HARD REAL-TIME ONSET DETECTION OF PERCUSSIVE SOUNDS

Luca Turchet *

Center for Digital Music
Queen Mary University of London
London, United Kingdom
luca.turchet@qmul.ac.uk

ABSTRACT

To date, the most successful onset detectors are those based on frequency representation of the signal. However, for such methods the time between the physical onset and the reported one is unpredictable and may largely vary according to the type of sound being analyzed. Such variability and unpredictability of spectrum-based onset detectors may not be convenient in some real-time applications. This paper proposes a real-time method to improve the temporal accuracy of state-of-the-art onset detectors. The method is grounded on the theory of hard real-time operating systems where the result of a task must be reported at a certain deadline. It consists of the combination of a time-base technique (which has a high degree of accuracy in detecting the physical onset time but is more prone to false positives and false negatives) with a spectrum-based technique (which has a high detection accuracy but a low temporal accuracy). The developed hard real-time onset detector was tested on a dataset of single non-pitched percussive sounds using the high frequency content detector as spectral technique. Experimental validation showed that the proposed approach was effective in better retrieving the physical onset time of about 50% of the hits detected by the spectral technique, with an average improvement of about 3 ms and maximum one of about 12 ms. The results also revealed that the use of a longer deadline may capture better the variability of the spectral technique, but at the cost of a bigger latency.

1. INTRODUCTION

The research field of Music Information Retrieval (MIR) focuses on the automatic extraction of different types of information from musical signals. One of the most common application domains of such a field is that of automatic music transcription [1]. Another domain is represented by the identification of timbral aspects [2], which might be associated to different expressive intents of a musician [3] or to a particular playing technique that generated a sound [4]. The retrieval of the instant in which a pitched or unpitched musical sound begins, generally referred to as *onset detection*, is a crucial step in a MIR process. Numerous time- and spectrum-based techniques have been proposed for this purpose (see e.g., [5, 6]), some of which are based on the fusion of various methods [7].

Up to now, the majority of MIR research on onset detection has focused on offline methods based on the analysis of large datasets of audio files. Nevertheless, different techniques have also been developed for real-time contexts [8, 9, 10], especially for retrieving information from the audio signal of a single musical instrument [11, 12]. Real-time implementations of some onset detection techniques have been made available in open source libraries (e.g., *aubio*[1] [13]). Typically, the performance of an onset detector is assessed against annotated datasets. Such annotations may define onset times in line with human perception [14] or with the actual physics (which are generally referred to as *perceptual* and *physical* onset times respectively [6]).

Once an onset has been detected, it is possible to apply, to the adjacent part of the signal, algorithms capable of extracting different types of information (e.g., spectral, cepstral, or temporal features [15, 16]). For instance, such information may be used to identify the timbre of the musical event associated to the detected onset. In turn, the identified timbre may be utilized for classification tasks by means of machine learning techniques [17]. A challenging timbral classification concerns the identification of different gestures performed on a same instrument. For this purpose, it is crucial to understand the exact moment in which an onset begins. Indeed lot of the timbral information is contained in the very first part of the signal of a musical event.

However, to date, the onset detection methods available in the literature are little sensitive to the challenge of retrieving the exact initial moment of a musical event (i.e., the physical onset time). For instance, the Onset Detection Task specifications of the Music Information Retrieval Evaluation eXchange (MIREX)[2], and most of the papers in the area of onset detection, consider detected onsets as true positives if they fall within a window of 50 ms around the onset time reported in an annotated dataset. Furthermore, the vast majority of freely available datasets for MIR research are not accurate at millisecond or sub-millisecond level, which would be useful to designers of real-time MIR systems.

Currently, the most successful onset detectors are those based on frequency representation of the signal [5, 6, 18] (as shown by the results of MIREX context between 2005 and 2017[3]). Typically, detecting efficiently and effectively an onset using spectral methods requires at least 5.8 milliseconds after the occurrence of the peak of the involved onset detection function (ODF), considering a window size of 256 samples for the Short Time Fourier Transform and a sampling rate of 44.1 kHz. However, for such methods the time between the actual onset and the reported onset is unpredictable and may largely vary according to the type of sound in question. This is due to the fact that spectral methods are not based on the actual initial moment of the hit but on the identification of the ODF's peak (or its beginning), which may occur some millisec-

* This work was supported by a Marie-Curie Individual fellowship from the European Union's Horizon 2020 research and innovation programme (749561).

[1] Available at www.aubio.org

[2] http://www.music-ir.org/mirex/wiki/2017:Audio_Onset_Detection

[3] http://www.music-ir.org/mirex/wiki/MIREX_HOME

onds after the physical onset. Such variability and unpredictability of spectrum-based onset detectors may not be convenient in some real-time applications. An example of such applications is represented by those hybrid acoustic-electronic musical instruments that must react with minimal latency to a performer's action, involving a response (such as the triggering of a sound sample) that accounts for the correct classification of the timbre of the sound acoustically produced (see e.g., [4]).

This paper addresses the improvement of existing onset detectors to achieve a less variable and more predictable time accuracy in real-time contexts. Specifically, we limit our investigation to sounds of single non-pitched percussive instruments (therefore implementing a "context-dependent" method, not a "blind" one). In more detail, we do not consider instruments capable of producing radically different sounds, such as those of a full drum kit, but rather all the possible gamut of sounds resulting from hits on a same instrument (which may be produced by the player using different gestures). This research originated while developing an improved version of the smart cajón reported in [19], which belongs to the family of smart musical instruments [20]. For that application it was fundamental to retrieve with a higher degree of temporal accuracy the onsets corresponding to each hit produced on the smartified acoustic cajón, since the portion of signal subsequent to each onset was utilized for gesture classification (using audio feature extraction methods and machine learning algorithms based on the extracted features). The classified gesture was then repurposed into a triggered sound sample concurrent with the acoustic sound.

Notably, the real-time repurposing of a hit in hybrid acoustic-electronic percussive instruments such as the smart cajón, poses very strict constraints in terms of accuracy of detection and temporal reporting: the system not only must guarantee that a produced hit is always detected, but also that the onset is reported within a certain latency as well as that such latency is constant. Any success rate of onset detection different from 100% or with a too high latency is simply not an option for professional musicians, who require a perfectly responsive instrument and feel that they can truly rely on it. This imposes that the latency between their action on the instrument and the digital sound produced in response to it must be imperceivable.

Such strict requirements parallel those of hard real-time operating systems where a task must be accomplished at the end of a defined temporal window (deadline), otherwise the system performance will fail [21]. Therefore, for the terminology's sake, to distinguish our method from other real-time algorithms less sensitive to temporal accuracy we introduce the notion of *hard real-time onset detector (HRTOD)* and *soft real-time onset detector (SRTOD)*[4]. The latter are those methods that have more tolerant constraints in terms of the accurate onset time identification as well as in the variability of such time. Examples of methods belonging to the SRTOD category are the implementations reported in [11] and [12], which present a real-time drum transcription system available for the real-time programming languages Pure Data and Max/MSP. Another example is represented by the study reported in [22], where a recurrent neural network is employed for the onset detection task. Notably, our proposed method does not intend to reduce the actual latency of state-of-the art methods. Instead it aims at guaranteeing that the time of an onset is reported more accurately at the end of a set time window computed from the physical onset, in the same way as it happens for tasks in a hard real-time operating system.

The remainder of the paper is organized as follows. Section 2 describes the proposed onset detector that meets the requirements mentioned above as well as an implementation for it in Pure Data. Section 3 presents the results of the technical evaluation performed on various datasets of single percussive non-pitched instruments, while Section 4 discusses them. Section 5 concludes the paper.

2. PROPOSED HARD REAL-TIME ONSET DETECTOR

The proposed onset detection algorithm relies on the combination of time- and spectrum-based techniques. This choice was motivated by our initial experimentations, which suggested that methods based on temporal features may have a higher degree of accuracy in detecting the physical onset time. On the other hand, onset detection methods based on the spectral content may be less prone to false positives and false negatives compared to methods based on temporal features if their parameters are appropriately tuned, although they may suffer from unpredictability and variability issues in timing accuracy.

The proposed onset detector aims to take advantage of the strengths of the two approaches. Specifically, a time-based technique capable of detecting more reliably the very initial moment of a hit, but also more sensitive to false positives and false negatives, was used in parallel with a spectrum-based technique that was tuned to optimize the performance in terms of F-measure. Moreover, our goal was not only to detect an onset with minimal delay after the initial moment of contact of the exciter (e.g., hand, stick, etc.) and the resonator (e.g., skin of a drum, wood of a cajón panel), but also to ensure a high temporal resolution in tracking two subsequent hits. We set such resolution to 30 ms since this is approximatively the temporal resolution of the human hearing system to distinguish two sequential sound events [23]. Such a resolution is also adopted by the real-time onset detector proposed in [22].

The implementation of the proposed onset detector was accomplished in Pure Data, considering as input a mono live audio signal sampled at 44.1 kHz. The implementation was devised to achieve high computational efficiency, and more specifically, to run on low-latency embedded audio systems with low computational power (e.g., the Bela board [24]), which may be involved in the prototypization of smart instruments. The next three sections detail the utilized time- and spectrum-based techniques as well as the adopted fusion policy.

2.1. Time-based method

The time-based method (TBM) here proposed is inspired by the approaches to onset detection described in [5] and [8]. It must be specified that this technique only provides as output an onset timing, not the associated peak. Notably, the time-based method proposed in [25], which employs the logarithm of the input signal's energy to model human perception, was not utilized. This was due to the fact that we were interested in the physical onset not in the perceptual one. Figure 1 illustrates the various steps in the onset detection process.

We generated an ODF as follows. Firstly, we filtered the input signal with a high pass filter whose cutoff frequency was tuned on the basis of the type of percussive instrument being analyzed. This is the main difference with the time-based methods reported in [5],

[4]This terminology should not be confused with that used to discriminate onsets as hard (usually by percussive instruments, pitched and unpitched) or soft (e.g., produced by bowed string instruments).

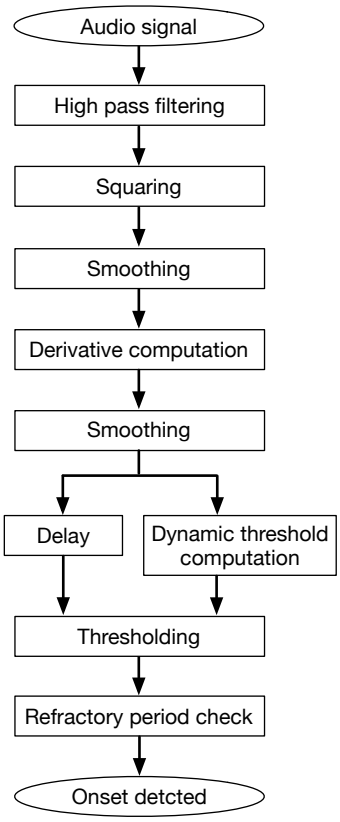

Figure 1: *Block diagram of the various steps involved in the time-based onset detector.*

which do not follow this initial step. Performing such a step allows one to drastically reduce the number of false positives while at the same time preserving (or only marginally affecting) the true positives. Secondly, we computed the energy by squaring the filtered signal. Subsequently, the energy signal underwent a smoothing process accomplished by a lowpass filter. This was followed by the calculation of the first derivative and again the application of a lowpass filter. The cutoff frequencies of the lowpass filters are configurable parameters.

Subsequently, a dynamic threshold (which is capable of compensating for pronounced amplitude changes in the signal profile) was subtracted from the signal. We utilized a threshold consisting of the weighted median and mean of a section of the signal centered around the current sample n:

$$\delta(n) = \lambda \cdot median(D[n_m]) + \alpha \cdot mean(D[n_m]) \qquad (1)$$

with $n_m \in [m-a, m+b]$ where the section $D[n_m]$ contains a samples before m and b after, and where λ and α are positive weighting factors. For the purpose of correctly calculating the median and the mean around the current sample, the pre-thresholded signal must be delayed of b samples before being subtracted from the threshold. The parameters a, b, λ and α are configurable. The real-time implementation of the median was accomplished by a Pure Data object performing the technique reported in [26].

The detection of an onset was finally accomplished by considering the first sample n of the ODF satisfying the condition:

$$n > \delta(n) \quad \& \quad n > \beta \qquad (2)$$

where β is a positive constant, which is configurable. To prevent repeated reporting of an onset (and thus producing false positive detections), an onset was only reported if no onsets had been detected in the previous 30 ms.

2.2. Spectrum-based onset detection technique

Various algorithms for onset detection available as external objects for Pure Data were assessed, all of which implemented techniques based on the spectral content. Specifically, we compared the objects i) *bonk~* [27], which is based on the analysis of the spectral growth of 11 spectral bands; ii) *bark~*, from the *timbreID* library[5], which consists of a variation of bonk~ relying on the Bark scale; iii) *aubioonset~* from the *aubio* library [13], which makes available different techniques, i.e., broadband energy rise ODF [5], high frequency content ODF (HFC) [28], complex domain ODF [29], phase-based ODF [30], spectral difference ODF [31], Kulback-Liebler ODF [32], modified Kulback-Liebler ODF [13], and spectral flux-based ODF [6]. Several combinations of parameters were used in order to find the best performances for each method.

All these spectral methods shared in common a variable delay between the actual onset time and the time in which the onset was detected. In the end *aubioonset~*, configured to implement the HFC was selected because it was empirically found to be capable of providing the best detection accuracy. This in line with Brossier's observations reported in [13]. A refractory period of 30 ms was applied after a detection to eliminate possible false positives within that window.

2.3. Fusion policy

Our strategy for combining the two onset detectors calculated in parallel consists in considering an onset as true positive if detected by HFC, and subsequently retrieving the initial moment by looking at the onset time of the corresponding onset (possibly) detected by TBM. The policy to fuse these two types of information highly depends on the deadline for reporting the onset after the physical one. In our HRTOD such a deadline is a configurable parameter, which must be greater than the duration of the window size chosen for HFC. On a separate note, we specify that while the time based method acts on a high-pass filtered version of the input signal, HFC uses the original signal.

The fusion policy is presented in the pseudocode of algorithm 1. For clarity's sake, the reader is referred to Figure 2. If HFC produces an onset and TBM has not yet, then the onset time is computed by subtracting the duration of HFC's window size from the time of the onset detected by HFC, and such an onset is reported after the difference between the deadline and the duration of HFC's window size. Any onset candidate deriving from TBM produced in the 30 ms subsequent to the reporting of HFC gets discarded.

Conversely, if TBM produces an onset and HFC has not yet, then the algorithms checks whether an onset is produced by HFC in the next amount of time corresponding to the duration of HFC's window size minus the *temporal error* that is estimated affecting TBM (i.e., the delay between the time of the physical onset and the time of the onset reported by TBM). If this happens, then such

[5] Available at `www.williambrent.com`

onset is reported after the amount of time corresponding to the deadline minus the duration of HFC's window size, and the onset time is computed by subtracting the duration of HFC's window size from the time of the onset detected by HFC. The error that affects TBM is a configurable parameter for the algorithm, whose value must be less than the duration of HFC's window size. Such an error is estimated on the basis of analyses performed on the input signal of the percussive instrument in question.

If HFC has not produced an onset in the time corresponding to the duration of HFC's window size minus the estimated error after the reporting of the onset by TBM, then the algorithm checks whether HFC has produced an onset in the next amount of time corresponding to deadline minus the duration of HFC's window size plus the estimated error. If this happens, then such onset is reported immediately and the onset time is computed by subtracting the estimated error from the time of the onset detected by TBM.

Critical to this fusion policy is the choice of the parameters governing the behavior of TBM. Indeed, if TBM produces too many false positives there is the risk of erroneous associations of onsets detected by TBM to onsets detected by HFC, as these might happen just before the actual physical onset. Conversely, if TBM produces too many false negatives, then HFC will be much less improved in terms of accuracy.

To estimate the TBM error while designing a real-time audio system, one could record the live audio produced by the system, apply the TBM configured to optimize the F-measure, and calculate the temporal distance between the time of the onset reported by TBM and the time of the physical onset (which can be determined by annotating the recorded dataset). Subsequently, the found minimum value could be used as the TBM error estimate. This guarantees that all onset times marked as improved with respect to the corresponding ones of the HFC, are effectively improved. Nevertheless, this would also limit the amount of improvement, as some onsets detected by HFC could be improved using a slightly greater TBM error estimate.

A less conservative strategy here recommended, consists in tolerating a small error on the time reporting of few onsets, such that the temporal accuracy for those onsets would be worsen only marginally, while at the same time increasing the temporal accuracy of a much greater number of HFC onsets. Specifically, our criterium adopted to determine an estimation of the TBM error is to select the minimum between the value of the first quartile and the result of the sum of 1 ms to the minimum delay found between the beginning of the sinusoid and the annotated physical onset:

$$TBM\_estimated\_error = min \begin{cases} 1^{st} quartile \\ 1 + min(error) \end{cases} \quad (3)$$

This allows one to tolerate in the worst case a maximum error of 1 ms for some of the hits (whose amount is lower or equal than the 25% of the total hits of the dataset). Therefore, the calculated onset times deriving from TBM can be effectively considered as an improvement compared to HFC in the majority of the cases.

3. EVALUATION

The temporal accuracy of the developed HRTOD was assessed on a dataset of recordings of four single percussive non-pitched instruments: conga, djembe, cajón, and bongo. In this evaluation we were not interested in assessing the detection accuracy of our HRTOD in terms of F-measure as this is fully determined by HFC (whose performance is well documented in the literature [28, 13]). Our focus was exclusively on the assessment of the actual improvement offered by HRTOD in terms of temporal accuracy compared to HFC. For this purpose, we carefully selected the parameters of TBM in order to maximize the F-measure and avoid any error in the fusion policy, likewise for HFC (see Table 1). In this investigation we were also interested in assessing whether the performance of HRTOD differed between the instruments and for two deadlines.

3.1. Procedure

In absence of accurate annotations of datasets of single percussive non-pitched instruments among those normally used by the MIR community, which could have served as a ground truth, we opted for using two freely available online libraries[6]. Such libraries were selected for the high quality recordings and the involvement of a large variety of playing styles and percussive techniques on the four investigated instruments. Those libraries contain 81 short recordings of hits on conga, 38 for djembe, 85 for cajon, and 31 for bongo.

To annotate the datasets we visually inspected the waveforms of the files and considered the first clear change in the waveform as an actual physical onset. Specifically, in this manual process we aimed at achieving an error tolerance of 0.5 ms. We did not annotate the whole database but only 100 hits per each instrument. Such annotated hits were those utilized to determine the estimated error of TBM. They were selected as follows. We recorded along with the file waveform, two additional tracks containing short sinusoidal waves beginning at the instants in which the onset were detected respectively by HFC and TBM (see Figure 2). Subsequently, for each sinusoid in the TBM track that was related to a true positive detected by HFC but happening before it, we calculated the time difference between the annotated physical onset and the beginning of the sinusoid. In this calculations one needs to add the time corresponding to b samples of which the waveform was delayed (in our case this corresponds to 0.045 ms as 2 samples were used for b).

For each instrument we randomly chose a subset of files and considered the first 100 hits satisfying the mentioned condition. For our purpose, an amount of 100 hits gives a reasonably accurate measurement in statistical sense and could be considered as the number that a designer of a real-time system would use to get the estimate of TBM error from analyzing live recordings of the system. Table 2 shows for each instrument the results of the analysis conducted on the 400 annotated hits to determine the estimate of TBM error, as well as the corresponding average and maximum error one would still get using it.

We configured HRTOD with two deadlines, at 11.6 and 18 ms, to compare its performance in the case of a short and long deadline. Indeed a longer deadline would have been able to capture those onsets detected by HFC after the short deadline is elapsed, given the HFC variability. The deadline of 11.6 ms was selected because it is equivalent to the time needed to compute analyses on 512 samples at 44.1 kHz sampling rate, therefore, the first 11.6 ms of the signal can be utilized without involving in the analysis any

[6] http://cdn.mos.musicradar.com/audio/samples/musicradar-percussion-samples.zip and http://www.stayonbeat.com/wp-content/uploads/2013/07/Bongo-Loops_StayOnBeat.com_.zip

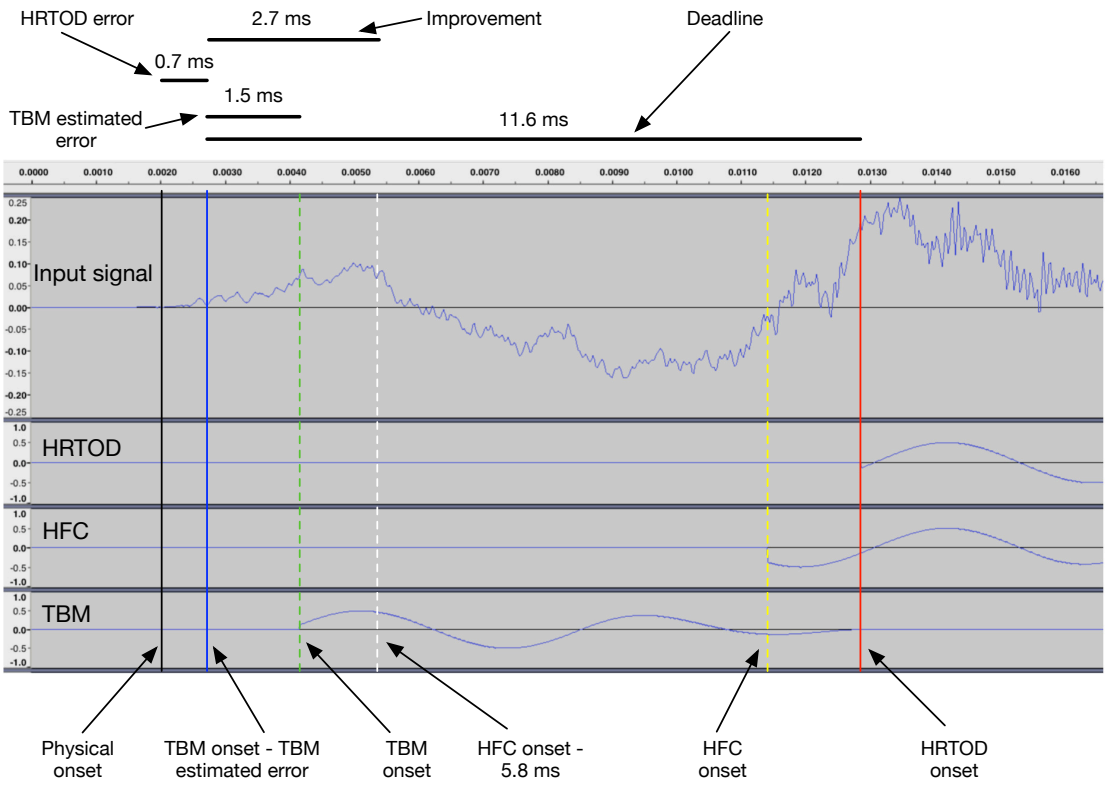

Figure 2: *Waveforms of the input signal of a hit on cajón and of three short sine waves triggered at the times of detecting the onsets using TBM, HFC, and HRTOD, with indications of the temporal events relevant to the HRTOD.*

pre-onset portion of the signal. The deadline at 18 ms was selected by considering a maximum reporting time of 20 ms for possible operations computed on such portion of the signal, which could take up to 2 ms (considering for instance real-time feature extraction, application of machine learning techniques, and repurposing of the analyzed sound). Specifically, this amount was justified by the results of the evaluation of the smart cajón prototype presented in [19]. These showed that a measured average latency of 20 ms between action and electronically generated sounds was deemed to be imperceivable by four professional cajón players. This was likely due to a masking effect in the attack of the acoustic sound that superimposes over the digital one.

3.2. Results

Table 3 presents the results of the application of the developed HRTOD to the dataset using the parameters for TBM reported in Table 2, and the two deadlines of 11.6 and 18 ms. For each instrument and for the whole dataset, we computed the number of hits detected by HFC, the number of hits affected by the temporal accuracy improvement of TBM, along with their percentage, their average improvement, and the maximum improvement. It is worth noticing that in calculating the improved performances of HRTOD compared to HFC we compared each onset time reported by HRTOD against the time reported by HFC minus 5.8 ms (this would be indeed the minimum time employed by HFC to report an onset after its actual occurrence given the 256-point window).

Table 3 also offers a comparison of the performances of HRTOD for the two deadlines by calculating their difference along the investigated metrics.

4. DISCUSSION

The first noticeable result emerging from Table 3 is that HRTOD effectively improved the temporal accuracy of HFC for all instruments and for both the investigated deadlines. The variability of HFC was drastically reduced since about 50% of the hits of the dataset were effectively improved for both the deadlines involved, with an average improvement of about 3 ms and maximum one of about 12 ms. Bongo was found to be the instrument most improved in terms of percentage of improved hits, although the average improvement was the lowest compared to the other instruments. Considering both the number of improved hits and the amount of average and maximum improvement, the cajón was found the instrument most positively affected by our HRTOD.

Furthermore, the results show that the use of a longer deadline generally improves all the considered metrics. Almost the 5% of the total hits were improved between the two deadlines, which shows the variability of HFC (and of spectral-based methods in general). Such a variability might constitute an issue in certain real-time applications. Indeed an error of more than 12 ms, as found for some hits on conga, may be critical when attempting to analyze in real-time the corresponding sound and classify it against other hits detected with no delay. The achieved average improvement due to the longer deadline was less than 0.5 ms compared

Algorithm 1: Pseudocode of the fusion policy of the involved TBM and HFC onset detection techniques in the developed HRTOD.

```
Input: Input signal, deadline, TBM_estimated_error, HFC_window_time
Output: Time of the detected onset reported when the deadline is elapsed
TBM_detected ← TBM(input_signal)
HFC_detected ← HFC(input_signal)
if HFC_detected == true && TBM_detected == false then
    HFC_onset_time ← get_time(HFC_detected)
    for the next 30 ms ignore any TBM_detected == true
    sleep(deadline - HFC_window_time)
    onset_time ← set_time(HFC_onset_time - HFC_window_time)
    return onset_time
else
    if HFC_detected == false && TBM_detected == true then
        TBM_onset_time ← get_time(TBM_detected)
        sleep(HFC_window_time - TBM_estimated_error)
        if HFC_detected == true then
            HFC_onset_time ← get_time(HFC_detected)
            sleep(deadline - HFC_window_time
            onset_time ← set_time(HFC_onset_time - HFC_window_time)
            return onset_time
        else
            sleep(deadline - HFC_window_time + TBM_estimated_error)
            if HFC_detected == true then
                onset_time ← set_time(TBM_onset_time - TBM_estimated_error)
                return onset_time
```

Table 1: *Values of parameters of TBM and HFC utilized for each instrument. Legend: HP = high-pass, LP = low-pass, f_c = cutoff frequency.*

	TBM								**HFC**		
	HP f_c (Hz)	**LP 1 f_c (Hz)**	**LP 2 f_c (Hz)**	a **(samples)**	b **(samples)**	β	λ	α	**threshold**	**window (samples)**	**hop (samples)**
Conga	4000	25	25	62	2	6e-09	0.8	0.8	0.2	256	64
Djembe	7500	25	25	62	2	7e-09	0.8	0.8	0.2	256	64
Cajón	7500	25	25	62	2	2e-09	0.8	0.8	0.2	256	64
Bongo	7500	25	25	62	2	2e-08	0.8	0.8	0.2	256	64

to the shorter one, but the maximum improvement was found to be more than 7 ms. The instrument that was mostly affected by such increment in the duration of the deadline was the cajón, while bongo was basically unaffected. This shows that for certain instruments a short deadline may be sufficient in capturing reliably the physical onset time of almost all hits.

Despite these encouraging results, it should be noticed that there are still margins for improvement as the method is affected by errors: as shown in the last two columns of Table 2, about the 75% of the hits would have needed a larger value for the TBM error estimate parameter. According to the analysis on the 400 annotated hits, the average error is below 2 ms but the maximum one could amount to about 11 ms. On a different vein, it is also worth noticing that the proposed method is context-dependent as it was built and tested by exploiting knowledge on the input signals investigated.

Although the algorithm has been conceived for real-time purposes, it can be applied to offline contexts as well. Offline algorithms have a number of advantages compared to real-time methods that might be exploited to refine the HRTOD here proposed. For instance, one could consider portions of the signal in the future, apply normalizations, use post-processing techniques, or utilize buffers larger than those here involved. A more timely accurate onset detector might have important implications not only for the design of musical instruments such as the smart ones [20], but also for automatic music transcription tasks [1], including those operating in real-time (see e.g., [11, 12]). Moreover, another application domain of the temporal accuracy improvements produced by the proposed method may be that of computational auditory scene analysis [33]. Although the sounds involved in this study belonged to the category of percussive non-pitched instruments, the method is expected to work well on several other categories of sounds (including the non musical ones as for instance footstep sounds, which have clearly discernible temporal characteristics like the sounds of percussive instruments [34]).

5. CONCLUSIONS AND FUTURE WORK

This paper proposed a real-time method to improve the temporal accuracy of state-of-the-art onset detectors. The study focused

Table 2: *Results of the analysis conducted on 100 annotated onsets for each instrument to determine the value of TBM estimated error, the expected average and maximum error of HRTOD.*

	mean±std err (ms)	**min (ms)**	**max (ms)**	**1st quartile (ms)**	**TBM estimated error (ms)**	**max error on 1st quartile (ms)**	**HRTOD mean error (ms)**	**HRTOD max error (ms)**
Conga	2.03±0.13	0.5	7	1	1	0.5	1.03	6
Djembe	1.7±0.14	0.5	7	1	1	0.5	0.7	6
Cajón	3.45±0.2	0.5	13	2	1.5	1	1.95	11.05
Bongo	2.98±0.09	1	6	2	2	1	1.98	4

Table 3: *Results of the proposed HRTOD involving the two deadlines and their differences.*

deadline (ms)	**instrument**	**# hits**	**# improved**	**% improved**	**mean improvement ± standard error (ms)**	**max improvement (ms)**
11.6	Conga	916	292	31.87	2.78±0.06	4.94
	Djembe	485	183	37.73	2.7±0.06	4.94
	Cajón	1094	532	48.62	3.7±0.05	4.94
	Bongo	965	643	66.63	2.02±0.04	4.94
	Total	**3460**	**1650**	**47.68**	**2.77±0.03**	**4.94**
18	Conga	916	325	35.48	3.33±0.11	12.2
	Djembe	485	200	41.23	3.1±0.11	10.75
	Cajón	1094	646	59.04	4.26±0.06	9.83
	Bongo	965	644	66.73	2.03±0.04	4.94
	Total	**3460**	**1815**	**52.45**	**3.17±0.04**	**12.2**
Difference	Conga	0	33	3.61	0.55	7.26
	Djembe	0	17	3.5	0.4	5.81
	Cajón	0	114	10.42	0.56	4.89
	Bongo	0	1	0.1	0.01	0
	Total	**0**	**165**	**4.77**	**0.38±0.12**	**7.26**

on percussive non-pitched sounds and for this purpose the spectral technique based on the high frequency content [28] was employed, which was reported in the literature to work the best for this type of sounds [13]. Experimental validation showed that the proposed approach was effective in better retrieving the physical onset time of about 50% of the hits in a dataset of four percussive non-pitched instruments compared to the performance of the onset detector based on high frequency content. The proposed method was inspired to hard real-time operating systems, which aim to guarantee that a task is accomplished at certain deadline. Our results revealed that the use of a longer deadline may capture better the variability of the spectral method (but at the cost of a bigger latency). Indeed, about 5% of the hits of the whole dataset could not be improved by involving a shorter deadline, although not all instruments were affected equally by a longer deadline.

The proposed method is expected to extend to sounds from other musical instruments as well as to non-musical sounds. Several directions for future work can be explored. Firstly, we plan to involve the proposed HRTOD in the development of percussive smart instruments such as the smart cajón reported in [19]. Secondly, future work will include experimenting with other types of data, in particular sounds from pitched instruments. An open question is whether the method would work for polyphonic pitched percussive instruments, where there can be one or more onsets roughly produced at the same time. Another future direction consists in exploring the performance of the proposed onset detector in noisy or multi-source environments, where for instance pitched onsets might be present. Finally, concerning context-awareness, it would be interesting to investigate whether the concepts presented in this study can be generalized to a more "blind" scenario.

The dataset involved in this study, the corresponding annotations, and the Pure Data source code are available online[7].

6. ACKNOWLEDGMENTS

Luca Turchet acknowledges supports from a Marie-Curie Individual fellowship of the European Union's Horizon 2020 research and innovation programme (grant nr. 749561).

7. REFERENCES

[1] E. Benetos, S. Dixon, D. Giannoulis, H. Kirchhoff, and A. Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, 2013.

[2] X. Zhang and W.R Zbigniew, "Analysis of sound features for music timbre recognition," in *IEEE International Conference on Multimedia and Ubiquitous Engineering*. IEEE, 2007, pp. 3–8.

[3] M. Barthet, P. Depalle, R. Kronland-Martinet, and S. Ystad, "Acoustical correlates of timbre and expressiveness in clarinet performance," *Music Perception: An Interdisciplinary Journal*, vol. 28, no. 2, pp. 135–154, 2010.

[4] K. Jathal, "Real-time timbre classification for tabletop hand drumming," *Computer Music Journal*, vol. 41, no. 2, pp. 38–51, 2017.

[7] https://github.com/lucaturchet

[5] 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.

[6] S. Dixon, "Onset detection revisited," in *Proceedings of the International Conference on Digital Audio Effects*, 2006, vol. 120, pp. 133–137.

[7] M. Tian, G. Fazekas, D. Black, and M.B Sandler, "Design and evaluation of onset detectors using different fusion policies," in *Proceedings of International Society for Music Information Retrieval Conference*, 2014, pp. 631–636.

[8] P. Brossier, J.P. Bello, and M.D Plumbley, "Real-time temporal segmentation of note objects in music signals," in *Proceedings of the International Computer Music Conference*, 2004.

[9] D. Stowell and M. Plumbley, "Adaptive whitening for improved real-time audio onset detection," in *Proceedings of the International Computer Music Conference*, 2007, pp. 312–319.

[10] S. Böck, F. Krebs, and M. Schedl, "Evaluating the online capabilities of onset detection methods," in *Proceedings of International Society for Music Information Retrieval Conference*, 2012, pp. 49–54.

[11] M. Miron, M.E.P. Davies, and F. Gouyon, "An open-source drum transcription system for pure data and max msp," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 221–225.

[12] M. Miron, M.E.P. Davies, and F. Gouyon, "Improving the real-time performance of a causal audio drum transcription system," in *Proceedings of the Sound and Music Computing Conference*, 2013, pp. 402–407.

[13] P. Brossier, *Automatic annotation of musical audio for interactive systems*, Ph.D. thesis, Queen Mary University of London, 2006.

[14] J. Vos and R. Rasch, "The perceptual onset of musical tones," *Perception & psychophysics*, vol. 29, no. 4, pp. 323–335, 1981.

[15] M. McKinney and J. Breebaart, "Features for audio and music classification," in *Proceedings of International Society for Music Information Retrieval Conference*, 2003, pp. 151–158.

[16] W. Brent, "Cepstral analysis tools for percussive timbre identification," in *Proceedings of the International Pure Data Convention*, 2009.

[17] W. Brent, "A timbre analysis and classification toolkit for pure data," in *Proceedings of the International Computer Music Conference*, 2010.

[18] C. Rosão, R. Ribeiro, and D.M de Matos, "Comparing onset detection methods based on spectral features," in *Proceedings of the Workshop on Open Source and Design of Communication*. ACM, 2012, pp. 71–78.

[19] L. Turchet, A. McPherson, and M. Barthet, "Co-design of a Smart Cajón," *Journal of the Audio Engineering Society*, vol. 66, no. 4, pp. 220–230, 2018.

[20] L. Turchet, A. McPherson, and C. Fischione, "Smart Instruments: Towards an Ecosystem of Interoperable Devices Connecting Performers and Audiences," in *Proceedings of the Sound and Music Computing Conference*, 2016, pp. 498–505.

[21] G.C Buttazzo, *Hard real-time computing systems: predictable scheduling algorithms and applications*, vol. 24, Springer Science & Business Media, 2011.

[22] S. Böck, A. Arzt, F. Krebs, and M. Schedl, "Online real-time onset detection with recurrent neural networks," in *Proceedings of the International Conference on Digital Audio Effects*, 2012.

[23] B.C.J Moore, *An introduction to the psychology of hearing*, Brill, 2012.

[24] A. McPherson and V. Zappi, "An environment for Submillisecond-Latency audio and sensor processing on BeagleBone black," in *Audio Engineering Society Convention 138*. 2015, Audio Engineering Society.

[25] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 1999, vol. 6, pp. 3089–3092.

[26] S. Herzog, "Efficient dsp implementation of median filtering for real-time audio noise reduction," in *Proceedings of the international conference on Digital Audio Effects*, 2013, pp. 1–6.

[27] M.S. Puckette, T. Apel, and D.D Ziccarelli, "Real-time audio analysis tools for pd and msp," in *Proceedings of the International Computer Music Conference*, 1998.

[28] P. Masri, *Computer modelling of sound for transformation and synthesis of musical signals*, Ph.D. thesis, University of Bristol, Department of Electrical and Electronic Engineering, 1996.

[29] C. Duxbury, J.P. Bello, M. Davies, and M.B Sandler, "Complex domain onset detection for musical signals," in *Proceedings of the Digital Audio Effects Conference*, 2003, pp. 1–4.

[30] J.P. Bello and M.B Sandler, "Phase-based note onset detection for music signals," in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 2003, vol. 5, pp. 441–444.

[31] J. Foote and S. Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *Proceedings of IEEE International Conference on Multimedia and Expo*. IEEE, 2001, pp. 881–884.

[32] S. Hainsworth and M. Macleod, "Onset detection in musical audio signals," in *Proceedings of the International Computer Music Conference*, 2003.

[33] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M.D Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[34] L. Turchet, "Footstep sounds synthesis: design, implementation, and evaluation of foot-floor interactions, surface materials, shoe types, and walkers' features," *Applied Acoustics*, vol. 107, pp. 46–68, 2016.