Experimental assessment of descriptive statistics and adaptive methodologies for threshold establishment in onset selection functions

Jose J. Valero-Mas, José M. Iñesta
Pattern Recognition and Artificial Intelligence Group
University of Alicante
{jjvalero,inesta}@dlsi.ua.es

ABSTRACT

Onset detection methods generally work on a two-stage basis: a first step which processes an audio stream and computes a time series depicting the estimated onset positions as local maxima in the function; and a second one which evaluates this time series to select some of the local maxima as onsets, typically with the use of peak picking and thresholding methodologies. Nevertheless, while the former stage has received considerable attention from the community, the second one typically ends up being one of a reduced catalogue of procedures. In this work we focus on this second stage and explore previously unconsidered methods based on descriptive statistics to obtain the threshold function. More precisely, we consider the use of the percentile descriptor as a design parameter and compare it to classic strategies, as for instance the median value. Additionally, a thorough comparison of methodologies considering the temporal evolution of the time series (adaptive techniques) against the use of static threshold values (non-adaptive) is carried out. The results obtained report several interesting conclusions, being the most remarkable two the fact that the percentile descriptor can be considered a competitive possible alternative for this task and that adaptive approaches do not always imply an improvement over static methodologies.

1. INTRODUCTION

Onset detection stands for the process of automatically retrieving the starting points of note events present in audio music signals [1]. This task has been commonly addressed in the Music Information Retrieval (MIR) discipline due to its remarkable relevance in related research fields such as *beat detection* [2], *tempo estimation* [3], or *automatic music transcription* [4].

However, in spite of its conceptual simplicity, this task remains unsolved. While the performance for cases of simple instrumentation and low polyphony degree may be considered sufficient for a number of MIR tasks, in more complex cases onset estimation still constitutes a challenge.

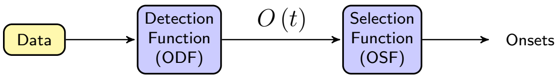


Figure 1: Graphical description of a two-stage onset detection process.

Except for some approaches based on end-to-end systems (e.g, [5]), classic onset estimation schemes base its performance on a process comprising two consecutive steps [6], as shown in Fig. 1. The first stage, usually known as *Onset Detection Function* (ODF), processes the target signal computing a time series whose peaks represent the positions of the estimated onsets, by measuring the change in one or more audio features. Features typically considered are signal energy [7], phase [8], or combinations of them [9].

The second stage, known as *Onset Selection Function* (OSF), evaluates the result from the ODF process and selects the most promising peaks as onsets. Ideally, if the ODF method would perfectly track the onset events in the signal, this stage would not be required. However, as no ODF process is neat, the OSF is required to discriminate between actual onsets and artifacts, thus constituting a key point in the system's performance [10, 11].

Due to its relevance, a number of OSF techniques have been proposed. In general, OSF methods seek for peaks above a certain threshold in the ODF. In this regard, while the maximum condition is typically unaltered, the different techniques differ in the way this threshold is obtained.

The most basic approach consists in manually establishing a static threshold value [12]. This technique does not consider any particularities of the signal for selecting the value, i.e., no prior knowledge of the signal is taken into account. Its value is heuristically set according to previous experimentation. Other more sophisticated approaches consider calculating the mean or the median value of the ODF and setting it as the static threshold value [13, 14].

Extending the premises of the aforementioned methods, adaptive techniques which consider the temporal evolution of the ODF are also used. Instead of obtaining a global static threshold value, these methods include a sliding window for performing the analysis at each point [15]. In this context, the median value of the elements in the sliding window is commonly considered a remarkably robust approach for onset detection [16].

Nevertheless, except for some particular works assessing the influence of the OSF process in the overall performance of the system (e.g., [10]), there is still a need to further study such processes. Thus, in this work we perform an experimental study assessing different ideas which, to the best of our knowledge, no previous work has considered. Particularly, we study the possibility of considering other percentile values (i.e., other statistical tendency measures) different to the median, somehow extending the work by Kauppinen [13] but particularised to onset detection. We also examine the influence of the window size in adaptive methodologies. Eventually, we compare static and adaptive threshold procedures to check their respective capabilities.

The rest of the paper is structured as follows: Section 2 introduces the ODF methods considered for this work; Section 3 explains the different OSF approaches assessed and compared; Section 4 describes the evaluation methodology considered; Section 5 introduces and analyses the results obtained; eventually, Section 6 remarks the conclusions and proposes future work.

2. ONSET DETECTION FUNCTIONS

Two different ODF algorithms were used for our experimentation, which were selected according to their good results reported in the literature: the Semitone Filter Bank (SFB) [7] and the SuperFlux (SuF) [17,18] algorithms. For the rest of the paper we refer to the time series resulting from the ODF process as $O(t)$.

As an energy-based ODF method, SFB analyses the evolution of the magnitude spectrogram with the particularity of considering that harmonic sounds are being processed. A semitone filter bank is applied to each frame window of the magnitude spectrogram, being the different filters centred at each of the semitones marked by the well temperament, and the energy of each band (root mean square value) is retrieved. After that, the first derivative across time is obtained for each single band. As only energy rises may point out onset information, negative outputs are zeroed. Finally, all bands are summed and normalised to obtain the $O(t)$ function.

The SuF method bases its performance on the idea of the Spectral Flux [19] signal descriptor and extends it. Spectral Flux obtains positive deviations of the bin-wise difference between two consecutive frames of the magnitude spectrogram for retrieving the $O(t)$ function. SuF substitutes the difference between consecutive analysis windows by a process of tracking spectral trajectories in the spectrum together with a maximum filtering process that suppresses vibrato articulations as they tend to increase false detections.

Given that the different ODF processes may not span in the same range, we apply the following normalisation process to $O(t)$ so that the resulting function $\hat{O}(t)$ satisfies $\hat{O}(t) \in [0,1]$:

$$\hat{O}(t) = \frac{O(t) - \min\{O(t)\}}{\max\{O(t)\} - \min\{O(t)\}} \tag{1}$$

where $\min\{\cdot\}$ and $\max\{\cdot\}$ retrieve the minimum and maximum values of $O(t)$, respectively.

Finally, it must be mentioned that the analysis parameters of both algorithms have been configured to a window size of $92.9\ ms$ with a 50 % of overlapping factor.

3. ONSET SELECTION FUNCTIONS

In this section we introduce the different OSF methods considered. Given that the contemplated ODF processes depict onset candidates as local maxima in $\hat{O}(t)$, the studied OSF methods search for peaks above a certain threshold value $\theta(t)$ for removing spurious detections.

As introduced, the idea in this work is to assess the following OSF criteria: a) considering percentile values different to the median one; b) assessing the influence of the window size in adaptive methodologies; and c) comparing the performance of static against adaptive methodologies. Mathematically, these criteria can be formalised as in Eq. 2:

$$\theta(t_i) = \mu\left\{\hat{O}(t_{w_i})\right\} + \mathcal{P}^{(n)}\left\{\hat{O}(t_{w_i})\right\} \tag{2}$$

where $t_{w_i} \in \left[t_i - \frac{W}{2}, t_i + \frac{W}{2}\right]$, W denotes the size of the sliding window, and $\mu\{\cdot\}$ and $\mathcal{P}^{(n)}\{\cdot\}$ denote the average and n_{th} percentile value of the sample distribution at issue, respectively.

For assessing the influence of the percentile, 20 values equally spaced in the range $n \in [0, 100]$ were considered. The particular case of $\mathcal{P}^{(50)}$ is equivalent to the median value of the distribution.

In terms of window sizes, 20 values equally space in the range $W \in [0.2, 5]\ s$ were also considered. Additionally, the value proposed by West and Cox [20] of $W = 1.5\ s$ was included as a reference.

In order to obviate the adaptive capabilities of the OSF, W was set to the length of the $O(t)$ to assess the static methodologies.

In addition, the case of manually imposing the threshold value $\theta(t_i) = \mathcal{T}$ was considered. For that we establish 20 values equally spaced in the range $\mathcal{T} \in [0,1]$ as $\hat{O}(t)$ is normalised to that range.

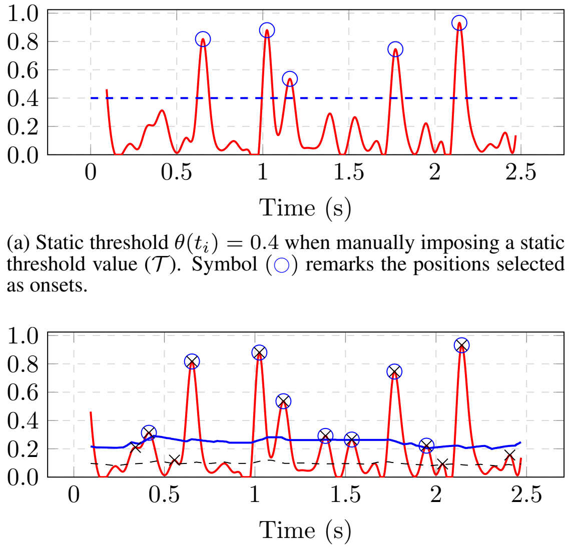
For the rest of the paper, the following notation will be used: μ and $\mathcal{P}$ denote the use of the mean or the percentile, respectively, whereas $\mu + \mathcal{P}$ stands for the sum of both descriptors; adaptive approaches are distinguished from the static ones by incorporating the t as a subindex, showing their temporal dependency (i.e., μ_t, $\mathcal{P}_t$, and $\mu_t + \mathcal{P}_t$ in opposition to μ, $\mathcal{P}$, and $\mu + \mathcal{P}$); $\mathcal{T}$ evinces the manually imposition of the threshold value; lastly, $\mathcal{B}$ is used to denote the case in which no threshold is applied and all local maxima are retrieved as onsets: $\mathcal{B} \equiv \theta(t_i) = 0$.

Finally, Fig. 2 shows some graphical examples of OSF processes applied to a given $\hat{O}(t)$ function.

4. METHODOLOGY

4.1 Corpus

The dataset used for the evaluation is the one proposed by Böck et al. [14]. It comprises 321 monaural recordings sampled at 44.1 kHz covering a wide variety of polyphony degrees and instrumentation.



(a) Static threshold $\theta(t_i) = 0.4$ when manually imposing a static threshold value ($\mathcal{T}$). Symbol (○) remarks the positions selected as onsets.

(b) Adaptive threshold $\theta(t_i)$ using the percentile scheme ($\mathcal{P}_t$): solid and dashed styles represent the 75th and 50th percentiles, respectively, while symbols (○) and (×) depict the positions selected as onsets for each criterion. The sliding window has been set to $W = 1.5\ s$.

Figure 2: Examples of OSF applied to a given $O(t)$.

The total duration of the set is 1 hour and 42 minutes and contains 27,774 onsets. The average duration per file is 19 seconds (the shortest lasts 1 second and the longest one extends up to 3 minutes) with an average figure of 87 onsets per file (minimum of 3 onsets and maximum of 1,132 onsets). For these experiments no partitioning in terms of instrumentation, duration, or polyphony degree was considered as we aim at assessing broader tendencies.

4.2 Performance measurement

Given the impossibility of defining an exact point in time to be the onset of a sound, correctness evaluation requires of a certain tolerance [21]. In this regard, an estimated onset is typically considered to be correct (*True Positive*, TP) if its corresponding ground-truth annotation is within a $\pm 50\ ms$ time lapse of it [1]. *False Positive* errors (FP) occur when the system overestimates onsets and *False Negative* errors (FN) when the system misses elements present in the reference annotations.

Based on the above, we may define the F-measure (F_1) figure of merit as follows:

$$\mathrm{F}_1 = \frac{2 \cdot \mathrm{TP}}{2 \cdot \mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \qquad (3)$$

where $\mathrm{F}_1 \in [0, 1]$, denoting the extreme values 0 and 1 the worst and best performance of the system, respectively.

Other standard metrics such as Precision (P) and Recall (R) have been obviated for simplicity.

5. RESULTS

In this section we introduce and analyse the results obtained for the different experiments proposed. All figures presented in this section show the weighted average of the results obtained for each of the audio files considering the number of onsets each one contains.

Table 1 shows the results obtained for the proposed comparative of static and adaptive threshold methodologies for both ODF processes considered. In this particular case, window size for adaptive methodologies has been fixed to $W = 1.5\ s$ as in West and Cox [20].

An initial remark to point out according to the results obtained is that, in general, the figures obtained with the sliding window methodology do not remarkably differ to the ones obtained in the static one, independently of the ODF process. This can be clearly seen when $\mathcal{P}$ and $\mu + \mathcal{P}$ are respectively compared to $\mathcal{P}_t$ and $\mu_t + \mathcal{P}_t$: all percentile values considered retrieve considerably similar results except for the case when high percentile values are considered, in which the $\mathcal{P}_t$ method shows its superior capabilities. Additionally, while $\mathcal{P}$ approaches show their best performance for percentile values in the range $[60, 75]$, the $\mu + \mathcal{P}$ methods obtain their optimal results in the lower ranges, more precisely around $[0, 30]$.

Regarding the use of the μ methodologies, results obtained for the static and adaptive methodologies did not differ for the SFB function. On the contrary, when considering the SuF function, the adaptive methodology performed slightly worse than the static one. Finally, as these methods do not depend on any external configuration, their performance does not vary with the percentile parameter.

In terms of the $\mathcal{T}$ strategy, its performance matched the static $\mathcal{P}$ and $\mu + \mathcal{P}$ methods in the sense that the performance degrades as the introduced threshold was increased. Nevertheless, this method showed its performance when the threshold value considered lies in the range $[0.10, 0.30]$.

As a general comparison of the methods considered, the μ methods showed a very steady performance paired with high performance results. Also, while $\mu + \mathcal{P}$ methods generally outperform $\mathcal{P}$ strategies in terms of peak performance, the particular case of $\mathcal{P}_t$ approaches showed less dependency on the percentile configuration value.

Given that all previous experimentation has been done considering a fixed window value of $W = 1.5\ s$, we need to study the influence of that parameter in the overall performance of the system. In this regard, Tables 2 and 3 show the results for the $\mathcal{P}_t$ and $\mu_t + \mathcal{P}_t$ methods considering different window sizes for the SFB and SuF processes, respectively. Additionally, Figures 3 and 4 graphically show these results for the $\mathcal{P}_t$ and $\mu_t + \mathcal{P}_t$ methods, respectively.

As the figures obtained for both ODF processes qualitatively show the same trends, we shall analyse them jointly. Checking Tables 2 and 3, results for the $\mathcal{P}_t$ method show that larger windows tend to increase the overall performance of the detection, at least when not considering an extreme percentile value. For instance, let us focus on the $\mathcal{P}_t^{(74)}$ method for the SuF process (Table 3): when considering a window of $W = 0.20\ s$, the performance is set on $\mathrm{F}_1 = 0.67$, which progressively improves as the window size is increased, getting to a score of $\mathrm{F}_1 = 0.73$ when $W = 5.00\ s$.

As commented, this premise is not accomplished when percentile values are set to its possible extremes. For low

Th/Pc	**SFB**							**SuF**						
	$\mathcal{T}$	μ	$\mathcal{P}$	$\mu + \mathcal{P}$	μ_t	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{T}$	μ	$\mathcal{P}$	$\mu + \mathcal{P}$	μ_t	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$
0.00	0.65	**0.73**	0.65	**0.73**	0.72	0.65	0.72	0.64	**0.77**	0.64	**0.77**	0.74	0.64	0.76
0.11	0.71	**0.73**	0.65	0.72	0.72	0.65	0.71	0.76	**0.77**	0.64	**0.77**	0.74	0.64	0.76
0.21	**0.75**	0.73	0.66	0.71	0.72	0.66	0.70	0.65	**0.77**	0.65	**0.77**	0.74	0.64	0.76
0.32	**0.73**	**0.73**	0.66	0.70	0.72	0.66	0.69	0.51	**0.77**	0.66	**0.77**	0.74	0.65	0.76
0.42	0.66	**0.73**	0.68	0.68	0.72	0.67	0.67	0.39	**0.77**	0.68	0.76	0.74	0.67	0.75
0.53	0.56	**0.73**	0.69	0.66	0.72	0.69	0.63	0.29	**0.77**	0.70	0.75	0.74	0.69	0.74
0.63	0.44	**0.73**	0.70	0.62	0.72	0.70	0.58	0.21	**0.77**	0.73	0.73	0.74	0.71	0.71
0.74	0.31	**0.73**	0.70	0.53	0.72	0.70	0.50	0.15	**0.77**	0.74	0.68	0.74	0.72	0.65
0.84	0.19	**0.73**	0.65	0.39	0.72	0.63	0.36	0.10	**0.77**	0.68	0.56	0.74	0.65	0.54
0.95	0.10	**0.73**	0.40	0.13	0.72	0.42	0.12	0.06	**0.77**	0.42	0.29	0.74	0.44	0.27
1.00	0.05	**0.73**	0.05	0.00	0.72	0.27	0.00	0.05	**0.77**	0.05	0.00	0.74	0.29	0.00

Table 1: Results in terms of the F_1 score for the dataset considered for the ODF methods introduced. A 1.5-second window has been used for the methods based on sliding window (μ_t, $\mathcal{P}_t$, $\mu_t + \mathcal{P}_t$) based on the work by West and Cox [20]. Bold elements represent the best figures for each ODF method considered. Due to space requirements we have selected the threshold and percentile parameters (Th/Pc) showing the general tendency. In this same regard, both threshold and percentile parameters are expressed in the range $[0, 1]$.

Pc	$W = 0.20$		$W = 0.71$		$W = 1.50$		$W = 2.22$		$W = 2.98$		$W = 3.99$		$W = 5.00$	
	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$
0	0.65	0.66	0.65	0.71	0.65	0.72	0.65	0.72	0.65	0.72	0.65	0.72	0.65	**0.73**
11	0.65	0.65	0.65	0.70	0.65	0.71	0.65	0.71	0.65	**0.72**	0.65	**0.72**	0.65	**0.72**
21	0.65	0.63	0.65	0.69	0.66	0.70	0.66	0.70	0.66	**0.71**	0.66	**0.71**	0.66	**0.71**
32	0.65	0.60	0.66	0.68	0.66	0.69	0.66	0.69	0.66	0.69	0.66	**0.70**	0.66	**0.70**
42	0.65	0.56	0.67	0.65	0.67	0.67	0.67	0.67	0.67	**0.68**	0.67	**0.68**	**0.68**	**0.68**
53	0.65	0.49	0.68	0.62	**0.69**	0.64	**0.69**	0.64	**0.69**	0.65	**0.69**	0.65	**0.69**	0.65
63	0.65	0.42	0.69	0.56	**0.70**	0.59	**0.70**	0.59	**0.70**	0.60	**0.70**	0.60	**0.70**	0.60
74	0.67	0.32	0.69	0.48	0.69	0.51	**0.70**	0.51	**0.70**	0.51	**0.70**	0.52	**0.70**	0.52
84	**0.67**	0.12	0.66	0.35	0.64	0.36	0.64	0.36	0.64	0.37	0.64	0.37	0.64	0.37
95	**0.67**	0.00	0.47	0.15	0.45	0.12	0.42	0.12	0.42	0.12	0.41	0.12	0.40	0.12
100	**0.67**	0.00	0.47	0.00	0.23	0.00	0.21	0.00	0.17	0.00	0.13	0.00	0.11	0.00

Table 2: Results in terms of the F_1 score for the SFB process. Bold elements represent the best figures obtained for each percentile value considered. Due to space requirements, the most representative percentile parameters (Pc) and window sizes (W) have been selected to show the general tendency.

Pc	$W = 0.20$		$W = 0.71$		$W = 1.50$		$W = 2.22$		$W = 2.98$		$W = 3.99$		$W = 5.00$	
	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$	$\mathcal{P}_t$	$\mu_t + \mathcal{P}_t$
0	0.64	0.72	0.64	0.74	0.64	0.76	0.64	0.76	0.64	**0.77**	0.64	**0.77**	0.64	**0.77**
11	0.64	0.72	0.64	0.75	0.64	0.76	0.64	**0.77**	0.64	**0.77**	0.64	**0.77**	0.64	**0.77**
21	0.64	0.72	0.64	0.75	0.65	0.76	0.65	0.76	0.65	**0.77**	0.65	**0.77**	0.65	**0.77**
32	0.64	0.71	0.65	0.75	0.65	**0.76**	0.65	**0.76**	0.65	**0.76**	0.66	**0.76**	0.66	**0.76**
42	0.64	0.70	0.66	0.74	0.67	0.75	0.67	0.75	0.67	**0.76**	0.67	**0.76**	0.68	**0.76**
53	0.64	0.65	0.68	0.73	0.69	0.74	0.69	0.74	0.70	0.74	0.70	0.74	0.70	**0.75**
63	0.64	0.61	0.70	0.69	0.71	0.71	0.71	0.71	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**	**0.72**
74	0.67	0.55	0.70	0.63	0.72	0.66	0.72	0.66	**0.73**	0.66	**0.73**	0.67	**0.73**	0.67
84	0.67	0.38	**0.68**	0.52	0.67	0.54	0.67	0.54	0.67	0.55	0.67	0.55	0.67	0.55
95	**0.67**	0.00	0.48	0.32	0.47	0.27	0.44	0.27	0.44	0.27	0.43	0.28	0.42	0.28
100	**0.67**	0.00	0.48	0.00	0.24	0.00	0.21	0.00	0.18	0.00	0.15	0.00	0.12	0.00

Table 3: Results in terms of the F_1 score for the SuF process. Bold elements represent the best figures obtained for each percentile value considered. Due to space requirements, the most representative percentile parameters (Pc) and window sizes (W) have been selected to show the general tendency.

percentile values, window size seems to be irrelevant to the system. For instance, when considering the SFB method, $\mathcal{P}_t^{(11)}$, the performance measure was always $\mathrm{F}_1 = 0.65$ independently of the W value considered.

On the other extreme, very high percentile values suffer a performance decrease as larger window sizes are used. As an example, for the SuF configuration, $\mathcal{P}_t^{(100)}$ with $W = 5.00\ s$ achieved an $\mathrm{F}_1 = 0.12$, while when $W = 0.20\ s$ this figure raised up to $\mathrm{F}_1 = 0.67$.

Results for the $\mu_t + \mathcal{P}_t$ method showed similar tendencies since, when not considering extreme percentile values, the overall performance increased as larger windows were used. Nevertheless, in opposition to the $\mathcal{P}_t$ case, this particular configuration showed an improvement tendency as W is progressively increased for low percentile values.

In general it was observed that, for all W window sizes, the best percentile configurations seemed to be in the range $[60, 70]$ for the $\mathcal{P}_t$ approach and in the range $[0, 20]$ for the $\mu_t + \mathcal{P}_t$ case. This fact somehow confirms the hypothesis suggesting that the median value may not always be the best percentile to consider.

Checking now the results considering Figs. 3 and 4, some additional remarks, which are more difficult to check in the aforementioned tables, may be pointed out.

The first one of them is that the selection of the proper parameters of windows size and percentile factor is crucial. For both $\mathcal{P}_t$ and $\mu_t + \mathcal{P}_t$ methods, there is a *turning point* in which the performance degrades to values lower than the considered baseline $\mathcal{B}$ (no OSF applied). For the $\mathcal{P}_t$ method there is a clear point for this change in tendency around the 85th percentile for any window size considered. However, for the $\mu_t + \mathcal{P}_t$ approach there was not a unique point but a range, which remarkably varied depending on the type of ODF and window size considered.

Another point to highlight is that the static methodologies ($\mathcal{P}$ and $\mu + \mathcal{P}$) consistently define the upper bound in performance before the so-called turning point. This fact somehow confirms the initial idea of using large windows (on the limit, one single window considering the whole $O(t)$ function) in opposition to small windows.

The results obtained when considering window sizes in the range $[0, 1]$ seconds, which include the performance of the considered reference window of $W = 1.5\ s$, achieved results similar to the obtained upper bound. The other considered window sizes showed a remarkable variability in the performance, ranging from achieving figures similar to the upper bound to figures close to the baseline $\mathcal{B}$.

As a general summary for all the experimentation performed, we may conclude that adaptive OSF methodologies may be avoided as static approaches obtain similar results with less computational cost. Particularly, in our experiments, methods considering percentile ($\mathcal{P}$) or mean and percentile ($\mu + \mathcal{P}$) reported the upper bounds in performance.

Nevertheless, the particular percentile value used remarkably affects the performance. For $\mathcal{P}$, the best results seem to be obtained when the percentile parameter was set in the range [60,70]. For $\mu + \mathcal{P}$ the best figures were obtained when this parameter was set to a very low value.

Finally, the commonly considered median value for the OSF did not report the best results in our experiments. These results point out that the median statistical descriptor may not always be the most appropriate to be used, being necessary to be tuned for each particular dataset.

5.1 Statistical significance analysis

In order to perform a rigorous analysis of the results obtained and derive strong conclusions out of them, we now perform a set of statistical tests. Specifically, these analyses were performed with the non-parametric Wilcoxon rank-sum and Friedman tests [22], which avoid any assumption about the distribution followed by the figures obtained. The former method is helpful for comparisons of different distributions in a pairwise fashion while the latter one generalises this pairwise comparison to a generic number of distributions.

In this work the Wilcoxon rank-sum test was applied to assess whether there are significant differences among all the methods proposed using the results in Table 1. The single scores obtained for each Th/Pc constitute a sample of the distribution for the OSF method to be tested. The results obtained when considering a significance level of $p < 0.05$ can be checked in Table 4.

Attending to the results of the Wilcoxon test, an initial point to comment is that method $\mathcal{T}$ the less robust of all considered due to its significantly lower performance in all cases. Oppositely, method μ can be considered the best strategy as it outperforms all other strategies. The adaptive equivalent to this technique, μ_t, achieves similar performances to the static one except for the case of SuF, in which it does not outperform the rest of the methods as consistently as in the static case. Methods $\mathcal{P}$ and $\mathcal{P}_t$ show an intermediate performance among the previously commented extremes. Their results are not competitive with any of the μ or μ_t strategies. These strategies are generally tied with methods $\mu + \mathcal{P}$ and $\mu_t + \mathcal{P}_t$ as they typically show significantly similar performances with punctual cases in which one of those methods improves the rest.

We now consider the statistical analysis of influence of the window size (W) and the percentile values (Pc) on the overall performance. Given that now we have two variables to be evaluated, we consider the use of the Friedman test. The idea with this experiment is to assess whether variations in these variables report statistically significant difference in the F_1 score, so we assume that the null hypothesis to be rejected is that no difference should be appreciated. Note that this type of analysis does not report whether a particular strategy performs better than another one but simply if the differences when using different parameters are relevant.

For performing this experiment we consider the data in Tables 2 and 3 as they contain all the information required for the analysis. In this regard Tables 5 and 6 show the p significance values obtained when measuring the influence of W and Pc, respectively.

Attending to the results obtained, we can check the remarkable influence of the W parameter in the overall results as all cases report very low p scores which reject the null hypothesis. The only exception is the $\mathcal{P}_t$ in the SFB case in which p is not that remarkably low, but still would reject the null hypothesis considering a typical significance threshold of $p < 0.05$.

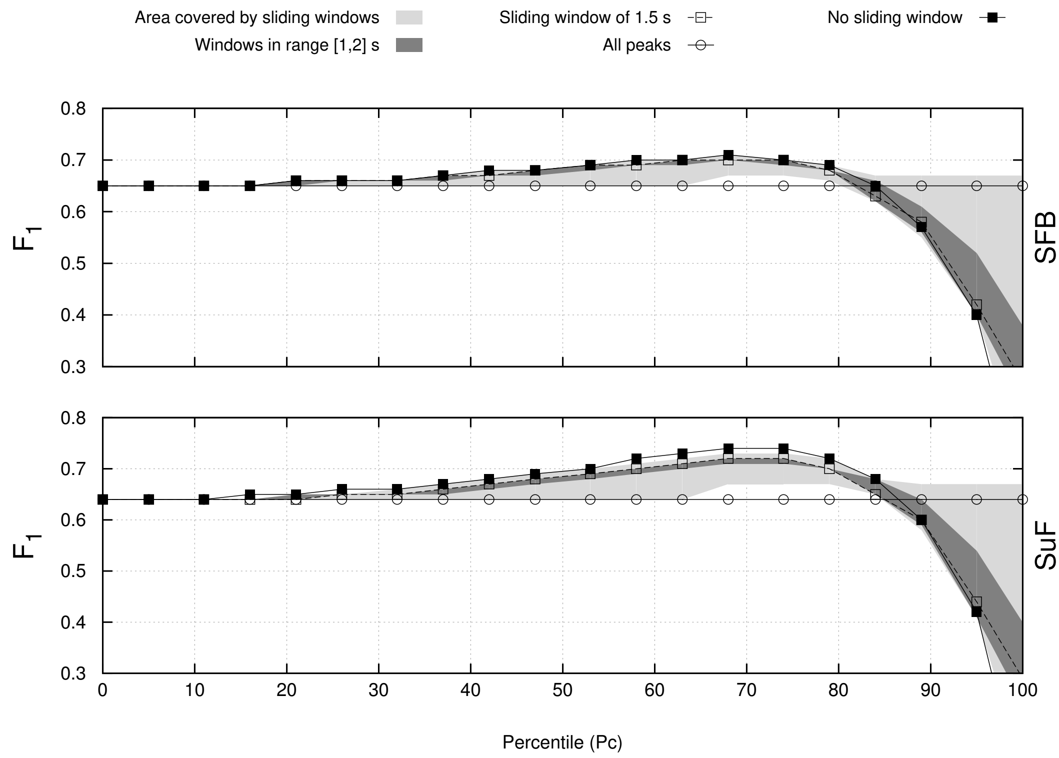


Figure 3: Evolution of the F_1 score when considering the $\mathcal{P}_t$ strategy for the dataset proposed.

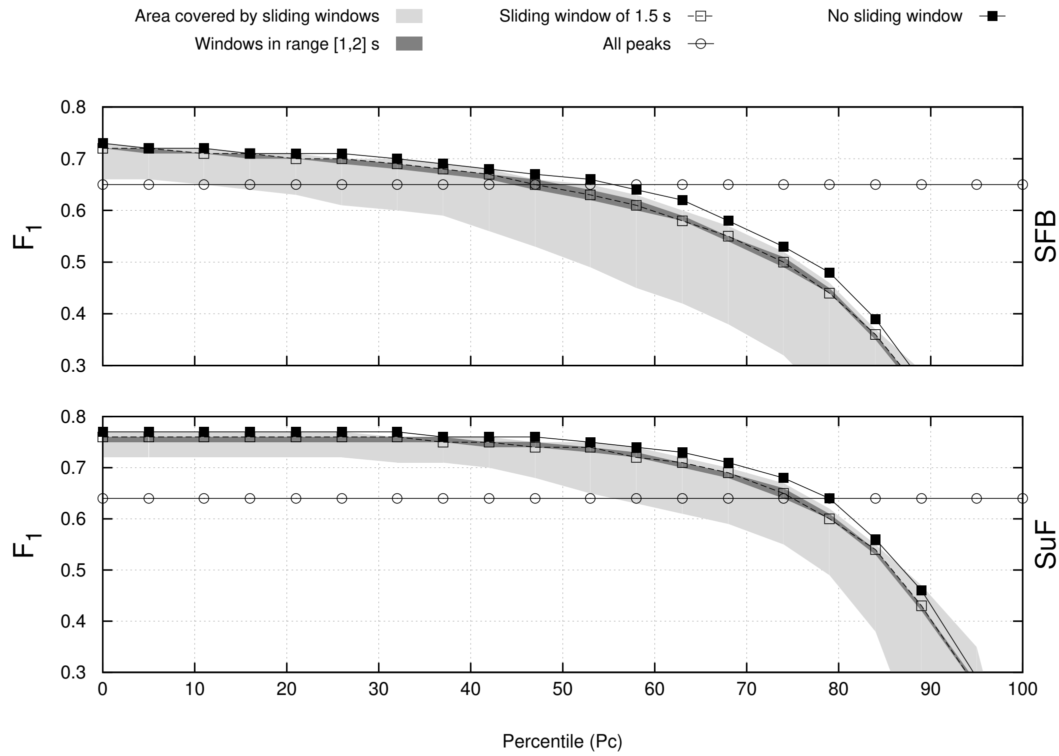


Figure 4: Evolution of the F_1 score when considering the $\mu_t + \mathcal{P}_t$ strategy for the dataset proposed.

In terms of the percentile value (Pc), the p significance values obtained clearly show the importance of this parameter in the design of the experiment. This points out the clear need to consider this as another parameter in the design of onset detection systems.

Finally, note that this statistical analysis confirms the initial conclusions depicted previously: static approaches are significantly competitive when compared to adaptive methods; window sizes for adaptive method remarkably influence the performance of the system; when considering

Method	SFB							SuF						
	$\mathcal{T}$	μ	$\mathcal{P}$	$\mu+\mathcal{P}$	μ_t	$\mathcal{P}_t$	$\mu_t+\mathcal{P}_t$	$\mathcal{T}$	μ	$\mathcal{P}$	$\mu+\mathcal{P}$	μ_t	$\mathcal{P}_t$	$\mu_t+\mathcal{P}_t$
$\mathcal{T}$	–	$<$	$<$	$<$	$<$	$<$	$<$	–	$<$	$<$	$<$	$<$	$<$	$<$
μ	$>$	–	$>$	$>$	$>$	$>$	$>$	$>$	–	$>$	$>$	$>$	$>$	$>$
$\mathcal{P}$	$>$	$<$	–	$=$	$<$	$=$	$=$	$>$	$<$	–	$=$	$<$	$>$	$=$
$\mu+\mathcal{P}$	$>$	$<$	$=$	–	$<$	$=$	$>$	$>$	$<$	$=$	–	$=$	$=$	$>$
μ_t	$>$	$<$	$>$	$>$	–	$>$	$>$	$>$	$<$	$>$	$=$	–	$>$	$=$
$\mathcal{P}_t$	$>$	$<$	$=$	$=$	$<$	–	$>$	$>$	$<$	$<$	$=$	$<$	–	$=$
$\mu_t+\mathcal{P}_t$	$>$	$<$	$=$	$<$	$<$	$<$	–	$>$	$<$	$=$	$<$	$=$	$=$	–

Table 4: Results obtained for the Wilcoxon rank-sum test to the comparison results depicted in Table 1. Symbols ($>$), ($<$), and ($=$) state that the onset detection capability of the method in the row is significantly higher, lower, or not different to the method in the column. Symbol (–) depicts that the comparison is obviated. A significance level of $p < 0.05$ has been considered.

Method	SFB	SuF
$\mathcal{P}_t$	0.03209	$4.917 \cdot 10^{-9}$
$\mu_t + \mathcal{P}_t$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$

Table 5: Statistical significance results of the Friedman test when measuring the influence of the window size (W) in the overall performance.

Method	SFB	SuF
$\mathcal{P}_t$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$
$\mu_t + \mathcal{P}_t$	$< 2.2 \cdot 10^{-16}$	$< 2.2 \cdot 10^{-16}$

Table 6: Statistical significance results of the Friedman test when measuring the influence of the percentile (Pc) in the overall performance.

OSF methods relying on statistical descriptions, the percentile should be considered as another design parameter due to its relevance in the results.

6. CONCLUSIONS

Onset detection, the retrieval of the starting points of note events in audio music signals, has been largely studied because of its remarkable usefulness in the Music Information Retrieval field. These schemes usually work on a two-stage basis: the first one derives a time series out of the initial signal whose peaks represent the candidate onset positions; the second assesses this function by setting a threshold to minimise the errors and maximise the overall performance. While the first of such processes has received significant attention from the community, the second stage has not been that deeply studied.

In this work we have studied the actual influence of a proper design of this second stage in the overall performance. More precisely we addressed three main points: measuring the benefits of static and adaptive threshold techniques, the use of different statistical descriptions for setting this threshold, and the influence of the considered window size in adaptive methodologies.

Our experiments show that static methodologies are clearly competitive when compared to adaptive methodologies. Additionally, our results report the remarkable influence that the different statistical descriptions and window sizes considered have on the overall performance, being thus important parameters to be considered in onset detection systems.

Future work considers extending this study to other descriptive statistic measures such as central tendency descriptors (e.g., harmonic and weighted means) or shape-based features (e.g., kurtosis and skewness), and considering the possibility of including specific weights for each of them. Also, the exploration of different machine learning methods to avoid use of hand-crafted rules by automatically inferring the optimal threshold curve are also considered as a line to pursue. Finally, end-to-end systems based on deep learning is also a line to further pursue.

Acknowledgments

This research work is partially supported by 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 project TIMuL (No. TIN2013–48152–C2–1–R, supported by EU FEDER funds). Authors would also like to thank Sebastian Böck for kindly sharing the onset dataset.

7. REFERENCES

[1] J. P. Bello, L. Daudet, S. A. Abdallah, C. Duxbury, M. E. Davies, and M. B. Sandler, "A Tutorial on Onset Detection in Music Signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[2] D. P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[3] M. Alonso, G. Richard, and B. David, "Tempo Estimation for Audio Recordings," *Journal of New Music Research*, vol. 36, no. 1, pp. 17–25, 2007.

[4] E. Benetos and S. Dixon, "Polyphonic music transcription using note onset and offset detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Prague, Czech Republic, 2011, pp. 37–40.

[5] J. Schlüter and S. Böck, "Improved musical onset detection with convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2014, pp. 6979–6983.

[6] R. Zhou, M. Mattavelli, and G. Zoia, "Music Onset Detection Based on Resonator Time Frequency Image," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1685–1695, 2008.

[7] A. Pertusa, A. Klapuri, and J. M. Iñesta, "Recognition of Note Onsets in Digital Music Using Semitone Bands," in *Proceedings of the 10th Iberoamerican Congress on Pattern Recognition, CIARP*, Havana, Cuba, 2005, pp. 869–879.

[8] 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, ICASSP*, Hong Kong, Hong Kong, 2003, pp. 49–52.

[9] E. Benetos and Y. Stylianou, "Auditory Spectrum-Based Pitched Instrument Onset Detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 1968–1977, 2010.

[10] C. Rosão, R. Ribeiro, and D. Martins de Matos, "Influence of peak picking methods on onset detection," in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR*, Porto, Portugal, 2012, pp. 517–522.

[11] S. Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects, DAFx-06*, Montreal, Canada, 2006, pp. 133–137.

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

[13] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals," in *Proceedings of the 14th International Conference on Digital Signal Processing, DSP*, vol. 2, 2002, pp. 967–970.

[14] S. Böck, F. Krebs, and M. Schedl, "Evaluating the Online Capabilities of Onset Detection Methods," in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR*, Porto, Portugal, 2012, pp. 49–54.

[15] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "Complex Domain Onset Detection for Musical Signals," in *Proceedings of the 6th International Conference on Digital Audio Effects, DAFx-03*, London, UK, 2003, pp. 90–93.

[16] J. P. Bello, C. Duxbury, M. Davies, and M. B. Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11, pp. 553–556, 2004.

[17] S. Böck and G. Widmer, "Maximum Filter Vibrato Suppression for Onset Detection," in *Proceedings of the 16th International Conference on Digital Audio Effects (DAFx-13)*, Maynooth, Ireland, September 2013, pp. 55–61.

[18] ——, "Local Group Delay based Vibrato and Tremolo Suppression for Onset Detection," in *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, November 2013, pp. 589–594.

[19] P. Masri, "Computer Modeling of Sound for Transformation and Synthesis of Musical Signals," PhD Thesis, University of Bristol, UK, 1996.

[20] K. West and S. Cox, "Finding An Optimal Segmentation for Audio Genre Classification," in *Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR*, London, UK, 2005, pp. 680–685.

[21] P. Leveau and L. Daudet, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *Proceeding of the 5th International Symposium on Music Information Retrieval, ISMIR*, Barcelona, Spain, 2004, pp. 72–75.

[22] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.