Valero-Mas and Iñesta *EURASIP Journal on Audio, Speech, and Music Processing* (2017) 2017:15
DOI 10.1186/s13636-017-0111-7

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Interactive user correction of automatically detected onsets: approach and evaluation

Jose J. Valero-Mas* and José M. Iñesta

Abstract

Onset detection still has room for improvement, especially when dealing with polyphonic music signals. For certain purposes in which the correctness of the result is a must, user intervention is hence required to correct the mistakes performed by the detection algorithm. In such interactive paradigm, the exactitude of the detection can be guaranteed at the expense of user's work, being the effort required to accomplish the task, the value that has to be both quantified and reduced. The present work studies the idea of interactive onset detection and proposes a methodology for assessing the user's workload, as well as a set of interactive schemes for reducing such workload when carrying out this detection task. Results show that the evaluation strategy proposed is able to quantitatively assess the invested user effort. Also, the presented interactive schemes significantly facilitate the correction task compared with the manual annotation.

Keywords: User interaction, Onset detection, Information retrieval, Audio analysis

Music signals may be decomposed into sound objects by means of signal processing techniques. Note events constitute an example of musical signal segmentation, and can be defined by both the moment the note starts—the onset—and its end—the offset [1]. *Onset detection*, defined as the automatic estimation of the starting points of note events in audio signals [2], has been of large interest to the Music Information Retrieval (MIR) community. Research areas such as *tempo and meter estimation* [3], *automatic music transcription* [4], *audio transformations* [5], or *real-time accompaniment* [6] often make use of onset information as a key part in their analysis process.

Due to that relevance, a considerable amount of effort has been made over the years to develop and improve onset detection algorithms. In this sense, although onset detection methods have typically addressed particular instrumentation cases [7], recent research outcomes have shown significantly precise results not limited to specific timbres [8, 9]. To check the performance of current state-of-the-art onset detection methods, the reader is referred to the results obtained in the annual Music Information Retrieval Evaluation eXchange (MIREX) contest.

The results obtained by current state of the art may be considered sufficiently accurate for applications such as *audio structure analysis* or *digital audio effects*, in which onset information simply constitutes a support information for the task rather than its main description. Nevertheless, for specific cases as *note tracking* in *automatic music transcription*, the preciseness of onset events remarkably influences the overall success of the task.

Note that, while onset estimation is generally used as an intermediate process within more complex MIR systems, this task may be also considered as a goal by itself. As an example, the work in [10] contemplates the use of onset information for identifying music pieces by comparing timing deviations between estimated onsets from interpretations of the pieces and its reference annotations from the scores.

The aforementioned cases constitute particular examples in which very precise onset times are required. Generally, research in such cases implies the manual annotation of corpora since no single onset estimation

*Correspondence: jjvalero@dlsi.ua.es
Pattern Recognition and Artificial Intelligence Group, Department of Software and Computer Systems, University of Alicante, Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain

algorithm guarantees a perfect retrieval. Whereas this performance limitation is inherent to any research topic, some authors in the MIR community suggest that a *glass ceiling* is being reached, at least in the case of some commonly addressed tasks [11, 12]. It thus appears interesting to explore alternative research paradigms that are capable of dealing with these limitations.

The so-called interactive paradigm and the efficient exploitation of *human expert knowledge* in the context of adaptive systems stands as a promising alternative to manual annotations [13]. Note that, within an interactive scheme, the correctness of the system output is no longer the main issue to assess. The challenge now is the development and creation of interactive schemes capable of efficiently exploiting the human feedback in the system in order to eventually reduce the user's workload [14]. While acknowledging that such interactive schemes are not applicable to the massive analysis of audio pieces, these methodologies should also contribute to further research into stand-alone onset detectors by facilitating the provision of annotated corpora, rather than the usual hand-made annotation.

A clear example of interactivity applied to MIR which concerned user-aided monotimbral polyphonic music transcription was reported in [15]. The authors proposed an interactive music transcription algorithm that allows the correction of the note onsets: a user interaction at a certain time point implicitly validates all the previous output (in time) and that user information is then employed by the system to adapt and recompute the output from that point on. However, although a qualitative improvement in the results was observed, there was still a need for a quantitative assessment.

This paper expands the aforementioned idea of interactive onset detection systems. The main contributions of this work are a) the proposal of a methodology and a set of measures for the assessment of the user workload in interactive systems; b) two interactive schemes capable of gathering information from each user correction in order to adapt the detection parameters so as to reduce the correction workload; and c) a thorough assessment of the interactive schemes using the methodology proposed.

The rest of the paper is organised as follows. Section 1 introduces the basis of onset detection; Section 2 describes the interaction methodologies proposed; Section 3 shows the evaluation methodology followed; Section 4 analyses the results obtained; and finally, Section 5 presents the conclusions and discusses about possible future work.

1 Introduction

Generally, every onset detection algorithm based on signal processing comprises two different stages: an initial phase, known as the *onset detection* or *novelty function*, and an *onset selection* stage that is employed to identify the onsets by using the output from the first step [16, 17]. Figure 1 shows this idea.

The objective of the *onset detection* or *novelty function* stage (ODF) is to compute a time series $O(t)$ from the initial audio stream, whose peaks represent the estimated position of the each single onset event in the signal. This representation is the result of a certain analysis process that measures the changes in one or more audio features. Characteristics typically considered in the literature comprise signal energy [18, 19], pitch [20], phase [21, 22], or even combinations of the previous three [23, 24].

The *onset selection* stage, commonly referred to as *Onset Selection Function* (OSF), selects the points (frames) in $O(t)$ detected as onsets. Its output is a sorted list of elements $\left(\hat{o}_i\right)_{i=1}^{L}$ representing the time positions of the estimated onsets.

A proper ODF process derives a function whose peaks represent potential onsets of the signal, while the OSF conceptually aims at discriminating peaks which represent onsets from spurious or noisy estimations. In that sense, the most straight-forward OSF approach is to search for local maxima of $O(t)$ above a global threshold value for discarding the spurious values [2].

Other methods consider the use of adaptive threshold functions for dealing with local changes in the signal as, for instance, alterations in dynamics. A commonly used technique is setting as threshold the mean or median value of $O(t)$ in a certain time lapse around the point under evaluation [25].

It is also important to highlight some works which use both supervised and unsupervised machine learning techniques in this context. Some examples are the use of recurrent neural networks [26] or clustering [27] to automatically estimate the most suitable and robust OSF for a set of data, or even the use of end-to-end systems based on convolutional neural networks (CNNs) [28] for directly integrating both stages into a single classification scheme.

In our case, we shall use a representative set of ODF and OSF techniques for assessing the usefulness of the interactive methodology proposed in this work. The selected methods will be introduced in Section 3 as it explains the evaluation methodology considered.

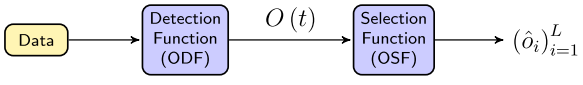

Fig. 1 Block diagram commonly used for onset detection

2 User interaction for onset detection

As aforementioned, onset detection algorithms rarely retrieve a perfect result in terms of precision. The two types of error which affect this performance are a) the algorithm misses onsets which should be detected (false negatives, FN) and b) the algorithm detects onsets than do not actually exist (false positives, FP).

Let N_{GT} denote the total number of onsets to be annotated in an audio file (ground truth). Let also be N_{OK} the number of correctly detected onsets, N_{FP} the amount of FP errors committed and N_{FN} the number of FN errors once the signal has been processed with an onset detection algorithm.

The amount of onsets obtained by a detection algorithm may be defined as $N_D = N_{OK} + N_{FP}$ whereas the total number of onsets to be estimated can be expressed as $N_{GT} = N_{OK} + N_{FN}$. Therefore, a user starting from the initial N_D analysis should manually eliminate the N_{FP} erroneous estimations and annotate the N_{FN} missed onsets, thus requiring a total of $C_T = N_{FP} + N_{FN}$ corrections to obtain the correct annotation.

User interaction, meaning that the system attempts to adjust its performance from what the user corrects, is proposed in order to reduce C_T. The idea is that the total number of corrections performed in an interactive system C_T^{int} is lower than, or in the worst-case scenario, equal to, the amount required in a complete manual correction C_T^{man}, i.e. $C_T^{int} \leq C_T^{man}$.

In a practical sense, the user interaction should *adapt* the system by changing the set of parameters involved in the ODF and/or OSF processes. Due to the influence of the OSF stage in the overall onset detection process [25], we assume that the detection errors are exclusively produced by considering an inappropriate configuration of this selection function. Although this constitutes a simplification, we restrict our work to this hypothesis.

The different interaction methodologies considered in this work are now introduced. Additionally, a set of measures for quantitatively assessing the user effort is also proposed.

2.1 Interaction methodologies

The premise behind these interactive methodologies is that the OSF process may not be properly parameterised: a particular OSF configuration may not be suitable for the entire $O(t)$ due to factors as, for instance, changes in instrumentation, dynamics, articulation, and so on. Thus, a given ODF should be examined by an OSF particularly tuned for different regions. These regions would be defined by the user as the FP and FN errors are pointed out, and the new local OSF parameters are estimated through the interactions.

In our case, as OSF we will restrict ourselves to variations of the strategy of finding local maxima above or equal to a certain threshold θ in the onset function $O(t)$. In that scheme, time frame t_i contains an onset if the following conditions are fulfilled:

$$\begin{aligned} O(t_{i-1}) < O(t_i) > O(t_{i+1}) \\ O(t_i) \geq \theta \end{aligned} \tag{1}$$

The idea is that, while the local maximum condition is kept unaltered, threshold θ now becomes a function $\theta \equiv \theta(t)$ whose value is defined according to one of the interactive policies to be explained.

Given that user interactions may not match the actual maxima in the ODF, the system needs to provide a particular tolerance window. Thus, given an interaction at t_{int}, the energy value retrieved from the ODF for the adaptation process is given by:

$$O(t_{int}) \equiv \max\{O(t_m)\} \text{ with } t_m \in [t_{int} - W_T,\ t_{int} + W_T] \tag{2}$$

where W_T represents the tolerance window considered. We consider a tolerance window of $W_T = 30\ ms$ since, as pointed out by [29], this time threshold represents a proper tolerance for human beings to perceive onsets.

Exceptionally, Eq. 2 may retrieve a value $O(t_{int}) = 0$ in the tolerance time lapse. This issue occurs when the ODF process has not obtained a proper $O(t)$ representation and some onsets are not represented by a peak in this function. In those cases, the correction is performed (the onset is added) but the threshold value is kept unaltered.

Finally, given the time dependency in the output of an onset detection algorithm, when the user interacts at position t_{int} of the $O(t)$, all information located at time frames $t < t_{int}$ is implicitly validated. Corrections are therefore only required in time frames $t \geq t_{int}$. This assumption of left-to-right correction is commonly considered in works involving data of sequential nature such as in interactive machine translation (IMT) or interactive speech transcription (IST) [30, 31]

The two policies proposed for propagating the effects of an interaction are described next.

2.1.1 *Threshold-based interaction*

This policy bases its performance on directly modifying the threshold value θ. This technique was already presented in [15] for interactive computer-user correction in polyphonic transcription.

In this case, the global threshold is substituted by an initial (*static*) proposal θ_0, and whenever the user interacts with an onset o_{int} (either an FP or an FN) located at a time frame t_{int}, its energy $O(t_{int})$ is retrieved. This value,

once modified by a small value ϵ compared to the variation range in $O(t)$, becomes the new threshold θ_{int} for the new detection process that will be performed for $t \geq t_{int}$:

$$\theta_{int} = \begin{cases} O(t_{int}) - \epsilon & \text{if} \quad o_{int} \notin \left(\hat{o}_i\right)_{i=1}^{L} \quad \text{(FN)} \\ O(t_{int}) + \epsilon & \text{if} \quad o_{int} \in \left(\hat{o}_i\right)_{i=1}^{L} \quad \text{(FP)} \end{cases} \tag{3}$$

where ϵ has been set to 0.001 for this work, as it constitutes a value an order of magnitude lower than the sensibility considered for the $O(t)$ functions. Additionally, ϵ is not relative to the range of $O(t)$ since, as explained in Section 3.1.1, these functions are normalised so that $O(t) \in [0, 1]$.

Figure 2 shows an example of the threshold variation as a result of the different interactions performed by the user.

2.1.2 Percentile-based interaction

This second approach is inspired by the idea of using an adaptive threshold for assessing the ODF. As previously introduced, a typical method for doing so consists of using an analysis window around the target point in $O(t)$ and setting as the, now variable, threshold $\theta(t)$ the median value of the window [25].

In our case, instead of using the median value of the sample distribution, we find useful the use of other percentiles for setting the threshold. The idea is that when the user performs an interaction at time frame t_{int}, its energy $O(t_{int})$ is retrieved for calculating the n_{th} percentile it represents with respect to the elements contained in a W-length window around that point, i.e.,

$$n_{th} | P_{n_{th}} \left\{ O\left(t_{w_{int}}\right) \right\} = O\left(t_{int}\right) \text{ with } t_{w_{int}} \in \left[t_{int} - \frac{W}{2},\ t_{int} + \frac{W}{2} \right] \tag{4}$$

where $P_{n_{th}}\{x\}$ obtains the value representing the n_{th} percentile of sample distribution x.

Then, for calculating threshold $\theta(t_i)$ for time positions $t \geq t_{int}$, the rest of the signal is evaluated with a W-length sliding window using the percentile index n_{th} obtained at the interaction point t_{int} as it follows:

$$\theta(t_i) = P_{n_{th}} \left\{ O(t_{w_i}) \right\} \text{ with } t_{w_i} \in \left[t_i - \frac{W}{2},\ t_i + \frac{W}{2} \right] \text{ and } t_i \in t \geq t_{int} \tag{5}$$

Conceptually, the premise of using this approach is that, when a correction at t_{int} is made, the particular threshold θ value is not relevant by itself but by its relation with the surrounding values. For example, if $O(t_{int})$ is a low value compared to the elements in the surrounding W-length window, the successive analysis windows should use low θ values as well, which can be obtained by using low percentiles. On the other hand, if $O(t_{int})$ is high compared to the surrounding elements, the percentile should be high. Ideally, this approach should adapt the performance of the OSF to the particularities of the ODF.

The duration of the W-length window has been set to cover 1.5 s, using as a reference the work by [32] in which windows ranging from 1 to 2 s were used.

Figure 3 graphically shows the evolution of threshold θ when using this approach.

2.2 User effort assessment

Having introduced the interactive correction methodologies, it is necessary to define some indicators able to quantitatively assess the user effort invested in the onset correction process.

In these measures, we assume the effort is represented by the amount of corrections C_T the user needs to perform. As previously commented, the intuitive idea is that an interactive scheme should require less or, in the worst-case scenario, the same effort than a complete manual correction, i.e., $C_T^{int} \leq C_T^{man}$. However, we find it necessary to quantify and formalise this idea so that future methodologies can be objectively compared.

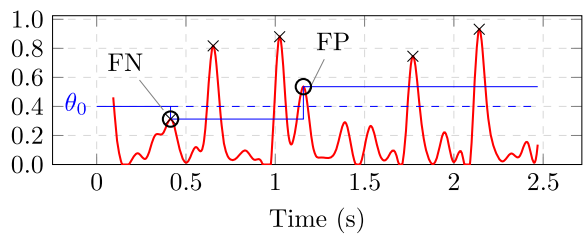

Fig. 2 Evolution of θ throughout time as the result of user interactions in the threshold-based approach: symbol (⊗) shows the ground truth onsets while (◯) represents the performed interactions. *Dashed* and *solid lines* represent the static (θ_0) and interactive thresholds, respectively

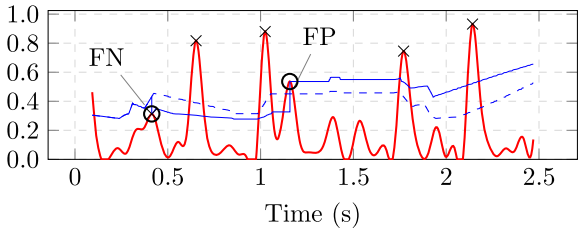

Fig. 3 Evolution of $\theta(t)$ throughout time as the result of user interaction in the sliding window percentile-based approach: symbol (⊗) shows the ground truth onsets while (◯) represents the performed interactions. *Dashed* and *solid lines* represent the static and interactive thresholds obtained with the sliding window approach, respectively. Initial percentile $\theta(t_i = 0)$ has been set to 50th (median value)

In the following sections, we introduce the two proposed measures for assessing the user effort invested in the correction process.

2.2.1 Total corrections ratio

The first of the two proposed metrics is the *Total corrections ratio*, R_{TC}. The idea behind this measure is comparing the amount of corrections a user needs to perform when using an interactive system $\left(C_T^{int}\right)$ to a manual correction $\left(C_T^{man}\right)$. This ratio is obtained as:

$$R_{TC} = \frac{C_T^{int}}{C_T^{man}} = \frac{N_{FP}^{int} + N_{FN}^{int}}{N_{FP}^{man} + N_{FN}^{man}} \tag{6}$$

Depending on the resulting ratio value, it is possible to assert whether the interactive scheme reduces the workload:

$$R_{TC} \begin{cases} > 1 \Rightarrow \textbf{Increasing} \text{ workload} \\ = 1 \Rightarrow \textbf{No} \text{ difference} \\ < 1 \Rightarrow \textbf{Decreasing} \text{ workload} \end{cases}$$

2.2.2 Corrections to ground truth ratio

Although the previous metric is able to assess whether an interactive scheme requires less effort than a manual correction, a certain premise is being assumed: an automatic onset detection stage reduces the annotation workload since it tracks, at least, part of the elements that must be annotated.

However, it is possible that the automatic detection algorithm will not be able to perform this task as expected (for instance, when dealing with a noisy signal). In such cases, the number of correctly tracked onsets N_{OK} may be negligible, or even non-existing, thus leading to $N_D = N_{OK} + N_{FP} \approx N_{FP}$. The user would be required to annotate the total number of onsets N_{GT} plus eliminating the N_{FP} errors committed, i.e., $C_T = N_{GT} + N_{FP} = N_{OK} + N_{FN} + N_{FP}$. Under these circumstances, it would be arguable that the need for an initial onset detection as the manual annotation of the signal from scratch would imply less workload.

To cope with this issue, the *corrections to ground truth ratio*, R_{GT}, compares the amount of interactions required C_T in relation to the total amount of ground truth onsets N_{GT} for both interactive systems (Eq. 7) and manual corrections (Eq. 8).

$$R_{GT}^{int} = \frac{C_T^{int}}{N_{GT}} = \frac{N_{FP}^{int} + N_{FN}^{int}}{N_{GT}} = \frac{N_{FP}^{int} + N_{FN}^{int}}{N_{OK} + N_{FN}} \tag{7}$$

$$R_{GT}^{man} = \frac{C_T^{man}}{N_{GT}} = \frac{N_{FP}^{man} + N_{FN}^{man}}{N_{GT}} = \frac{N_{FP}^{man} + N_{FN}^{man}}{N_{OK} + N_{FN}} \tag{8}$$

Bearing in mind that a ratio of 1 is equivalent to manually annotating all the onsets, the results depict whether the system forces the user to make more corrections than without any initial detection, thus making the system useless in practice:

$$R_{GT} \begin{cases} > 1 \Rightarrow \textbf{More} \text{ than manual} \\ = 1 \Rightarrow \textbf{Same} \text{ as manual} \\ < 1 \Rightarrow \textbf{Less} \text{ than manual} \end{cases}$$

Finally, it must be pointed out the existing relation among measures R_{GT}^{int} (Eq. 7) and R_{GT}^{man} (Eq. 8) with measure R_{TC} (Eq. 6) by using the following expression:

$$R_{TC} = \frac{R_{GT}^{int}}{R_{GT}^{man}} = \frac{N_{FP}^{int} + N_{FN}^{int}}{N_{FP}^{man} + N_{FN}^{man}} \tag{9}$$

3 Evaluation methodology

In order to assess the proposed interactive strategies, the scheme shown in Fig. 4 has been implemented. First of all, the input data is processed by an *initial onset detection* algorithm (an ODF method that retrieves an $O(t)$ function and a OSF algorithm that processes it) retrieving a list of estimated onsets $\left(\hat{o}_i\right)_{i=1}^{L}$; both the $O(t)$ signal and the estimations $\left(\hat{o}_i\right)_{i=1}^{L}$ are the input to the *user interaction* process. In that last stage, the user validates and interactively corrects those estimations.

In order to avoid the need for a person to manually carry out the corrections, ground truth annotations were used to automate the process as in other works addressing interactive methodologies [30].

We shall now describe the different onset detection algorithms, datasets, and performance metrics considered for assessing our proposal.

3.1 Initial onset detection

This section introduces the different ODF and OSF strategies considered for the evaluation of the work.

3.1.1 Onset detection functions

The considered ODF methods cover the different principles and methodologies introduced in Section 1 with the aim of exhaustively assessing the behavior of the proposed interactive methodologies with different analysis principles.

We now introduce and describe the different functions considered:

1. **Sum of Magnitudes (SM)**: This first approach bases its performance on measuring changes in the energy of the signal. Using the magnitude part of the spectrogram of the signal, this process estimates the energy for each analysis window as the sum of the magnitude component of each frequency bin [33].
2. **Power Spectrum (PS)**: This second method also bases its performance on measuring changes in energy. The approach is identical to the previous one

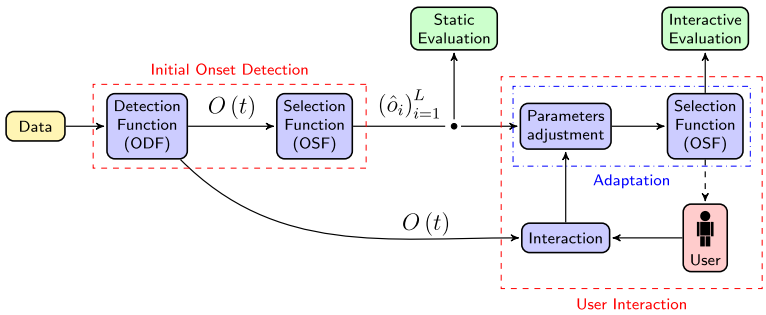

Fig. 4 Block diagram of the proposed scheme: an initial onset detection is performed on the input signal (*Data*) in the *initial onset detection* block; *static evaluation* assesses the performance of the stand-alone algorithm; the *user interaction* block introduces human verification, interaction and correction; *interactive evaluation* assesses the performance of the interactive scheme

but performing the sum of the squared value of the magnitude components of the spectrogram [33].

3. **Semitone Filter Bank (SFB)**: This energy-based approach analyses the evolution of the magnitude spectrogram assuming a harmonic sound is being processed. The algorithm applies a harmonic semitone filter bank to each analysis window of the magnitude spectrogram and retrieves the energy of each band (root mean square value); then, consecutive semitone bands in time are substracted to find energy differences; negative results are filtered out as only energy increases may point out onset information; finally, all bands are summed to finally obtain the detection function [19].
4. **Phase Deviation (PD)**: This method relies exclusively on phase information. The idea is that discontinuities in the phase component of the spectrogram may depict onsets. With that premise, this approach basically predicts what the value of the phase component of the current frame should be using the information from previous frames; the deviation between that prediction and the actual value of the phase spectrum models this function [23].
5. **Weighted Phase Deviation (WPD)**: A major flaw in the previous phase method is that it considers all frequency bins to have the same relevance in the prediction. This severely distorts the result as low energy components which should have no relevance in the process are considered equal to more relevant elements. In order to avoid that, each phase component is weighted by the correspondent magnitude spectrum value [34].
6. **Complex Domain Deviation (CDD)**: Extends the principle introduced in the *Phase Deviation* method by estimating both magnitude and phase components for the analysis window at issue using the two preceding frames and assuming steady-state behaviour with a complex domain representation. The difference between the prediction and the actual value of the frame defines the function [35].
7. **Rectified Complex Domain Deviation (RCDD)**: In the *Complex Domain Deviation* method no distinction in the type of deviation between the predicted spectrum and the one at issue is made. In such case, the algorithm does not distinguish between energy rises, which depict onsets, and energy decreases, which point out offsets. Hence, a slight modification based on half-wave rectification is performed on the method to avoid tracking offsets. The difference between predicted and real values is now carried out when the spectral bins increase their energy along time; in case the energy decreases, a zero is retrieved [34].
8. **Modified Kullback-Leibler Divergence (MKLD)**: This method also measures energy changes between consecutive analysis frames in the magnitude spectrum of the signal. The particularity of this approach lies in the use of the Kullback-Leibler divergence for measuring such changes, which allows tracking large energy variations while inhibiting small ones [36].
9. **Spectral Flux (SF)**: This approach depicts the presence of onsets by measuring the temporal evolution of the magnitude spectrogram of the signal. The idea is obtaining the bin-wise difference between the magnitude of two consecutive analysis windows and summing only the positive deviations for retrieving the detection function [37].
10. **SuperFlux (SuF)**: Modifies the *Spectral Flux* method by substituting the difference between consecutive analysis windows by a process of tracking spectral trajectories in the spectrum together with a maximum filtering process. This allows the suppression of vibrato articulations in the signal which generally tend to increase false detections in classic algorithms [38, 39].

Given the different principles in which the presented processes are based on, the resulting $O(t)$ functions may not span for the same range. Thus, a normalisation process is directly applied to the $O(t)$ time series once it has been obtained from the initial audio piece so that its spans in the range $[0, 1]$. This normalization is a Min-Max scaling applied as it follows:

$$O(t) = \frac{\hat{O}(t) - \min\{O(t)\}}{\max\{O(t)\} - \min\{O(t)\}}$$

where $\hat{O}(t)$ and $O(t)$, respectively, stand for the initial and normalized times series, and $\min\{\cdot\}$ and $\max\{\cdot\}$ represent the minimum and maximum operators, respectively.

Finally, since all these functions rely on a spectrogram representation obtained as a Short-Time Fourier Transform (STFT), we set the same analysis parameters to all of them: an analysis window size of 92.8 *ms* samples with a 50% of overlapping factor.

3.1.2 Onset Selection Functions

In order to process the considered detection functions, we have used different OSF methods. As in the case of the interactive methodologies (cf. Section 2.1), these methods are based on finding local maxima above, or equal to, a certain threshold θ. In line with those cases, the maximum condition will remain unaltered, being threshold θ the parameter to be set.

The two methods considered are:

1. **Global threshold**: The threshold θ is manually set as a user parameter to a value $\theta = \theta_o$ for analysing the entire $O(t)$ function.
2. **Sliding window with percentile index**: A W-length sliding window is used to analyse the detection function $O(t)$ and obtain a time-dependent function $\theta \equiv \theta(t)$ adapted to the particularities of $O(t)$. For analysing time frame t_i, we take the elements of $O(t)$ in the range $\left[t_i - \frac{W}{2},\ t_i + \frac{W}{2}\right]$ and we calculate a percentile value using that sample distribution with Eq. 5.[1] In this case, W has been also set to 1.5 seconds considering the results in [32].

In order to assess the influence of the parameterisation of the considered selection functions in the overall performance, 25 values equally distributed in the range $[0, 1]$ have been used as either threshold or normalised percentile index.

Finally, it must be pointed out that these selection functions are equivalent to the interactive policies in Section 2.1. This has been intentionally done as we want to assess two different configurations in this experimentation: on one hand using the same detection functions for both the static onset detection and the interactive scheme; on the other hand, using different detection functions for both parts.

3.2 Dataset

The dataset used for the evaluation is the one introduced in [29]. It comprises a set of 321 monaural real world recordings sampled at 44.1 kHz covering a wide range of timbres and polyphony degrees. The total duration of the set is 1 h and 42 min containing 27,774 onsets with an average duration of 19 s per file (the shortest lasts 1 s and the largest one extends up to 3 min) and an average figure of 87 onsets per file (minimum of three onsets and maximum of 1132 onsets).

However, as pointed out in [29], these precise annotations (raw onsets) do not necessarily represent human perceptions of onsets in spite of being musically correct. Thus, as this work addresses the human effort in the annotation/correction of onsets, the dataset was processed following the previous reference: all onsets within 30 *ms* were combined into one located at the arithmetic mean of their single positions. This process reduced the total number of elements to 25,996 onsets (approximately, 81 onsets per file).

A detailed description of the set in terms of the instrumentation and number of raw and combined onsets is shown in Table 1.

In our case, no particular partitioning in terms of instrumentation, duration, or polyphony degree has been done to the data, as the idea is to check the usefulness of the interactive approach disregarding the nature of the data used.

3.3 Performance measurement

In order to assess the proposed onset detection and correction strategies, we have considered two different sets of evaluation measures.

The set of metrics proposed in Section 2.2 will be used, as they actually aim at assessing the usefulness of the interactive schemes in terms of the user effort.

Table 1 Description of the dataset in terms of instrumentation used for evaluation. Reproduced from [29]

Instrumentation	Files	Raw onsets	Combined
Complex mixtures	193	21,091	19,492
Pitched percussive	60	2981	2795
Wind instruments	25	822	1376
Bowed strings	23	1180	820
Non-pitched percussive	17	1390	1177
Vocal	3	310	306
Total	321	27,774	25,996

Nevertheless, we also find necessary the use of measures evaluating the accuracy of static onset detectors. This way, we may relate the accuracy of the onset detection approaches considered and the effort required to correct the errors committed, either manually or within an interactive scheme.

3.3.1 Static metrics

For evaluating the goodness of the onset detectors, we have considered the three figures of merit commonly used in this context (e.g., works by [24] and [17]): Precision (P), Recall (R) and F-measure (F_1). Let N_{OK} be the amount of correct onsets detected by the algorithm and N_{FP} and N_{FN} the number of FP and FN errors committed, respectively. These measures can be defined as:

$$\begin{aligned} P &= \frac{N_{OK}}{N_{OK}+N_{FP}} \qquad R = \frac{N_{OK}}{N_{OK}+N_{FN}} \\ F_1 &= \frac{2\cdot P\cdot R}{P+R} = \frac{2\cdot N_{OK}}{2\cdot N_{OK}+N_{FP}+N_{FN}} \end{aligned} \tag{10}$$

As frequently commented on literature, the start of a musical event is not a specific point in time, but rather a time lapse known as a rise or transient time [40]. A certain point in this span must therefore be chosen as the onset. Given the reported variability in the notation of rhythmic aspects of music even among expert transcribers [41], it is assumed that onset annotation is highly dependent on the person, imprecise and difficult to generalise.

Owing to this *loose* definition, onset detection algorithms are given a certain time lapse in which the detection is considered to be correct. Most commonly, this acceptance window has been set to 50 *ms*, which is the same as the one used in the MIREX contest. Nevertheless, in this work we adopt a more restrictive tolerance window of 30 *ms* since, as pointed out by [29], this value represents a proper time lapse for human beings to be able to detect onsets.

Finally, it must be pointed out that this assessment does not consider doubled onsets (two detections for a single ground truth element) and merged onsets (one detection for two ground truth elements) as they constitute subsets of N_{FP} and N_{FN}, respectively.

3.3.2 User-centred metrics

In terms of user-centred metrics, as aforementioned, we shall restrict ourselves to the set of measures proposed in Section 2.2. As a reminder to the reader, the two metrics considered were a) *Total Corrections ratio* (R_{TC}) that compares the amount of corrections required for correcting a sequence under the interactive paradigm with respect to complete a manual correction; and b) *Corrections to Ground Truth ratio* (R_{GT}) which contrasts the total amount of interactions performed (either manually or in an interactive scheme) with the total number of onsets to be annotated.

4 Results

In this section, we present the results obtained when assessing our interactive proposals with the evaluation procedure described above. For each particular pair of onset detection and selection function plus either the manual correction or the interactive scheme at issue, the figure of merit shows the average and standard deviation of the 25 selection function configurations.

Results obtained in the static assessment of the considered onset detection algorithms are shown in Table 2. Additionally, Fig. 5 graphically shows the results obtained for the F-measure values for a better comprehension.

Figures achieved by the different configurations considered show the intrinsic difficulty of the dataset: focusing on the F-measure score, results are far from being perfect as all the scores are lower than 0.6. In that sense, the Phase Deviation method showed the lowest accuracy, possibly due to relying only on phase information and its reported disadvantage of considering all frequency bins equally relevant. Methods such as Semitone Filter Bank or SuperFlux showed good responses as, although mostly relying on an energy description of the signal, the information is

Table 2 Results obtained in terms of Precision, Recall, and F-measure for the static evaluation of the different detection (ODF) and selection (OSF) methods considered

ODF	OSF	Precision	Recall	F-measure
SFB	Threshold	0.82 ± 0.12	0.4 ± 0.2	0.5 ± 0.2
	Percentile	0.63 ± 0.10	0.64 ± 0.13	0.59 ± 0.07
PS	Threshold	0.69 ± 0.07	0.4 ± 0.2	0.4 ± 0.2
	Percentile	0.65 ± 0.06	0.57 ± 0.12	0.55 ± 0.08
SM	Threshold	0.66 ± 0.07	0.4 ± 0.2	0.4 ± 0.2
	Percentile	0.64 ± 0.06	0.56 ± 0.12	0.55 ± 0.08
CDD	Threshold	0.36 ± 0.03	0.18 ± 0.11	0.19 ± 0.10
	Percentile	0.33 ± 0.02	0.26 ± 0.06	0.26 ± 0.05
RCDD	Threshold	0.70 ± 0.08	0.4 ± 0.3	0.4 ± 0.2
	Percentile	0.63 ± 0.07	0.61 ± 0.13	0.57 ± 0.08
PD	Threshold	0.29 ± 0.02	0.17 ± 0.15	0.14 ± 0.11
	Percentile	0.35 ± 0.02	0.37 ± 0.10	0.32 ± 0.06
WPD	Threshold	0.66 ± 0.06	0.4 ± 0.2	0.4 ± 0.2
	Percentile	0.64 ± 0.06	0.56 ± 0.12	0.54 ± 0.08
MKLD	Threshold	0.45 ± 0.16	0.3 ± 0.3	0.2 ± 0.2
	Percentile	0.61 ± 0.07	0.63 ± 0.14	0.56 ± 0.08
SF	Threshold	0.53 ± 0.10	0.3 ± 0.2	0.30 ± 0.19
	Percentile	0.51 ± 0.08	0.55 ± 0.12	0.48 ± 0.06
SuF	Threshold	0.93 ± 0.08	0.3 ± 0.3	0.4 ± 0.2
	Percentile	0.67 ± 0.11	0.74 ± 0.15	0.64 ± 0.08

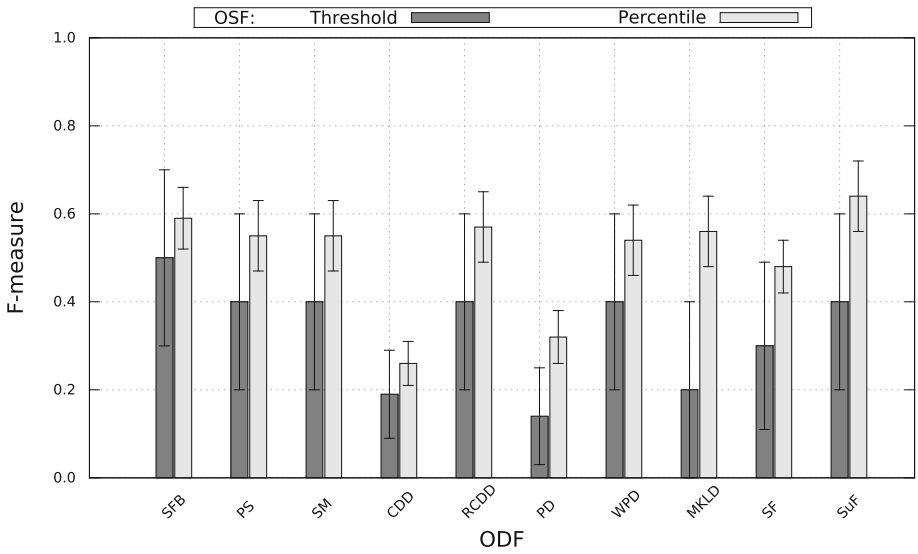

Fig. 5 Graphical representation of the results obtained in terms of F-measure for the static evaluation of the different detection (ODF) and selection (OSF) methods considered

processed in very sophisticated ways to avoid estimation errors.

In general terms, the relatively high precision scores achieved suggest that FP may not be the most common type of error in the considered systems. However, recall scores were low, especially when considering the global threshold selection process, thus pointing out a considerable amount of FN errors.

These results also show the clear advantage of adaptive threshold methods in the onset selection process when compared to a global initial value. In general, the former paradigm achieved better detection figures with lower deviation values than the latter, thus stating its robustness.

Once we have gained a general insight of the performance of the considered onset detection and selection schemes, we shall study them from the interactive point of view. Table 3 and Fig. 6 introduce the effort results in terms of the *Corrections to Ground Truth ratio* ($\mathrm{R_{GT}}$) measure when considering the manual and interactive corrections of the errors.

As an initial remark, it can be seen that the workload figures for manual correction (R_{GT}^{man}) are close to a value of 0.5 for all the ODF and OSF considered. These results suggest that an initial onset estimation process is indeed beneficial for lowering the manual annotation since such figures depict that half of the total number of onsets are properly handled by the autonomous detection system. The reported low deviation values also suggest that only for some particular cases in which the OSF parameters are badly selected, the required effort may be higher.

In terms of the threshold-based interaction scheme, there is a consistent workload reduction when compared to the manual procedure. Figures obtained are almost always under the 0.5 value, getting to the point of 0.26 for the SuF algorithm (which broadly means annotating just a fourth of the total number of onsets), showing the workload reduction capabilities of the scheme. Additionally, the very low standard deviation values obtained point out the robustness of the method: independently of the initial performance of the ODF and OSF at issue, the threshold-based interaction scheme performs consistently solves the task within a fixed figure of effort. This fact could suggest that, when considering this scheme, the performance of the initial onset estimation by the autonomous algorithm may not be completely relevant as the interactive scheme is able to solve the task within the same figure of effort.

Regarding the percentile-based scheme, the effort figures obtained are clearly worse than in the case of the threshold-based scheme, with up to 0.14 points of difference between the two schemes for this measure, and are qualitatively similar to the figures by the manual correction. This premise can be also seen in the deviation values obtained: in spite of being quite low, in some cases these figures show less consistency than in the threshold-based approach (e.g., SuF or RCDD); nevertheless, it should be noted that when compared to the manual correction, percentile-based interaction shows a superior robustness since for this scheme the deviation figures are consistently lower than those obtained when considering the manual approach.

Finally, the results obtained in terms of the *Total Corrections* ratio are shown in Table 4 and Fig. 7. This figure of merit helps us to compare the different interactive

Table 3 Comparison of the user effort invested in correcting the initial estimation of static onset detectors in terms of the R_{GT}. The F-measure column shows the performance of the static detection method, whereas R_{GT}^{man} refers to the effort invested when considering a complete manual correction of the results. R_{GT}^{thres} and R_{GT}^{pctl} stand for the user effort in the threshold-based and percentile-based correction approaches, respectively

ODF	OSF	F-measure	R_{GT}^{man}	R_{GT}^{thres}	R_{GT}^{pctl}
SFB	Threshold	0.5 ± 0.2	0.41 ± 0.05	0.34 ± 0.01†	0.43 ± 0.02
	Percentile	0.59 ± 0.07	0.45 ± 0.03	0.34 ± 0.01†	0.44 ± 0.01
PS	Threshold	0.4 ± 0.2	0.44 ± 0.03	0.37 ± 0.01†	0.43 ± 0.01
	Percentile	0.55 ± 0.08	0.44 ± 0.02	0.37 ± 0.01†	0.43 ± 0.01†
SM	Threshold	0.4 ± 0.2	0.45 ± 0.03	0.38 ± 0.01†	0.43 ± 0.01†
	Percentile	0.55 ± 0.08	0.45 ± 0.01	0.38 ± 0.01†	0.44 ± 0.01†
CDD	Threshold	0.19 ± 0.10	0.54 ± 0.04	0.51 ± 0.01†	0.57 ± 0.01†
	Percentile	0.26 ± 0.05	0.57 ± 0.02	0.52 ± 0.01†	0.57 ± 0.01†
RCDD	Threshold	0.4 ± 0.2	0.44 ± 0.04	0.35 ± 0.01†	0.44 ± 0.02
	Percentile	0.57 ± 0.08	0.45 ± 0.02	0.36 ± 0.01†	0.44 ± 0.01
PD	Threshold	0.14 ± 0.11	0.54 ± 0.04	0.52 ± 0.01†	0.59 ± 0.01†
	Percentile	0.32 ± 0.06	0.59 ± 0.03	0.52 ± 0.01†	0.59 ± 0.01†
WPD	Threshold	0.4 ± 0.2	0.45 ± 0.03	0.38 ± 0.01†	0.43 ± 0.01†
	Percentile	0.54 ± 0.08	0.45 ± 0.01	0.38 ± 0.01†	0.44 ± 0.01†
MKLD	Threshold	0.2 ± 0.2	0.48 ± 0.02	0.36 ± 0.01†	0.46 ± 0.01†
	Percentile	0.56 ± 0.08	0.46 ± 0.02	0.36 ± 0.01†	0.46 ± 0.01†
SF	Threshold	0.30 ± 0.19	0.48 ± 0.02	0.44 ± 0.01†	0.46 ± 0.01†
	Percentile	0.48 ± 0.06	0.52 ± 0.03	0.44 ± 0.01†	0.47 ± 0.01†
SuF	Threshold	0.4 ± 0.2	0.42 ± 0.07	0.26 ± 0.01†	0.40 ± 0.03
	Percentile	0.64 ± 0.08	0.42 ± 0.07	0.26 ± 0.01†	0.40 ± 0.02

Symbol † denotes the cases in which the deviation is lower than the second significant decimal figure

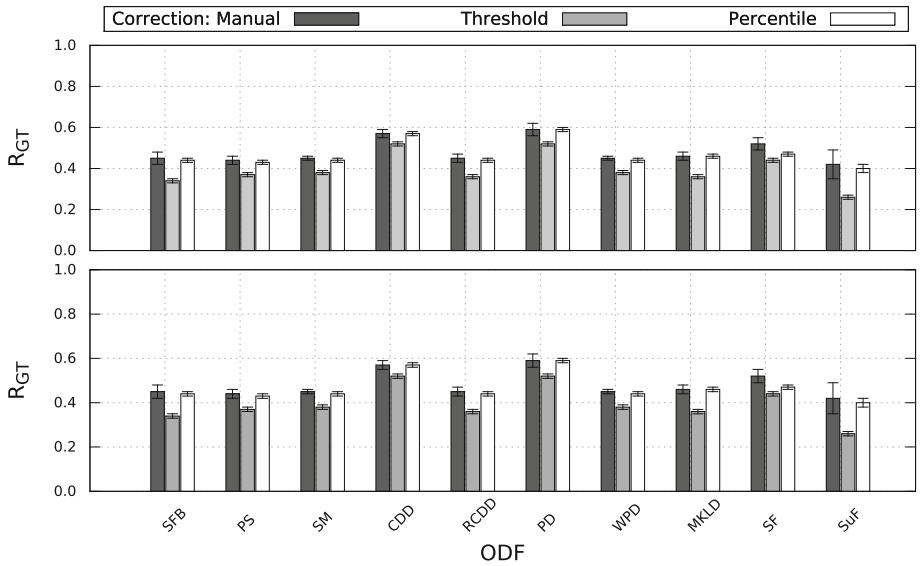

Fig. 6 Graphical representation of the user effort results obtained in terms of the R_{GT} measure for the manual correction and the threshold-based and percentile-based interactive schemes. *Top* and *bottom* figures represent the results obtained when considering either threshold-based or percentile-based OSF respectively

Table 4 R_{TC} figures for the different onset detectors considered. R_{xy} represents each R_{TC} score, where x refers to the selection function (OSF) used and y to the interactive approach

ODF	R_{TT}	R_{PT}	R_{TP}	R_{PP}
SFB	0.75 ± 0.11	0.69 ± 0.07	1.3 ± 0.2	1.00 ± 0.14
PS	0.77 ± 0.06	0.80 ± 0.03	1.10 ± 0.11	0.98 ± 0.05
SM	0.78 ± 0.06	0.80 ± 0.03	1.11 ± 0.12	0.99 ± 0.06
CDD	0.92 ± 0.11	0.83 ± 0.06	1.2 ± 0.2	1.00 ± 0.09
RCDD	0.73 ± 0.07	0.73 ± 0.04	1.3 ± 0.3	1.01 ± 0.11
PD	0.96 ± 0.14	0.79 ± 0.09	1.4 ± 0.3	1.0 ± 0.2
WPD	0.77 ± 0.05	0.80 ± 0.03	1.10 ± 0.11	0.97 ± 0.06
MKLD	0.69 ± 0.04	0.72 ± 0.05	1.2 ± 0.2	1.01 ± 0.13
SF	0.87 ± 0.06	0.78 ± 0.08	0.99 ± 0.07	0.86 ± 0.10
SuF	0.54 ± 0.12	0.56 ± 0.07	1.61 ± 0.2	1.0 ± 0.2

configurations among them to gain some insights about their differences in behavior.

Checking the figures obtained, and disregarding the initial selection function, the threshold-based interaction scheme (R_{xT}) clearly outperforms the percentile-based one (R_{xP}) as the R_{TC} results are always lower in the former one. In the same sense, threshold-based figures always achieved values under the unit whereas the other scheme was clearly not capable of doing so. Deviation figures also proved threshold-based interaction as more robust, given that in general they were lower than the ones obtained in the percentile-based scheme.

Focusing on the threshold-based schemes, it can be seen that scores (both in terms of average and deviation) were quite similar independent of the initial selection methods (OSF). This fact suggests that this straight-forward modification of the threshold value could be considered a rather robust method capable of achieving good effort figures independently of the estimation given by the initial selection process (OSF).

On the contrary, attending to the difference in the results among the percentile-based interaction schemes, the initial estimation has a clear influence for this type of interaction. As observed, using an initial selection process (OSF) based on either threshold or percentile, results in terms of the R_{TC} get to diverge in 0.3 points (case of SFB) or even 0.6 points (as in SuF). Thus, given the dependency of this interaction scheme with the initial selection process (OSF), results suggest that the best particular configuration for this percentile-based interaction approach is the case in which the initial static selection is based on percentile as well, i.e., R_{PP}.

4.1 Analysis

In order to statistically assess the reduction of the user effort a Wilcoxon signed-rank test [42] has been performed comparing each interactive method proposed against manual correction. This comparison has been performed considering the *Corrections to Ground Truth ratio* (R_{GT}) values. Table 5 shows the results of this test when considering a significance $p < 0.05$.

Figures obtained show that threshold-based interaction significantly reduced the correction workload when compared to the manual correction. It is especially remarkable the fact that this approach consistently reduced the user effort for all the combinations of onset selection and detection methods considered.

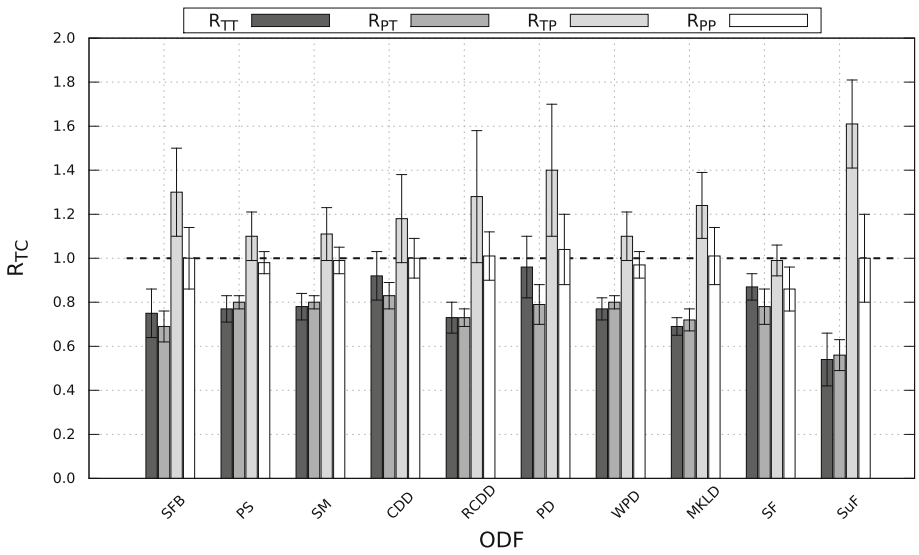

Fig. 7 Graphical representation of the user effort measure R_{TC} for all the combinations of OSF and interactive correction schemes. $R_{TC} = 1$ is highlighted

Table 5 Statistical significance results of the user effort invested in the correction of the detected onsets. Manual correction (R_{GT}^{man}) is compared against the threshold-based (R_{GT}^{thres}) and percentile-based (R_{GT}^{pctl}) interactive correction methods. Symbols $<$, $>$ and $=$ state that effort invested with the interactive methodologies is significantly lower, higher or not different to the results by the manual correction. Significance has been set to $p < 0.05$

ODF	OSF	R_{GT}^{thres} *vs* R_{GT}^{man}	R_{GT}^{pctl} *vs* R_{GT}^{man}
SFB	Threshold	$<$	$>$
	Percentile	$<$	$<$
PS	Threshold	$<$	$<$
	Percentile	$<$	$<$
SM	Threshold	$<$	$<$
	Percentile	$<$	$<$
CDD	Threshold	$<$	$>$
	Percentile	$<$	$=$
RCDD	Threshold	$<$	$=$
	Percentile	$<$	$<$
PD	Threshold	$<$	$>$
	Percentile	$<$	$=$
WPD	Threshold	$<$	$<$
	Percentile	$<$	$<$
MKLD	Threshold	$<$	$<$
	Percentile	$<$	$<$
SF	Threshold	$<$	$<$
	Percentile	$<$	$<$
SuF	Threshold	$<$	$=$
	Percentile	$<$	$<$

Results for the percentile-based interaction also show that for most of the cases there was a significant reduction in terms of workload. However, this statistical evaluation also proves that, for some particular configurations as for instance CDD with the percentile-based selection function or the SuF with the global threshold selection function, this interactive scheme may not be useful if percentiles are used for adapting the system from the user corrections, as the resulting workload does not significantly differ from the manual correction. In addition, a particular mention must be done to the SFB, CDD, and PD algorithms with the global threshold selection function as they constitute the particular cases in which the interactive algorithm implies more user effort than the manual correction.

Finally, figures obtained with this statistical analysis state the robustness of the threshold-based interaction when compared to the percentile-based scheme: while results for the former method consistently presented a reduction in workload, the latter one did not show such steady behavior.

5 Conclusions

The present work focuses on user-assisted onset detection and correction. Given that no method is able to retrieve perfect results in terms of accuracy, human correction is required for situations in which the correctness in the onset information is a must, like in a database annotation or in music teaching environments. In such cases, the user can be considered as an *active* part of the detection process rather than a verification agent. Therefore, it is necessary to propose and evaluate interactive systems capable of reducing user effort in these tasks.

Following this premise, and assuming that estimation errors occur because of an incorrect configuration of the peak selection function, two different schemes have been proposed: a first one that directly sets a new threshold value for processing the onset detection function as an interaction is performed; and a second one which combines a sliding-window analysis of the detection function with statistical information with the idea of adapting its performance to the particularities of the function.

Due to the lack of methodology for evaluating such systems, a series of measures for the quantitative assessment of the user effort in interactive onset detection schemes have been proposed. A first metric compares the required workload to complete the task when using an interactive system against the complete manual correction. The second one compares the workload required in the correction of a sequence, either manually or within an interactive scheme, to the number of annotations a user would make if no initial detection was employed.

Experimentation was carried out using a dataset comprising close to 25,000 onset events in roughly 2 h of audio in more than 300 files. A comprehensive list of onset detection and selection algorithms has also been considered. Results show that, in general terms, interactive onset detection schemes significantly reduce the workload required for the user to correct the errors in the estimation, exhibiting some particular configurations a reduction of a 40% of the workload in terms of the proposed method. Also, this human effort is also reduced when compared to the case of annotating the signal from scratch without any initial estimation, which is the typical situation when a new dataset has to be annotated.

When comparing the two proposed interaction schemes, the one directly modifying the threshold value shows a more remarkable workload reduction and superior robustness than the one pairing percentile information with the sliding window analysis.

In view of these results, it might be interesting to pursue new lines in order to further develop this work. A major drawback of these interactive systems is that they

still cannot be considered to be practical tools for a massive database annotation: for instance, consider a case in which a reduction of the 50% of the amount of interactions is achieved; if 1000 onsets had to be annotated, the user would have to deal with 500 elements, which still constitutes a significant workload.

A first aspect to reduce the amount of interactions required would be to consider the possibility of the user *moving* a certain False Positive to the position of a False Negative. We have observed that users tend to do this kind of corrections in practice for neighbour errors in order to correct both with a single interaction. In our simulation scheme this interaction has been considered as two corrections (one False Negative and one False Positive), so the figures obtained are pessimistic, based on the kind of interactions that have been simulated, and can be improved in practice.

Another point that could be considered is that, rather than starting every single correction from scratch, machine learning techniques could be used to *learn* how the scheme can be adapted from the interactions performed by the user. Given the *plasticity* of *data-driven* methods, progressive user interactions could refine a model initially trained with generic data so that it could be used to process types of sound not considered previously. In this context, a possible path to explore would be the use of *reinforcement learning*, which stands for the family of algorithms whose performance is adapted to the problem at issue with the successive use of rewards and penalties as the task is either accomplished or not.

Finally, a point of remarkable interest is to consider different costs for the False Negative and False Positive errors. While in this work it is assumed that the cost of including a missed onset is totally equivalent to the cost of removing an extra onset, in practical terms there is a difference. Thus, it seems interesting to further extend and generalise the proposed evaluation methodology to consider different weights for the different types of errors to model the real human effort invested.

Endnote

[1] Note that in this case $t_i \in t$ since there is no interaction point t_{int} for the OSF.

Acknowledgements
This research work is partially supported by the Vicerrectorado de Investigación, Desarrollo e Innovación de la Universidad de Alicante through FPU program (UAFPU2014–5883) and the Spanish Ministerio de Economía y Competitividad through the TIMuL project (No. TIN2013–48152–C2–1–R, supported by EU FEDER funds). Authors would like to thank Juan P. Bello, Sebastian Böck and Andre Holzapfel for kindly sharing their datasets.

Authors' contributions
Both authors have equally contributed in proposing the ideas, discussing the results, and writing and proofreading the manuscript. JJVM implemented the algorithms and carried out the experiments.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 21 October 2016 Accepted: 12 June 2017
Published online: 27 June 2017

References

1. P Brossier, JP Bello, MD Plumbley, in *Proceedings of the International Computer Music Conference, ICMC*. Real-time temporal segmentation of note objects in music signals (ICMC, Florida, 2004), pp. 458–461
2. JP Bello, L Daudet, SA Abdallah, C Duxbury, ME Davies, MB Sandler, A Tutorial on Onset Detection in Music Signals. IEEE Trans Speech Audio Process. **13**(5), 1035–1047 (2005)
3. M Alonso, G Richard, B David, Tempo Estimation for Audio Recordings. J New Music Res. **36**(1), 17–25 (2007)
4. E Benetos, S Dixon, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. Polyphonic music transcription using note onset and offset detection (ICASSP, Prague, 2011), pp. 37–40
5. D Dorran, R Lawlor, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Time-scale modification of music using a synchronized subband/time-domain approach (ICASSP, Montreal, 2004), pp. 225–228
6. A Robertson, MD Plumbley, in *Proceedings of the International Conference on New Interfaces for Musical Expression*. B-Keeper : A Beat-Tracker for Live Performance, (New York City, NY, 2007), pp. 234–237
7. W Wang, Y Luo, J Chambers, S Sanei, Note Onset Detection via Nonnegative Factorization of Magnitude Spectrum. EURASIP J Adv Signal Process. **2008**(1), 1–15 (2008)
8. F Eyben, S Böck, B Schuller, A Graves, in *Proceedings of the 11th International Society for Music Information Retrieval Conference*. Universal Onset Detection with Bidirectional Long Short-Term Memory Neural Networks (ISMIR, Utrecht, 2010), pp. 589–594
9. E Marchi, G Ferroni, F Eyben, L Gabrielli, S Squartini, B Schuller, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multi-resolution linear prediction based features for audio onset detection with bidirectional LSTM neural networks (ICASSP, Florence, 2014), pp. 2164–2168
10. J Serrà, TH Özaslan, JL Arcos, Note Onset Deviations as Musical Piece Signatures. PLoS ONE. **8**(7), 69268 (2013)
11. W Bas de Haas, F Wiering, Hooked on Music Information Retrieval. Empir Musicol Rev. **5**(4), 176–185 (2010)
12. E Benetos, S Dixon, D Giannoulis, H Kirchhoff, A Klapuri, in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR*. Automatic Music Transcription: Breaking the Glass Ceiling (ISMIR, Porto, 2012), pp. 379–384
13. R Rossi, A Faria, Profiling New Paradigms in Sound and Music Technologies. J New Music Res. **40**(3), 191–204 (2011)
14. E Benetos, S Dixon, D Giannoulis, H Kirchhoff, A Klapuri, Automatic music transcription: challenges and future directions. J Intell Inf Syst. **41**(3), 407–434 (2013)
15. JM Iñesta, C Pérez-Sancho, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*. Interactive multimodal music transcription (ICASSP, Vancouver, 2013), pp. 211–215
16. R Zhou, M Mattavelli, G Zoia, Music Onset Detection Based on Resonator Time Frequency Image. IEEE Trans Audio, Speech, Language Process. **16**(8), 1685–1695 (2008)
17. J Glover, V Lazzarini, J Timoney, Real-time detection of musical onsets with linear prediction and sinusoidal modeling. EURASIP J Adv Signal Process. **2011**(1), 1–13 (2011)
18. A Klapuri, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. Sound onset detection by applying psychoacoustic knowledge, vol. 6 (ICASSP, Phoenix, 1999), pp. 3089–3092
19. A Pertusa, A Klapuri, JM Iñesta, in *Proc 10th Iberoamerican Congress on Pattern Recognition, CIARP*. Recognition of Note Onsets in Digital Music Using Semitone Bands (CIARP, Havana, 2005), pp. 869–879

20. N Collins, in *Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR*. Using a Pitch Detector for Onset Detection (ISMIR, London, 2005), pp. 100–106
21. JP Bello, MB Sandler, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. Phase-based note onset detection for music signals (ICASSP, Hong Kong, 2003), pp. 49–52
22. A Holzapfel, Y Stylianou, AC Gedik, B Bozkurt, Three dimensions of pitched instrument onset detection. IEEE Trans Audio, Speech, Language Processing. **18**(6), 1517–1527 (2010)
23. JP Bello, C Duxbury, M Davies, MB Sandler, On the use of phase and energy for musical onset detection in the complex domain. IEEE Signal Processing Letters. **11**, 553–556 (2004)
24. E Benetos, Y Stylianou, Auditory Spectrum-Based Pitched Instrument Onset Detection. IEEE Trans Audio, Speech, Language Process. **18**(8), 1968–1977 (2010)
25. C Rosão, R Ribeiro, D Martins de Matos, in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR*. Influence of peak picking methods on onset detection (ISMIR, Porto, 2012), pp. 517–522
26. S Böck, J Schlüter, G Widmer, in *Proceedings of the 6th International Workshop on Machine Learning and Music*. Enhanced Peak Picking for Onset Detection with Recurrent Neural Networks (MML, Prague, 2013)
27. S Abdallah, M Plumbley, in *Proceedings of the Cambridge Music Processing Colloquium*. Unsupervised onset detection: a probabilistic approach using ICA and a hidden Markov classifier (CMPC, Cambridge, 2003)
28. J Schlüter, S Böck, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*. Improved Musical Onset Detection with Convolutional Neural Networks (ICASSP, Florence, 2014), pp. 6979–6983
29. S Böck, F Krebs, M Schedl, in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR*. Evaluating the Online Capabilities of Onset Detection Methods (ISMIR, Porto, 2012), pp. 49–54
30. AH Toselli, E Vidal, F Casacuberta, *Multimodal Interactive Pattern Recognition and Applications*, 1st. (Springer, New York, USA, 2011)
31. J Calvo-Zaragoza, J Oncina, An efficient approach for Interactive Sequential Pattern Recognition. Pattern Recognition. **64**, 295–304 (2017)
32. K West, S Cox, in *Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR*. Finding An Optimal Segmentation for Audio Genre Classification (ISMIR, London, 2005), pp. 680–685
33. D Stowell, MD Plumbey, in *Proceedings of the International Computer Music Conference, ICMC*. Adaptive whitening for improved real-time audio onset detection (ICMC, Copenhagen, 2007), pp. 312–319
34. S Dixon, in *Proceedings of the 9th International Conference on Digital Audio Effects, DAFx-06*. Onset detection revisited (DAFx, Montreal, 2006), pp. 133–137
35. C Duxbury, JP Bello, M Davies, M Sandler, in *Proceedings of the 6th International Conference on Digital Audio Effects, DAFx-03*. Complex Domain Onset Detection for Musical Signals (DAFx, London, 2003), pp. 90–93
36. P Brossier, *Automatic Annotation of Musical Audio for Interactive Application. Phd thesis*. (Centre for Digital Music Queen Mary, University of London, UK, 2007)
37. P Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals. Phd thesis,*. (Department of Electrical and Electronic Engineering, University of Bristol, UK, 1996)
38. S Böck, G Widmer, in *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR*. Local Group Delay based Vibrato and Tremolo Suppression for Onset Detection (ISMIR, Curitiba, 2013), pp. 361–366
39. S Böck, G Widmer, in *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*. Maximum Filter Vibrato Suppression for Onset Detection, (Maynooth, Ireland, 2013), pp. 55–61
40. A Lerch, I Klich, On the Evaluation of Automatic Onset Tracking Systems. Technical report, Berlin, Germany (2005)
41. G List, The Reliability of Transcription. Ethnomusicology. **18**(3), 353–377 (1974)
42. J Demsar, Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research. **7**, 1–30 (2006)