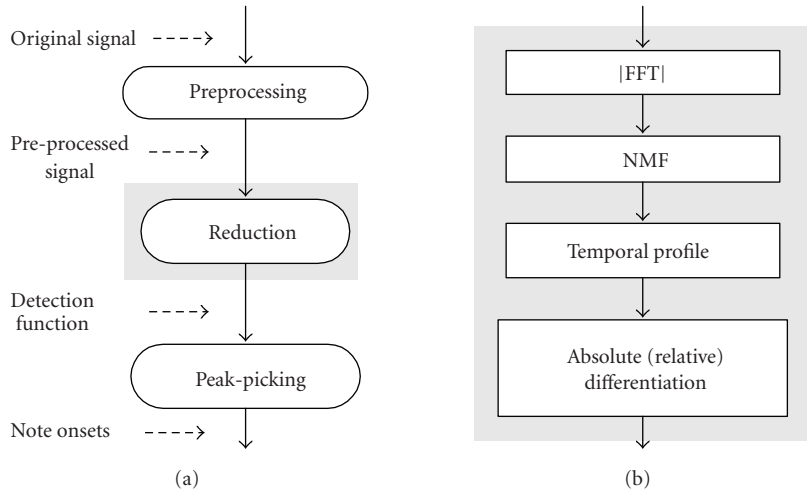

Figure 1: Diagram of the onset detection: (a) the general scheme, (b) the proposed reduction strategy, that is, the scheme for deriving the detection functions in this work.

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. Therefore, our main attention here is paid to the construction method of detection functions.

Although similar concepts relevant to human perception have been used in most existing approaches to detect onset changes, they are essentially very distinctive with regards to the various types of signal information being employed in the construction of detection functions. These include the intensity change-based methods using temporal features, for example, [7, 8]; the timbre change-based methods using spectral features, for example, [9]; model based detection methods using statistical properties, for example, [10], and methods based on phase and pitch information of signals, for example, [11, 12], 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 nonnegative decomposition of the musical spectra. The overall detection scheme is shown in Figure 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 nonnegative 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, thanks to the temporal features obtained implicitly from the NMF decomposition, an explicit computation of the signal envelope or energy function, which is required for many existing intensity-based detection approaches, is no longer necessary. Additionally, the NMF-based temporal feature is more robust for both first-order difference and relative difference as compared with direct envelope detection-based approaches (this will be highlighted in the subsequent simulation section).

Note that, due to the scalability issue with the generated nonnegative matrix (see Section 3 for more details), the proposed approach will only be applied to process relatively short recordings in our experiments. Long recordings are therefore not considered in this paper as more computing time is required by the algorithm for handling the increased size of the nonnegative matrix. Additionally, we focus only on single instrument (or voice) recordings, even though the proposed approach can, theoretically, be applicable to multiple instrument (or voice) recordings.

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 nonnegative 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, conclusions are drawn in Section 6.

2. NONNEGATIVE MATRIX FACTORISATION

NMF is an emerging technique for data analysis that was proposed recently [13, 14]. Given an $M \times N$ nonnegative

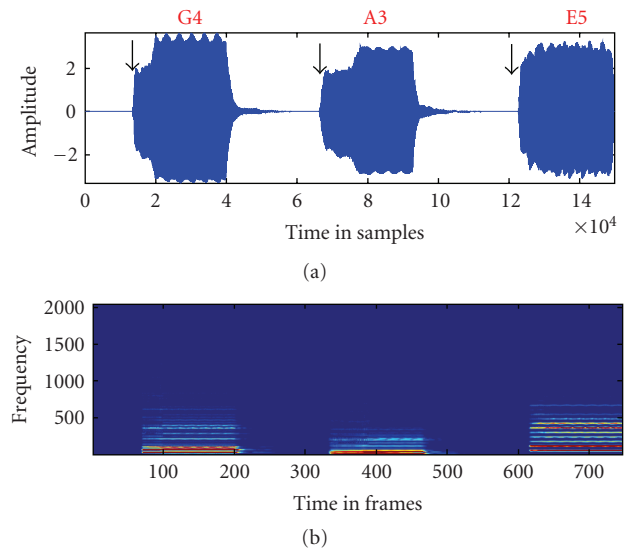

FIGURE 2: The waveform of the original audio signal (a) and the generated nonnegative magnitude spectrum matrix $\mathbf{X}$ (b). The onset locations are marked manually with arrows.

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 factorisation, generally chosen to be smaller than M (or N), or a value which satisfies $(M + N)R < MN$, 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 [13–16]. For instance, an appropriate choice is to use the criterion based on the squared Frobenius norm,

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

where $\widehat{\mathbf{W}}$ and $\widehat{\mathbf{H}}$ 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,

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

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

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

where $\odot$ and $\oslash$ denote the Hadamard (elementwise) product and division, respectively, that is, each entry of the resultant matrix is a product and division of the corresponding entries from two individual matrices, respectively. Although gradient descent and conjugate gradient approaches can

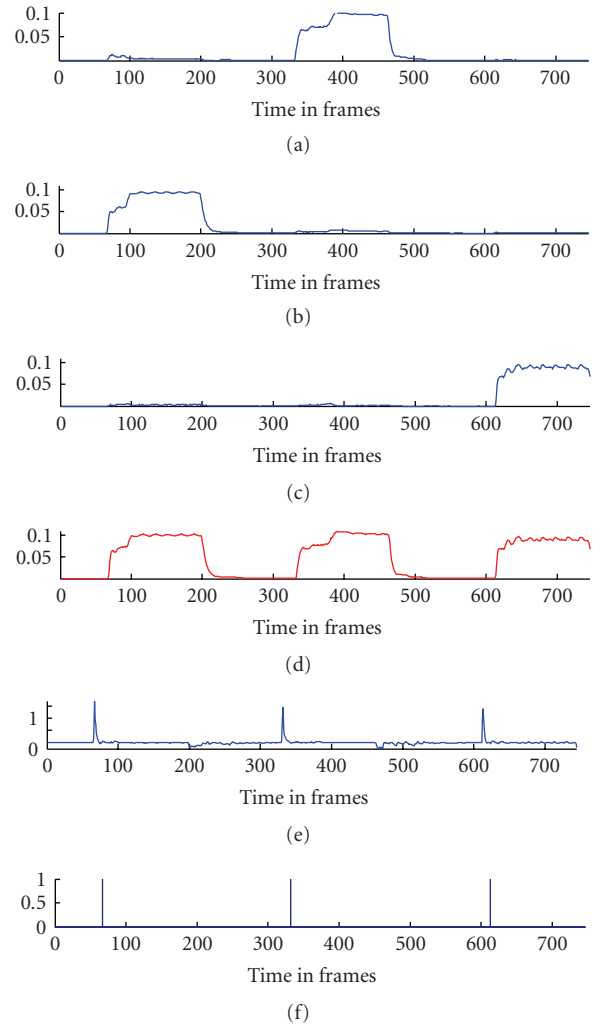

FIGURE 3: Detection results of the signal depicted in Figure 2. Figures 3(a)–3(c) are the visualizations of row vectors of the matrix $\mathbf{H}^o$; (d) denotes the temporal profile of $h^o(k)$, that is, (9); (e) visualizes the detection function (13); and (f) represents the final onset locations.

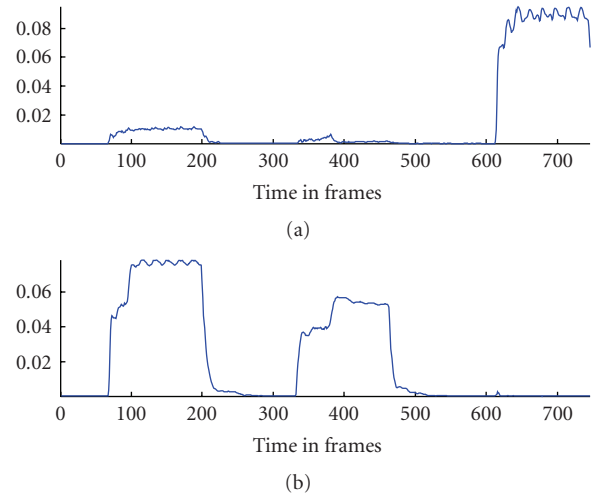

FIGURE 4: The visualisation of row vectors of $\mathbf{H}^o$ for rank $R = 2$.

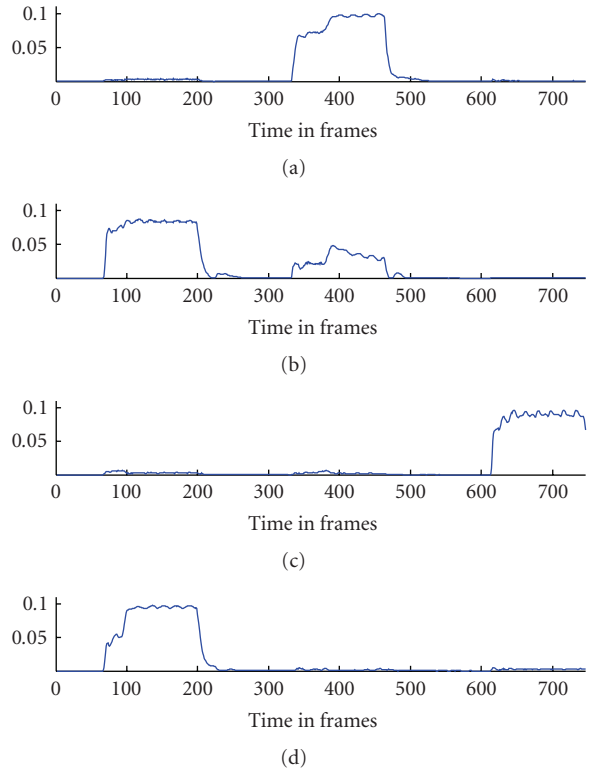

Figure 5: The visualisation of row vectors of the matrix $\mathbf{H}^o$ for rank $R = 4$.

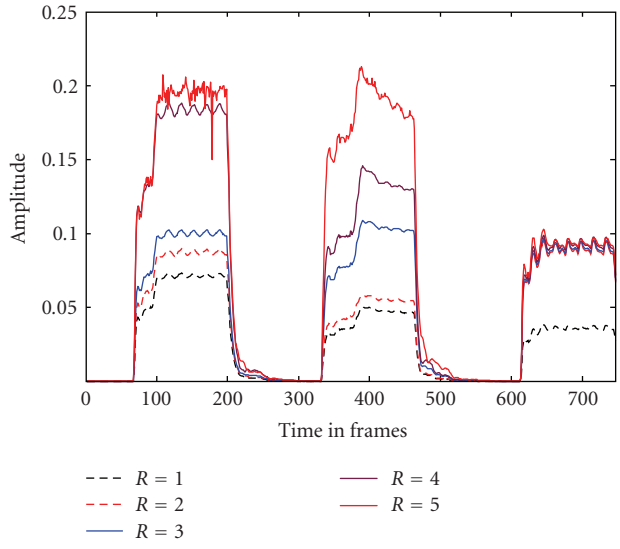

Figure 6: Temporal profile $h^o(k)$ changes with various R varying from 1 to 5.

both be applied to minimize these cost functions, we are particularly interested in the multiplicative rules developed by Lee and Seung [14, 15]. 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. In compact form,

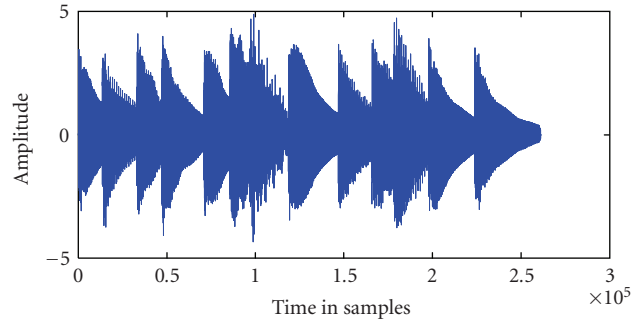

Figure 7: A real piano signal containing twelve onsets is used for showing the effect of the choice of R on the detection performance.

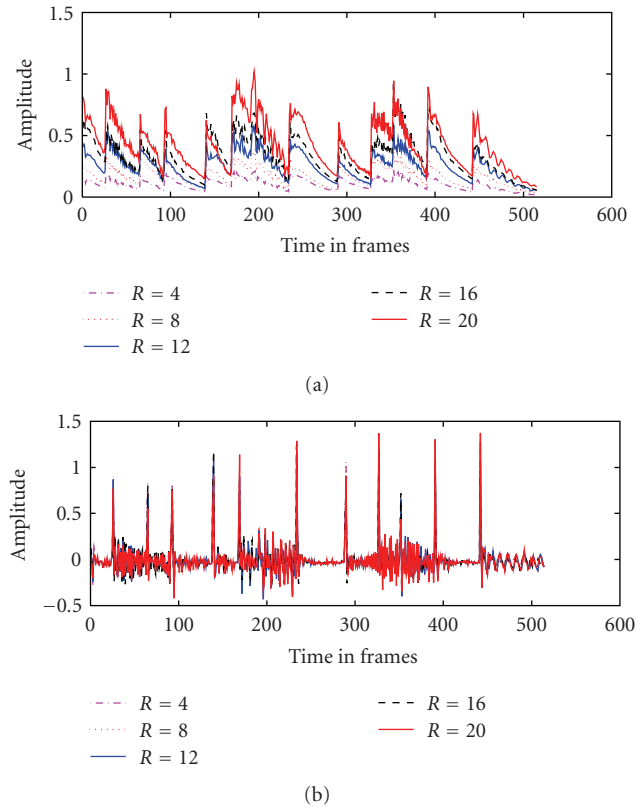

Figure 8: Detection performance in terms of $h^o(k)$ (upper subplot) and $h^o_r(k)$ (below subplot) remains relatively constant despite the variable rank R.

the multiplicative update rules for minimizing criterion (2) can be rewritten as

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

$$\mathbf{W} \longleftarrow \mathbf{W} \odot (\mathbf{X}\mathbf{H}^T) \oslash (\mathbf{W}\mathbf{H}\mathbf{H}^T), \tag{6}$$

where $(\cdot)^T$ is the matrix transpose operator, and $\leftarrow$ denotes iterative evaluation. The iteration of these update rules is guaranteed to converge to a locally optimal matrix factorization [15]. The rules (5) and (6) are used in our work.

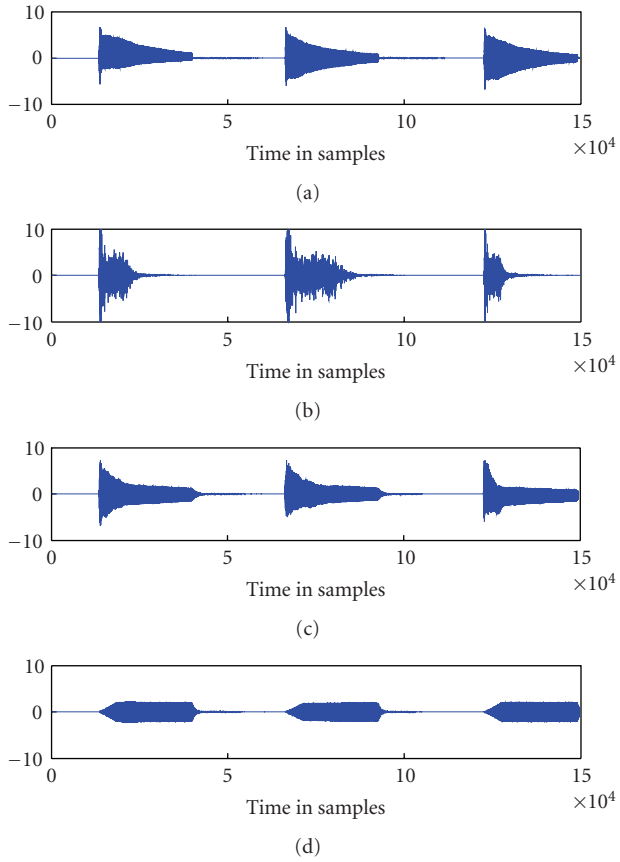

FIGURE 9: Four music audio signals played (or generated) by a (a) guitar, (b) gun, (c) piano, and (d) whistle, respectively, containing the same notes G4, A3, and E5 as those in the violin signal used in Section 5.1.

3. NONNEGATIVE DECOMPOSITION OF MUSICAL SPECTRA

For the NMF algorithm to be applied, we should first prepare a nonnegative matrix that contains an appropriate representation of the original data to be analyzed. Unlike the image data analyzed in [14], musical audio data cannot be directly used as they contain negative-valued samples. In our problem, the nonnegative matrix $\mathbf{X}$ is generated as the magnitude spectra of the input data, similar to [17]. 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

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

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, $f = 0, 1, \ldots, T-1$. 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, that is, K, is given

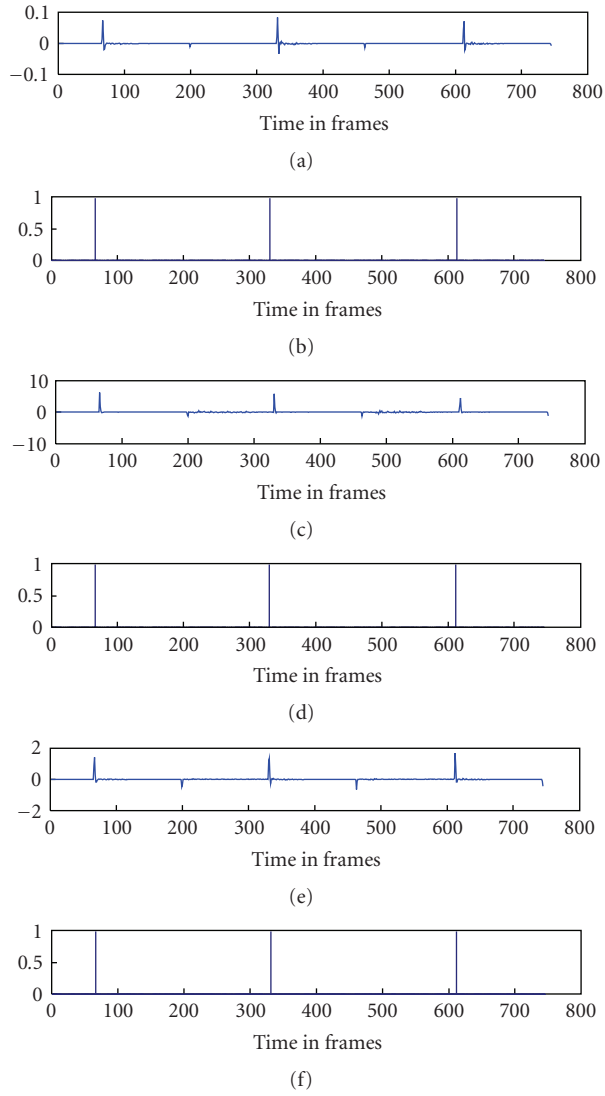

FIGURE 10: Comparison between the detection functions for the guitar signal (see Figure 9(a)). Plots (a), (c), and (e) are detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$, respectively, and plots (b), (d), and (f) are the onsets localised correspondingly using these detection functions.

as $K = \lfloor (L-T)/\delta \rfloor$, where $\lfloor \cdot \rfloor$ is an operator taking the maximum integer no greater than its argument. (In practice, zero-padding may be required to allow the remaining p ($0 \le p < \delta$) samples in the end of the signal to be covered by the analysis window.) Let $\widetilde{S}(f,k)$ be the absolute value of $S(f,k)$, we can then generate the following nonnegative matrix by packing $\widetilde{S}(f,k)$ together,

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

where only half of the frequency bins (from 0 to $T/2+1$) are used since the magnitude spectra are symmetrical along

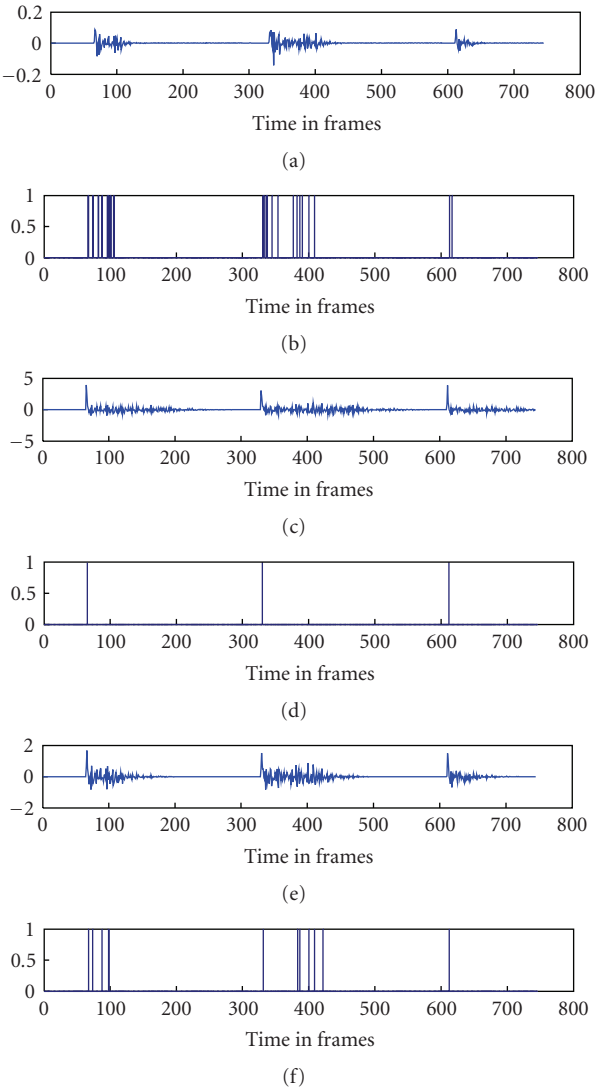

Figure 11: Comparison between the detection functions for the gunshot signal (see Figure 9(b)). Plots (a), (c), and (e) are detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$, respectively, and plots (b), (d), and (f) are the onsets localised correspondingly using these detection functions. Gunshot signals fluctuate more strongly as compared with violin, guitar, and piano signals. The onset peaks revealed by functions $h_a^o(k)$ and $h_b^o(k)$ are not as strong as those revealed by $h_r^o(k)$.

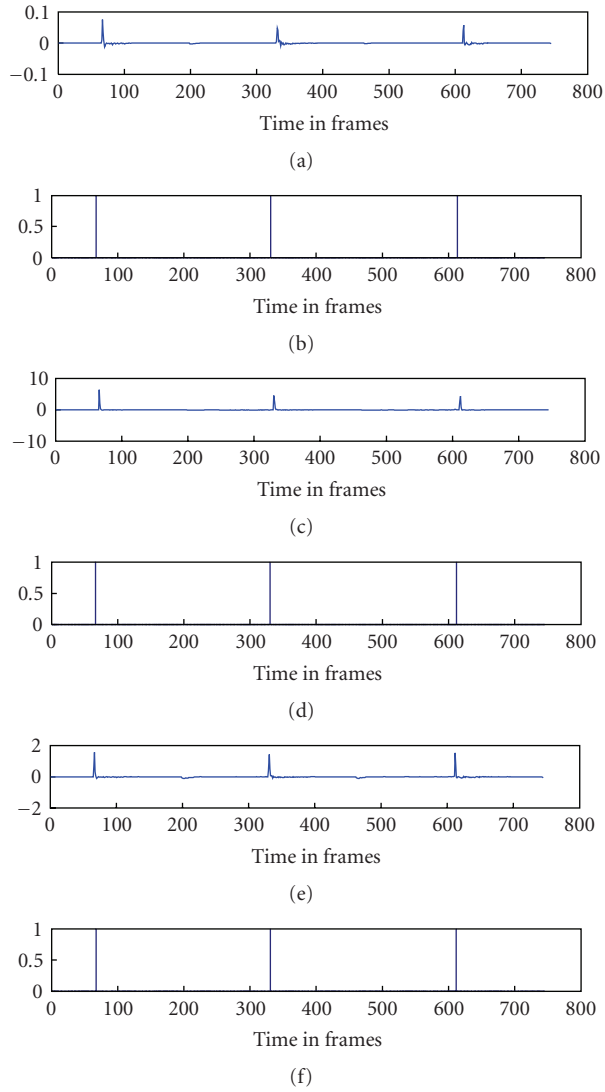

Figure 12: Comparison between the detection functions for the piano signal (see Figure 9(c)). Plots (a), (c), and (e) are detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$, respectively, and plots (b), (d), and (f) are the onsets localised correspondingly using these detection functions. The detection functions reveal strong peaks at the onset locations, while remaining relatively flat for the period of note decaying, due to the relatively small variations of dynamics of the piano signal.

the frequency axis, and the dimension of $\mathbf{X}$, that is, $M \times N$, then becomes $(T/2+1) \times K$ [18]. This non-negative matrix containing the magnitude spectra of the input signal will be used for decomposition. It is worth noting that there is a scalability issue with the generated matrix $\mathbf{X}$, that is, if the signal to be processed is very long, the constructed data matrix $\mathbf{X}$ can be very large in dimension. In this work, we focus on relatively short signals for which NMF does not pose a problem in terms of computational loads.

Using the learning rules (5) and (6), $\mathbf{X}$ in (8) can be effectively decomposed into the product of two nonnegative matrices, denoted as $\mathbf{W}^o \in \mathbb{R}^{\geq 0,(T/2+1)\times R}$ and $\mathbf{H}^o \in \mathbb{R}^{\geq 0,R\times K}$, that is, 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 part-based representation of $\mathbf{X}$ [14]. In practice, whether the learned parts reveal that the true (very often latent) patterns of the input data depend on

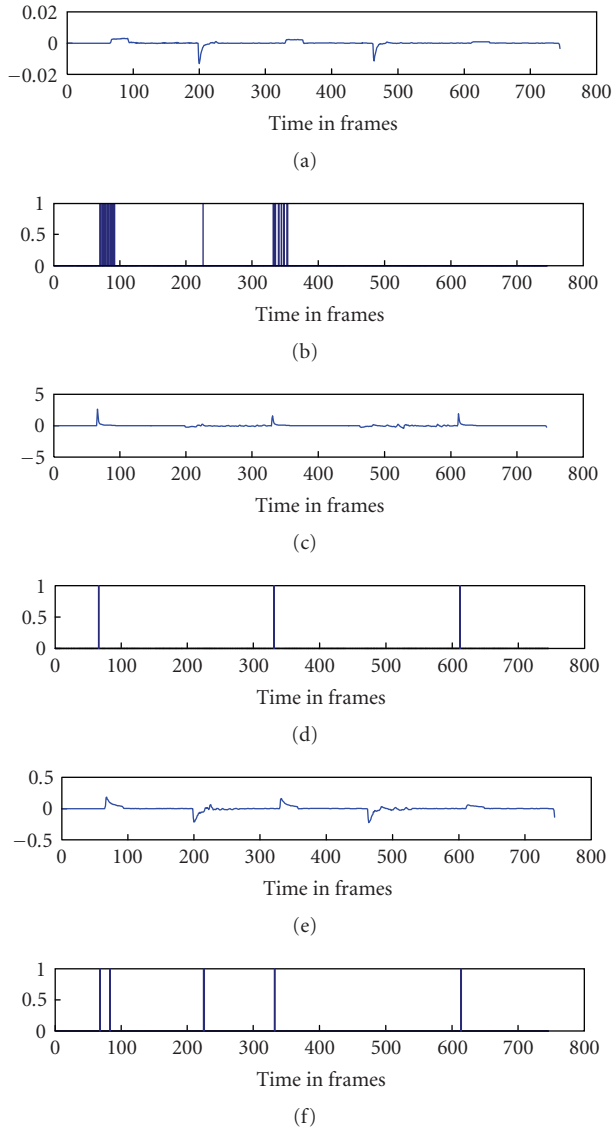

Figure 13: Comparison between the detection functions for the whistle signal (see Figure 9(d)). Plots (a), (c), and (e) are detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$, respectively, and plots (b), (d) and (f) are the onsets localised correspondingly using these detection functions. The attack of the notes of the whistle signal is not as strong as percussive audio, for example, guitar signal. Detection functions $h_a^o(k)$ and $h_b^o(k)$ are less accurate than $h_r^o(k)$ for revealing the peaks of the onset attack.

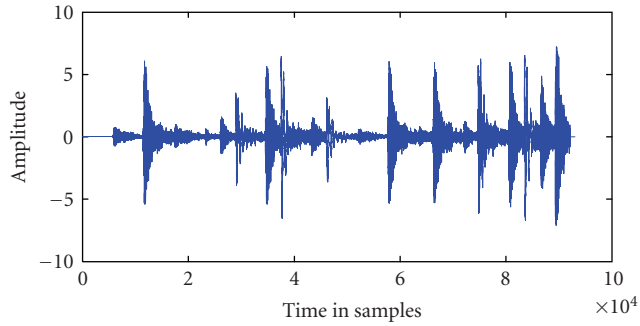

Figure 14: A realistic music signal played by a guitar.

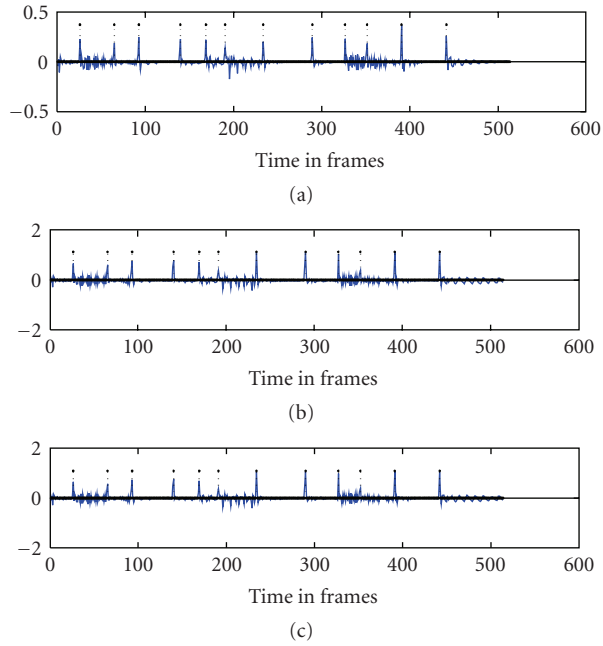

Figure 15: Comparison between the detection functions for the real piano signal (see Figure 7). Plots (a), (b), and (c) are detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$, respectively, where the detected onsets using these functions are marked with stars.

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. It is worth noting that by using the magnitude spectrum, we have actually ignored the phase information, which can be useful for improving the detection performance especially for the algorithms considering spectral features, as examined in [1, 11]. However, as will be clear in the next section, our detection functions are constructed from the temporal basis of the factorization, which has the form of a temporal feature. Therefore, phase information does not have the same impact for the detection functions in this work as those based on spectral features.

4. CONSTRUCTION OF DETECTION FUNCTIONS

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

$$h^o(k) = \sum_{r=1}^{R} \mathbf{H}_{rk}^o, \tag{9}$$

where $k = 0, \ldots, K-1$, and $h^o(k)$ provides an alternative approach for the construction of an onset detection function. 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,

$$h_a^o(k) = \frac{d}{dk} h^o(k), \quad k = 0, \ldots, K-1, \tag{10}$$

where d/dk is a *difference* operator for a discrete time series (taken from its continuous counterpart *derivative*), that is,

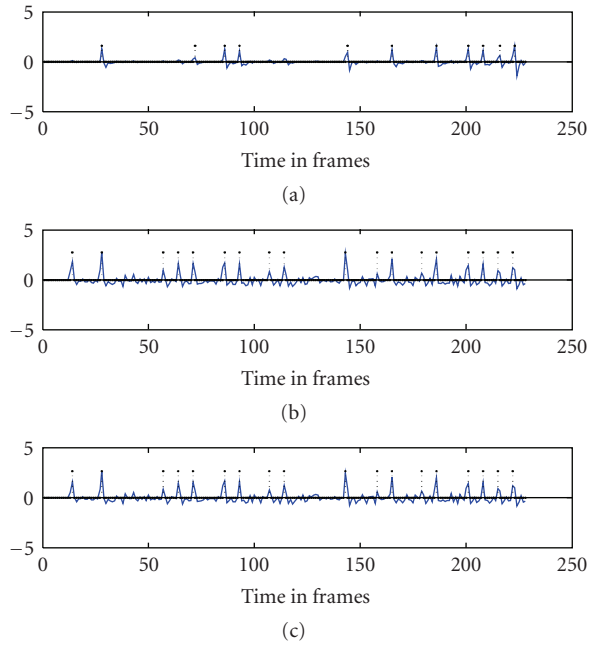

Figure 16: Comparison between the detection functions for the real guitar signal (see Figure 14). Plots (a), (b) and (c) are detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$, respectively, where the detected onsets using these functions are marked with stars.

taking the difference between two consecutive samples of the series. Therefore, $h_a^o(k) = h^o(k) - h^o(k-1)$. In other words, $h_a^o(k)$ takes the absolute difference between the neigbouring samples of $h^o(k)$ at discrete time instant k, hence it is able to reveal sudden intensity changes in the signal. However, there exists psychoacoustic evidence showing that human hearing is generally more sensitive to the relative than to the absolute intensity changes [19]. Therefore, we can also use a detection function based on the first-order relative difference, that is,

$$h_r^o(k) = \frac{(d/dk)h^o(k)}{h^o(k)}. \tag{11}$$

Note that, the major difference between $h_r^o(k)$ in (11) and the detection function proposed by Klapuri [8] lies in the different strategies taken for the construction of the temporal profile. In [8], it is formed from the energy or amplitude envelope of a group of subband signals obtained from the original signal using a filterbank decomposition.

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

$$h_b^o(k) = \frac{(d/dk)h^o(k)}{\eta + h^o(k)}, \tag{12}$$

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, that is, $\eta \to 0$, in other words, $\eta \ll h^o(k)$, we have $h_b^o(k) \approx h_r^o(k)$. On the other hand, if $\eta \gg h^o(k)$,

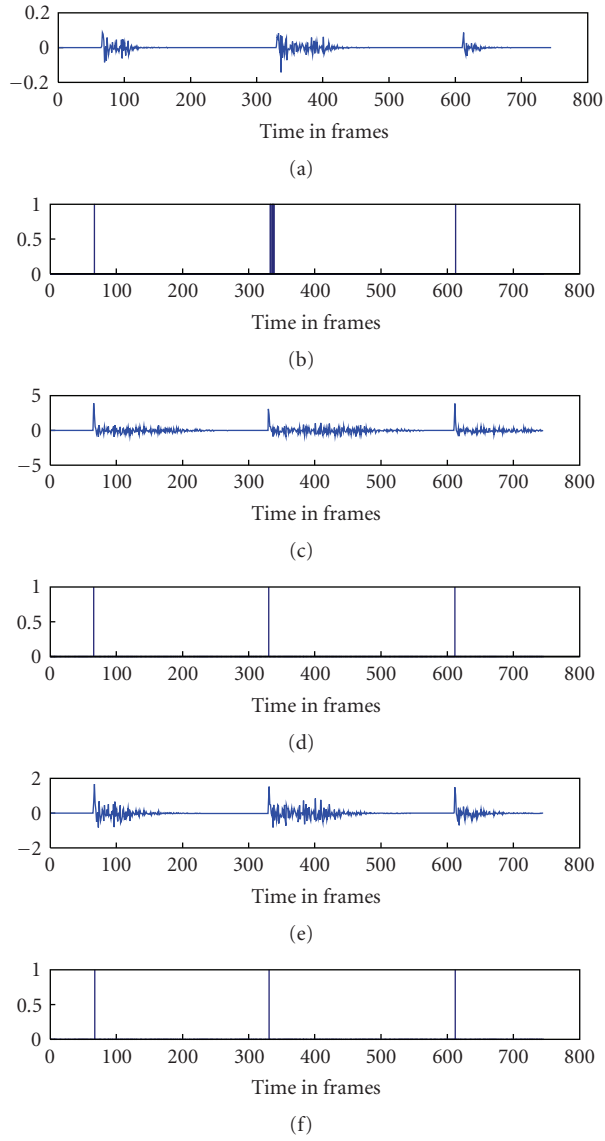

Figure 17: Increasing the threshold used for localisation of the onsets can improve the robustness against the instrumental dynamics. In this example, the threshold is set to 0.6 for onset detection of the gunshot signal (see Figure 9(b)). Plots (a), (c) and (e) are detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$, respectively, and plots (b), (d), and (f) are the onsets localised correspondingly using these detection functions. This figure is in contrast to Figure 11, where the threshold is set to 0.3.

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 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) from being zero. Effectively, (12) can also be written as the logarithm,

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

where $\log(\cdot)$ is a natural logarithm-based function of its argument.

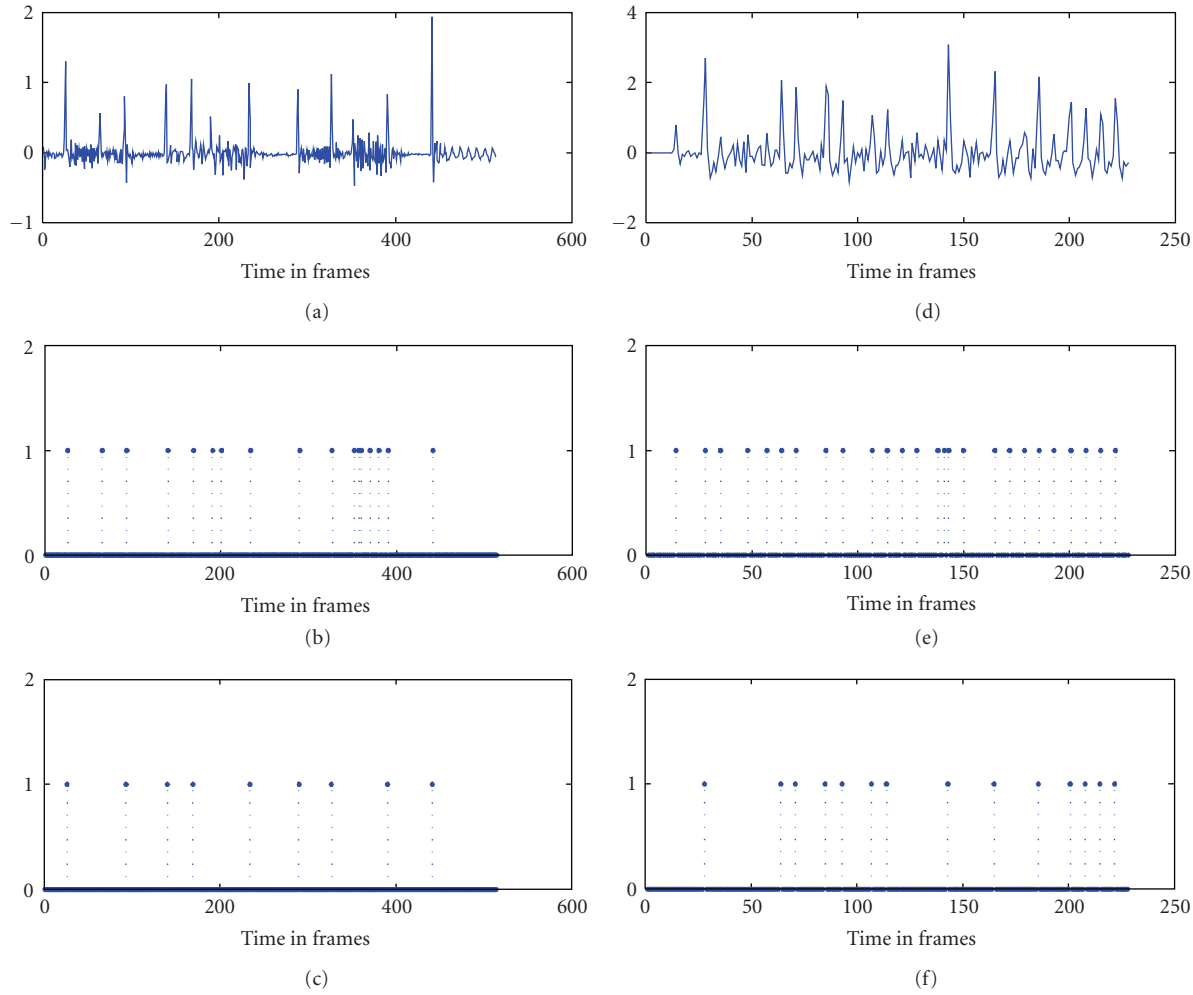

FIGURE 18: Adjusting the threshold used for localisation of the onsets can improve the robustness against the instrumental dynamics. In this example, two different values of the threshold, that is, 0.1 (corresponding to subplots (b) and (e)) and 0.3 (corresponding to subplots (c) and (f)) were used for onset detection of the piano and guitar signals, whose detection functions $h_r^o(k)$ are plotted in (a) and (d), respectively. Subplots (b) and (c) show the locations of the onsets detected using the relative difference functions with the threshold set to 0.1 and 0.4, respectively, for the piano signal, and (e) and (f) for the guitar signal.

5. NUMERICAL EXPERIMENTS

5.1. Detection example for a music 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 Figure 2(a)), whose note numbers are 55, 45, and 64, respectively, and whose frequencies are 196.0 Hz, 110.0 Hz, and 329.6 Hz, respectively. (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.) The choice of the simplistic signal, together with some others used in subsequent sections, is dictated by a particular application scenario, where MIDI commands may be used as controlling keys in some advanced music consoles and software packages for hand-free but voice or music assisted control of a mobile handset. In such applications, the music audio signals adopted can be relatively short and simple. However, realistic signals have also been tested for thorough evaluations of the proposed approach. The sampling frequency f_s for this signal is 22050 Hz. The whole signal has $L = 149800$ samples with an approximate length of 6794 milliseconds. 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, that is, the frequency resolution is approximately 5.4 Hz. The signal is segmented by a Hamming window with the window size set to 400 samples (approximately 18 milliseconds), and the time shift δ to 200 samples (approximately 9 milliseconds), that is, a half-window overlap between the neighbouring 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 (FFT number of points) T. Each segment is then zero-padded to have the same size as T for FFT

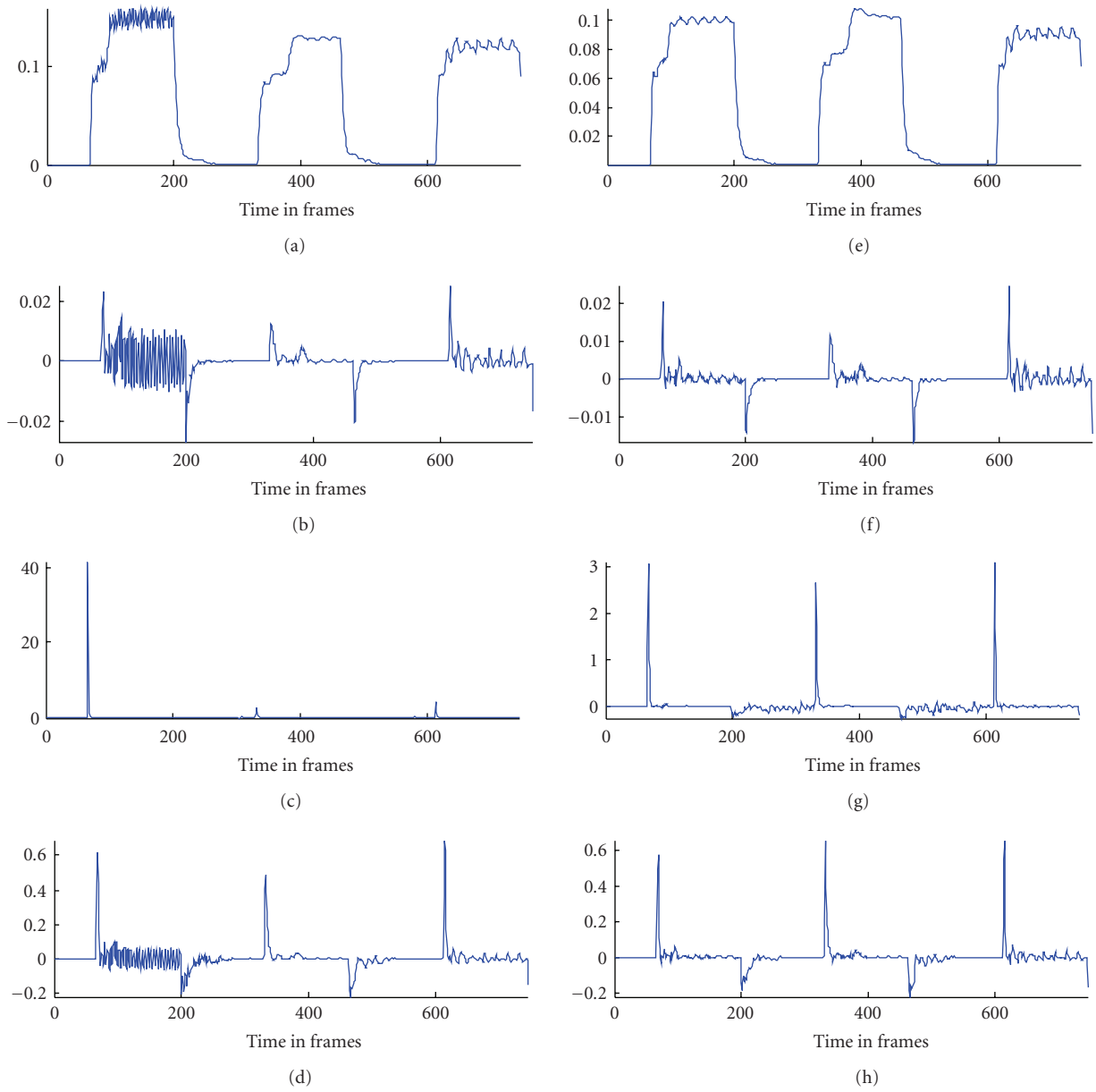

Figure 19: Comparison between the results of the proposed detection method and that based on RMS, where the plots are (a) $h^{\mathrm{RMS}}(k)$, (b) $h_a^{\mathrm{RMS}}(k)$, (c) $h_r^{\mathrm{RMS}}(k)$, (d) $h_b^{\mathrm{RMS}}(k)$, (e) $h^o(k)$, (f) $h_a^o(k)$, (g) $h_r^o(k)$, and (h) $h_b^o(k)$, respectively.

operation. The factorization rank R is set to 3, that is, exactly the same as the total number of the notes in the signal. The matrices $\mathbf{W}$ and $\mathbf{H}$ were initialized as two matrices whose elements are absolute values of zero mean real i.i.d. Gaussian random variables. The NMF algorithm was running over 100 iterations. In fact, the algorithms only took 11 iterations to converge to a local minimum in this experiment. The generated nonnegative magnitude spectrum matrix $\mathbf{X}$ is visualized in Figure 2(b). Figure 3 demonstrates the process described in Sections 3 and 4 (see also Figure 1(b)), where the detection function (13) was applied, and the constant η is set to 0.01. From Figures 3(a)–3(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 up these three parts using (9), the overall temporal profile $h^o(k)$ of the original signal is reconstructed, as shown in Figure 3(d). After applying (13) to this profile, the detection function $h_b^o(k)$ reveals apparent peaks on the locations where the notes start to strike, see Figure 3(e). The onset locations can thereby be easily determined by thresholding the local maxima of $h_b^o(k)$, see Figure 3(f), which are 630 milliseconds, 3016 milliseconds, and 5574 milliseconds, respectively.

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

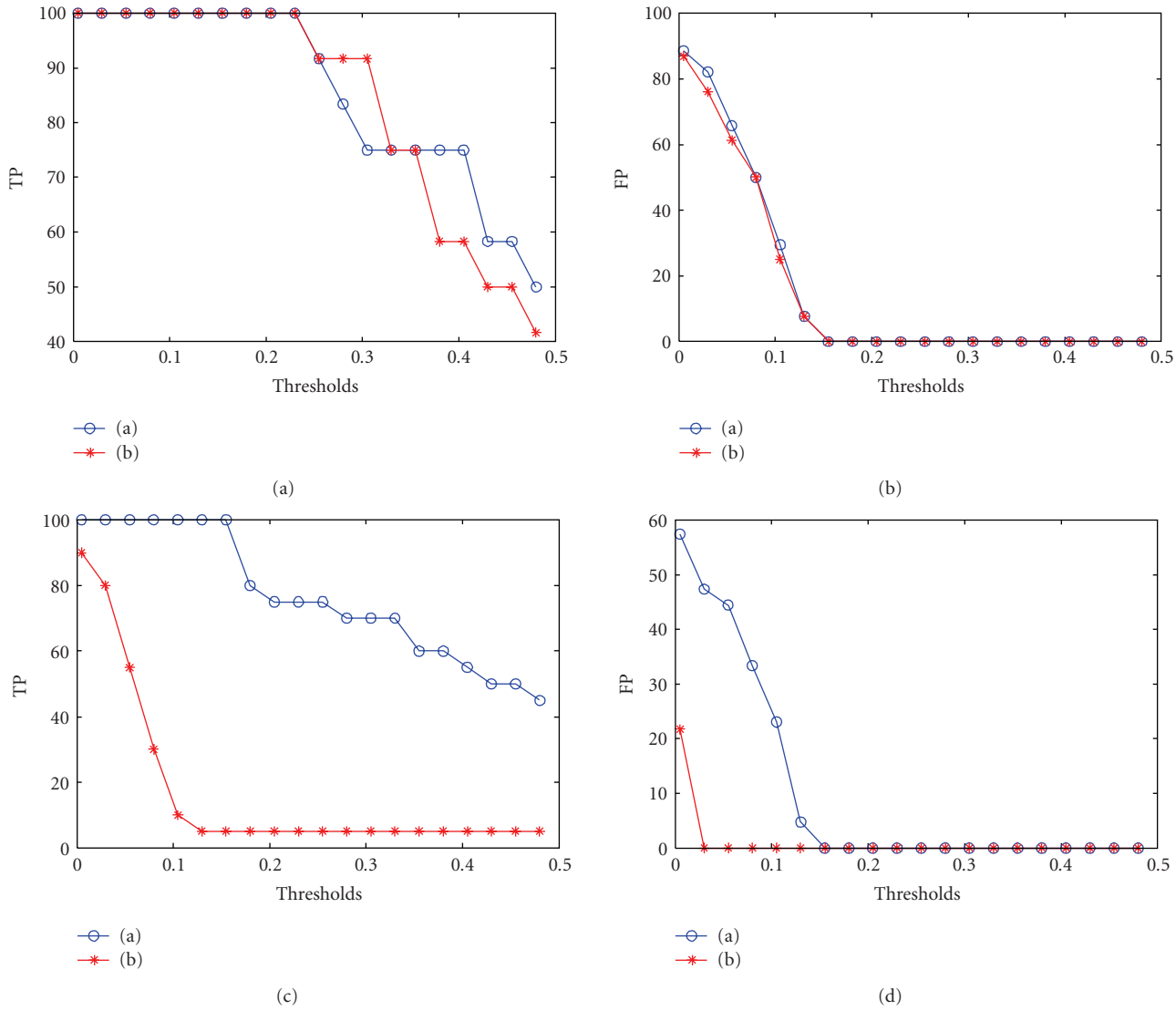

FIGURE 20: Test results of TP and FP against different thresholds for the two real music signals in Figures 7 and 14. Plot (a) corresponds to the proposed approach, and (b) to the RMS approach. The upper two plots correspond to the test results for the piano signal in Figure 7, while the below two plots correspond to the test results for the guitar signal in Figure 14. In this test, twenty different thresholds between 0.025 and 0.5 were used.

this case. In many practical situations, however, the number of hidden parts is 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 setup for the parameters as above, except for R, which we change from 1 to 5. Figures 4 and 5 are the visualizations of matrix $\mathbf{H}^o$ with R equal to 2 and 4, respectively. Figure 4(b) indicates that the total parts have not been fully separated, as there are two parts bound together in one row. Figure 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. Figure 6 plots $h^o(k)$ changing with various 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 give the same detection result.

Although a relatively simple signal was used in the above experiment, the observations found here are also valid for realistic music signals, for which we have performed extensive numerical tests. As an example, a segment of such a signal is shown in Figure 7, and $h^o(k)$ and $h_r^o(k)$ changing with various R are shown in Figure 8. Although the temporal profiles are obtained using various R differ in their amplitude, $h_r^o(k)$ remains relatively the same for different R. This promising property implies that a consistent detection

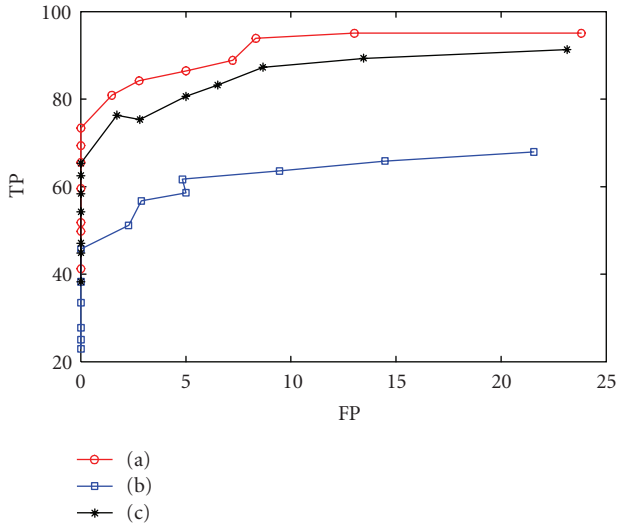

Figure 21: Average results of TP against FP for a dataset containing realistic music signals. Plots (a), (b), and (c) correspond to the proposed approach, the RMS approach, and the method in [20], where 14 different thresholds (shown as marks in each plot) were used.

performance can be achieved even though R is not known accurately.

5.3. Robustness to instruments

In Section 5.1, we have shown the good performance of the proposed approach for the stimulus played by violin. However, the performance may vary for the stimuli played by various instruments, or generated in some other ways. Figure 9 shows four audio signals containing three consecutive music notes G4, A3, and E5, which were played (generated) by a guitar, gun (gunshot), piano, and whistle, respectively. (The choice of the instruments in this experiment is dictated by a specific application scenario, as described in Section 5.1.) Figures 10(a), 10(c), and 10(e) show the detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$ obtained by applying (10), (11), and (13) to the profile of the guitar signal in Figure 9(a), and Figures 10(b), 10(d), and 10(f) show the onset locations determined by thresholding the local maxima of $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$ respectively. Similarly, Figures 11, 12, and 13 are the plots of the results of detection functions and the onset locations of the gunshot, piano, and whistle signals in Figures 9(b), 9(c), and 9(d), respectively. Note that, we use the same threshold as that in Section 5.1 for the localisation of the onsets for all these instruments. Clearly, for guitar and piano signals, $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$ all provide robust estimates of the note onsets. However, for gunshot and whistle signals, the onsets detected using $h_a^o(k)$ and $h_b^o(k)$ appear not only at the correct location, but also at some false positions, while the robustness of the detection function $h_r^o(k)$ remains relatively consistent. These experiments indicate that the robustness of the proposed method may vary with the different instruments, due to their various dynamics. For the onsets to be robustly detected, the detection functions are expected to provide instrument relatively independent performance. In this respect, $h_r^o(k)$ provides more robust detection performance against the variations of instrumental dynamics, as compared with those of detection functions $h_a^o(k)$ and $h_b^o(k)$.

To show the performance of the proposed method for more realistic signals, we have performed tests based on a commercial dataset containing signals played by different instruments (see Section 5.6 for objective performance measurements). As illustrative examples, apart from the signal in Figure 7, we show another music signal played by a guitar in Figure 14. The detection functions obtained for the piano (Figure 7) and guitar (Figure 14) signals are plotted in Figures 15 and 16, respectively, where subplots (a), (b), and (c) show the detection functions $h_a^o(k)$, $h_r^o(k)$, and $h_b^o(k)$, respectively. From the detected onsets (marked with stars) in each subplot, we can compare the performance of each detection function. Note that the threshold in the peak-picking stage was set to 0.2 for both tests. The observations made for simplistic music signals are also valid for these realistic signals played with different instruments.

5.4. Effect of thresholding

From the above section, we understand that the performance of the proposed approach may be affected by the instruments. Apart from using better detection functions, the robustness can also be improved by applying additional constraints, such as removing the false onsets if they fall into a certain distance to a detected onset, as onsets may occur in the order of one after another with a certain period of time between each other. Another effective yet simple way of improving the robustness against the stimuli is to appropriately adjust the threshold used for the localisation of onsets. Figure 17 shows that by increasing the threshold from 0.3 to 0.6, most of the false onsets detected in the gunshot signal, that is, Figure 11, have almost been removed, and the detection accuracy is greatly improved for detection functions $h_a^o(k)$ and $h_b^o(k)$. In Figure 18, applying two different thresholds in the peak-picking stage for the relative detection function $h_r^o(k)$ obtained from the real piano and guitar signals (see Figures 7 and 14), the detected onsets may vary. A small threshold may lead to some erroneous onsets, while a big threshold may result in some true onsets being missed out. It remains a practical challenge for finding optimal thresholds which are relatively immune to signal dynamics. In the literature, there are generally two main approaches for choosing thresholds, that is, using either fixed or adaptive thresholds [1]. In some situations, it may be required to develop an adaptive thresholding scheme. However, these schemes normally involve a smoothing (low-pass filtering) process [1], and therefore lead to higher computational complexity. Additionally, new methods (or parameters) may be required to be introduced (or to be tuned) for removing the fluctuations due to the smoothing process [1]. As the aim of this work is to evaluate the performance of the proposed detection functions, it is our interest to focus on the fixed thresholding scheme. For this reason, the overall performance evaluations in Section 5.6 are all based on

Table 1: Onset detection results by the proprosed approach as compared with the true values marked manually. The deviations between the estimated and the actual onset time are denoted in brackets.

Onset time (s)	G4	A3	E5
Estimated by (10)	0.630 (0.016)	3.016 (0.007)	5.583 (0.023)
Estimated by (11)	0.612 (−0.002)	3.007 (−0.002)	5.556 (−0.004)
Estimated by (13)	0.630 (0.016)	3.016 (0.007)	5.574 (0.014)
Actual	0.614	3.009	5.560

fixed thresholds. However, we have tested many different thresholds with the hope that such evaluations may provide a general guideline for choosing an optimal threshold, and also give useful clues for future development of an adaptive scheme.

5.5. 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), that is,

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

where δ is the time shift, k denotes the frame index, and T is the frame length. Expression (14) is a variation of the detection function in [7]. For simplicity, the detection functions derived from (14), corresponding to those described by (10), (11), and (13), respectively, in Section 4, are denoted as $h_a^{\mathrm{RMS}}(k)$, $h_r^{\mathrm{RMS}}(k)$, and $h_b^{\mathrm{RMS}}(k)$, respectively, which are obtained simply by replacing $h^o(k)$ with $h^{\mathrm{RMS}}(k)$. To make an appropriate comparison, the parameters are set to be identical for both approaches, as in Section 5.1. In practical implementation, (11) is approximated by (13) through setting η to be 10^{-22} (a trivial value approximating zero). Figure 19 shows the results. From this figure, we can see that, surprisingly, although the temporal profiles look similar for both the RMS and NMF approaches, the derived detection functions are relatively different, especially the behaviours of $h_r^o(k)$ and $h_r^{\mathrm{RMS}}(k)$ are very different. $h_r^o(k)$ tends to be more balanced over the different onsets, while $h_r^{\mathrm{RMS}}(k)$ is seriously unbalanced which would make the final step "peak-picking" depicted in Figure 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 Figures 19(a) and 19(e), it appears that $h^o(k)$ is smoother than $h^{\mathrm{RMS}}(k)$. This is a good property for $h^o(k)$, as we find from Figures 19(b) and 19(f) that the fluctuations in (b) may be too large to apply global thresholding for peak-picking. Since the same window size has been used for generating $h^o(k)$ and $h^{\mathrm{RMS}}(k)$, it is likely that $h^o(k)$ is less sensitive to the choice of window size. 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 by Figure 19.

To show the accuracy of the proposed approach, we list in Table 1 the estimated locations of the onsets in Figures 19(f)–19(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 5 milliseconds, which means the detection accuracy is perfect in this case, as the human auditory system is not capable of detecting gaps in sinusoids under 5 milliseconds [19]. Although the difference function appears to be less accurate, considering the window size and overlap are 18 milliseconds and 9 milliseconds in our experiment, respectively, the accuracy of the first-order difference function is also acceptable. This is because all the proposed detection functions operate framewise on the spectrum data, and an onset can be considered as correctly detected if it falls within a window size of the predetermined onset position [1, 21]. Clearly, in this experiment, all the onsets detected by the three detection functions can be deemed as accurate since they all fall within a 25-millisecond window around the true onset position. However, it is worth noting that a sample-accurate onset detection may be obtained by preselecting just those frames (and their surrounding frames) in which the onsets are detected and by processing these frames in sample-accuracy [22]. We would also like to point out that the proposed approach is especially useful for percussive audio signals, as the consistently informative amplitude changes within the signals have been effectively used for the formulation of the detection functions.

5.6. Objective performance evaluation

In this section, we evaluate the performance of the proposed approach more objectively. Two performance indices were used for this purpose, namely, the percentage of true positives (i.e., the number of correct detections relative to that of total existing onsets, denoted as TP for brevity) and the percentage of the false positives (i.e., the number of erroneous onsets relative to that of the total detected onsets, denoted as FP for brevity) [1]. A detected note is considered to be a true positive if it falls into one analysis window within the original onset. Otherwise, it is considered as a false

positive. In practice, there may exist a few missing notes not being detected at all, which is reflected by the index of TP.

In the first experiment, the two signals in Figures 7 and 14 were used. The thresholds used for peak-picking were increased gradually from 0.025 to 0.5 with a step size 0.025, that is, 20 different thresholds were tested. The proposed approach is compared with the RMS approach as described in Section 5.5. The performance analysis in the previous sections suggests that the relative difference function provides the best results in most cases, we therefore focus only on this detection function. As shown in Section 5.2, the performance of the proposed algorithm is not sensitive to the choice of rank R, we therefore set R to 12 for both signals. Figure 20 shows the result. From this figure, we can see that the proposed approach performs much better especially for the guitar signal; though for the piano signal, the performance difference between the two approaches is trivial. In accordance with the observations made in Section 5.4, an optimal threshold may be found by considering TP and FP simultaneously, that is, maximizing TP while minimizing FP. For example, for the piano signal, 0.2, can be regarded as an approximately optimal threshold for both the proposed approach and the RMS approach.

To evaluate the performance more substantially, apart from the RMS method, we have also considered another approach in the literature [20]. All the approaches were applied to a collection of realistic signals from a commercial dataset, where 21 testing signals with each containing a particular number of notes were tested. The thresholds used for peak-picking were increased gradually from 0.1 to 0.425 with a step size 0.025, that is, 14 different thresholds were tested. Note that, unlike the 20 thresholds used in the previous experiment, we discarded the relatively small (e.g., 0.025) and big (e.g., 0.5) thresholds in these tests as they either give a large number of false detections or miss many correct notes. The average performances based on these test signals are shown in Figure 21, which shows the change of TP versus FP for all 14 tested thresholds. The closer the plot approaches to the top-left corner of the figure, the better performance the approach may have. It is clear from this sense that the proposed approach performs better than the method in [20] and the RMS approach. From this figure, an optimal threshold can also be found if the TP-FP point for this particular threshold approaches the top-left corner. As is well known, music signals are composed of different notes, no matter whether they are complicated or not, from one instrument or multiple instruments. Each note can be regarded as a "part" of the whole signal. This agrees conceptually with the promising property of the NMF technique, that is, decomposing data into a part-based representation. For music signals, it naturally decomposes the data into different musical events, that is, individual parts of the musical signals. This might be the reason why NMF features perform well for the purpose of onset detection.

6. CONCLUSIONS

We have presented a new onset detection approach for musical audio by using nonnegative decomposition of a magnitude spectrum matrix. Based on the nonnegative basis learned from the factorization, we have constructed three feasible detection functions, in which the relative difference detection function provides the best performance against instrumental dynamics. The proposed technique has also been compared with the RMS envelope-based approach and its advantages have been shown. The numerical examples provided have supported the good performance of the proposed technique for onset detection.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their very helpful comments. Some preliminary results of this work appeared partly in IEEE International Workshop on Machine Learning for Signal Processing, Maynooth, Ireland, September 6–8, 2006.

REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[2] A. Lacoste and D. Eck, "A supervised classification algorithm for note onset detection," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, Article ID 43745, 13 pages, 2007.

[3] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1009–1020, 2007.

[4] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Computer Music Journal*, vol. 29, no. 1, pp. 34–54, 2005.

[5] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[6] B. Supper, T. Brookes, and F. Rumsey, "An auditory onset detection algorithm for improved automatic source localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 1008–1017, 2006.

[7] W. A. Schloss, *On the automatic transcription of percussive music: from acoustic signal to high-level analysis*, Ph.D. Dissertation, Department of Hearing and Speech, Stanford University, Stanford, Calif, USA, 1985.

[8] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '99)*, vol. 6, pp. 3089–3092, Phoenix, Ariz, USA, March 1999.

[9] P. Masri, *Computer modeling of sound for transformation and synthesis of musical signal*, Ph.D. Dissertation, University of Bristol, Bristol, UK, 1996.

[10] S. Abdallah and M. D. Plumbley, "Probability as metadata: event detection in music using ICA as a conditional density model," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA '03)*, pp. 233–238, Nara, Japan, April 2003.

[11] J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '03)*, vol. 5, pp. 441–444, Hong Kong, April 2003.

[12] C. Roads, Ed., *The Music Machine: Selected Readings from Computer Music Journal*, MIT Press, Cambridge, Mass, USA,

1989.

[13] P. Paatero, "Least squares formulation of robust non-negative factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 37, no. 1, pp. 23–35, 1997.

[14] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[15] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13 (NIPS '00)*, MIT Press, Cambridge, Mass, USA, 2001.

[16] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[17] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA '03)*, pp. 177–180, New Paltz, NY, USA, October 2003.

[18] W. Wang, Y. Luo, J. A. Chambers, and S. Sanei, "Non-negative matrix factorization for note onset detection of audio signals," in *Proceedings of the 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing (MLSP '07)*, pp. 447–452, Maynooth, Ireland, September 2006.

[19] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, San Diego, Calif, USA, 5th edition, 2003.

[20] D. L. Wang, "Feature-based speech segregation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds., IEEE Press/Wiley, New York, NY, USA, 2006.

[21] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proceedings of the 5th International Conference on Digital Audio Effects (DAFx '02)*, Hamburg, Germany, September 2002.

[22] H. Thornburg, R. J. Leistikow, and J. Berger, "Melody extraction and musical onset detection via probabilistic models of framewise STFT peak data," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1257–1272, 2007.