Onset Detection in Polyphonic Signals by means of Transient Peak Classification

A. Röbel
IRCAM, 1, place Igor Stravinsky
75004, Paris, France
roebel@ircam.fr

ABSTRACT

This abstract describes an onset detection algorithm that is based on a classification of spectral peaks into transient and non-transient peaks and a statistical model of the classification results to prevent detection of random transient peaks due to noise. A special feature of the proposed algorithm is that it is suitable for real time analysis with a maximum delay of only an 8-th part of the analysis window.

Keywords: real time, onset detection, polyphonic signals, peak classification.

1 INTRODUCTION

In the following article we are going to describe a transient detection algorithm that has been developed for a special application, the detection of transients to prevent transformation artifacts in phase vocoder based (real time) signal transformations (Röbel, 2003a,b). This application requires a number of special features that distinguishes the proposed algorithm from general case onset detection algorithms: The detection delay should be as short as possible, frequency resolution should be high such that it becomes possible to distinguish spectral peaks that are related to transient and non transient signal components, for proper phase reinitialization the onset detector needs to provide a precise estimate of the location of the steepest ascend of the energy of the attack. In contrast to this constraints the application does not require the detection of soft onsets, where a soft onset is characterized by time constants equal to or above the length of the analysis window. This is due to the fact that such onsets are sufficiently well treated by the standard phase vocoder algorithm. False positive detections are not very problematic as long as they appear in noisy time frequency regions. A major distinction is that a single onsets may be (and very often is) composed of multiple transient parts, related either to a slight desynchronization of polyphonic onsets or due to sound made during the preparation of the sound (gliding fingers on a string). While these desynchronized transients are generally not considered as independent onsets they nevertheless constitute transients which should be detected for the intended application.

The evaluation of the transient detection algorithm for onset detection and music segmentation tasks has revealed that the detection results are comparable with existing algorithms for onset detection or signal segmentation tasks and, therefore, it has been implemented as a means for signal segmentation and onset detection in IRCAMS AudioSculpt application (Bogaards et al., 2004).

In the following article we will first describe the algorithm and comment on its features and expected performance, will then describe the optimal parameters that have been selected for the MIREX comparison and then discuss the evaluation results obtained for the MIREX 2005 audio onset detection task.

2 Fundamental Strategy

There exist many approaches to detect attack transients. For a number of current approaches see the other articles that have been proposed in the MIREX 2005 onset detection contest and furthermore Bonada (2000); Masri and Bateman (1996); Duxbury et al. (2002); Rodet and Jaillet (2001). In contrast to the evaluation of energy evolution in integral frequency bands, a criterion that most of the approaches are relying on, the following article proposes a two stage strategy which first classifies the spectral peaks in a standard DFT spectrum into peaks that potentially may be part of an attack transient and those that are not. Based on this classification a statistical model of background transient peak activity is employed to detect transient events. The advantage of this two stage approach is that the transient components of the signal are classified with rather high frequency resolution, allowing a precise distinction between transient and non transient signal components.

The basic idea of the proposed transient detection scheme is straightforward. A peak is detected as potentially transient whenever the center of gravity (COG) of the time domain energy of the signal related to this peak is at the far right side of the center of the signal window. Note, that it will be shown in section **5** that the COG of the energy of the time signal and the normalized energy slope are two quantities with qualitatively similar evolution and, therefore, the use of the COG of the energy for transient detection instead of the energy evolution appears to be of minor importance. Still it would be interesting to compare a COG based and amplitude slope based implementation of the algorithm.

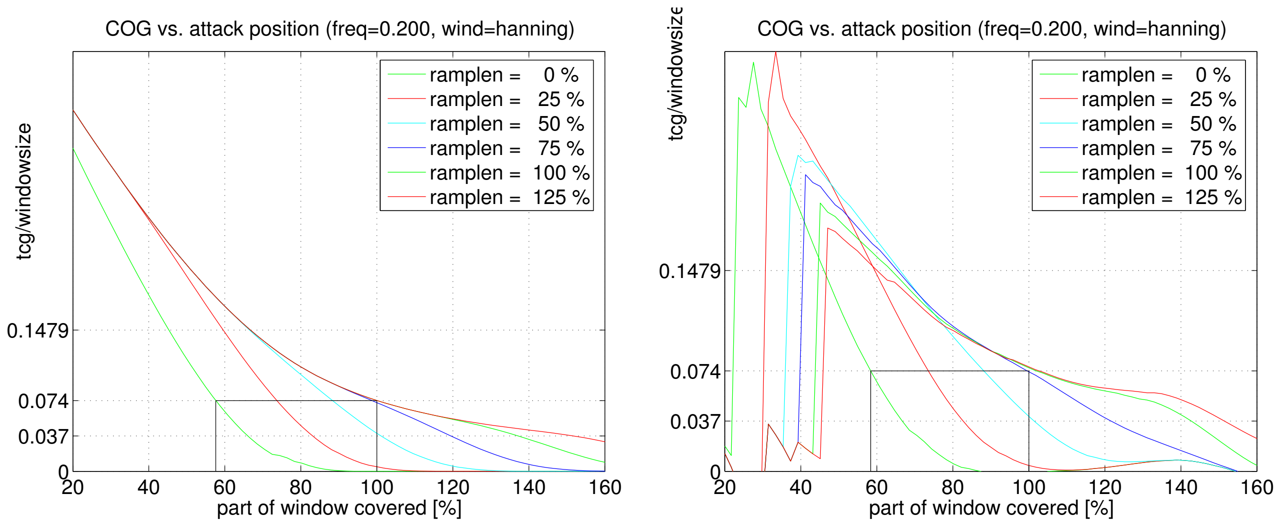

Figure 1: Center of gravity of partial energy (according to eq. (4)) as a function of transient position under the analysis window for transient partials with fixed frequency $w = 0.2\pi$ and different length of linear ramp (in percent of window size). Window type used is a Hanning window and SNR is $-\infty$ (left) and 10dB (right). The minimum thresholds C_e (see text) used for transient classification are marked.

3 Classification of spectral peaks

To understand the analysis procedure we investigate into the COG of a single transient sinusoid having an amplitude envelope representing a linear ramp with saturation. The signal is analyzed by means of moving the analysis window over the attack and performing a STFT. Without loss of generality we assume that the time origin is moving and is always in the center of the analysis window.

We denote the Fourier spectrum of the signal $s_h(t, t_m)$ which is the signal $s(t)$ windowed with the analysis window centered at time position t_m, $h(t, t_m)$, to be

$$S_h(w, t_m) = A(w, t_m) e^{j\phi(w,t_m)}. \tag{1}$$

Here w is the frequency in rad and $A(w, .)$ and $\phi(w, .)$ are the amplitude and phase spectrum respectively. As shown in Cohen (1995) the center of gravity (COG) of the instantaneous energy of the windowed signal $s_h(t, t_m)$ defined as

$$t_{cg} = \frac{\int t s_h(t, t_m)^2 dt}{\int s_h(t, t_m)^2 dt}, \tag{2}$$

can be calculated by means of

$$t_{cg} = \frac{\int -\frac{\partial\phi(w,t_m)}{\partial w} A(w, t_m)^2 dw}{\int A(w, t_m)^2 dw}. \tag{3}$$

The negative phase derivative, called group delay, determines the contribution of a frequency to this position. While equations (1) - (3) are derived for time continuous signals the same type of relations can be established for the DFT of discrete time signals where the integrations have to be replaced by summations and the differentiation with respect to frequency is understood to be performed using the properly interpolated DFT spectrum. The origin of the coordinate system for the sample positions has to be chosen consistently when calculating the DFT. Note, that the differentiation of the phase with respect to frequency, the group delay, is equal to the time reassignment operator, which can be calculated efficiently by means of a Fourier transform (or DFT) of the signal using a modified analysis window Auger and Flandrin (1995).

To be able to determine transient positions for sinusoidal components that are part of a conglomerate spectrum we need to modify the estimation of the COG such that it operates local in frequency. This is achieved by means of considering each spectral peak independently and limit the integral in eq. (3) to the frequencies located between the amplitude minimum surrounding each peak. Consequently, the COG is calculated using

$$t_{cg} = \frac{\int_{w_l}^{w_h} -\frac{\partial\phi(w,t_m)}{\partial w} A(w, t_m)^2 dw}{\int_{w_l}^{w_h} A(w, t_m)^2 dw}, \tag{4}$$

where w_l and w_h are the positions of the amplitude minima below and above the current maximum respectively. Due to the amplitude weighting taking place the difference between eq. (3) and eq. (4) will be small as long as the partial is sufficiently resolved. For sinusoids that are to close in frequency to be individually resolved the treatment of individual peaks performs a somewhat arbitrary signal decomposition which nevertheless will correctly detect transient situations as long as all the sinusoids that are contributing to the same peak are transient.

If the analysis window is moved from the left over the attack of a sinusoid the COG is first located to the right of the window center. Moving the window further to the right results in the COG moving to the left such that the absolute value of the phase slope is decreasing together with the bandwidth of the peak. Finally, the phase slope becomes zero and the peak reaches its minimum bandwidth if the window has completely moved over the attack transient in which case we have reached the stationary part of the sinusoid. In fig. 1 the decrease of the COG is shown that results if the analysis window moves over sinusoids having an attack transient of different ramp length. The analysis window that has been used in fig. 1 is a Hanning window, however qualitatively similar results are obtained for all other analysis windows. The ramp length of the attack phase is given in percent of the analysis window length and the window position is given in terms of the position of the right end of the window relative to the start of the attack. The window position is normalized by window length and expressed in percent of the window length. The left figure in fig. 1 shows the evolution of the COG in case of a pure sinusoid, while the right figure shows the COG if the sinusoid is embedded in white background noise with a SNR of 10dB relative to the stationary part of the sinusoid. In both cases the transient part of the sinusoid can be detected by simply thresholding the COG curve. The base threshold level C_e has been selected according to (Röbel, 2003a,b) such that the transient will be close to the signal center if the COG is close to C_e. If the attack takes place in background noise the increase of the COG is delayed due to the fact that the noise that fills the analysis window will offset the COG to the center of the window. Nevertheless, for a stationary sinusoid with amplitude of about 20dB above the background noise amplitude (and similarly for the attack of a chirp signal) the maximum COG still exceeds C_e.

4 From transient peaks to onsets

Unfortunately not every spectral peak detected as transient indicates the existence of a transient signal component. Further inspection reveals that spectral peaks related to noise signals quite often have a COG far of the center of the window. In contrast to spectral peaks related to signal attacks the transient peaks in noise are not synchronized between each others and this synchronization of a sufficient number of transient peaks will be our final means to avoid detection of noise peaks as transient events.

In the following we will extend the deterministic transient peak model described above by means of a statistical model that treats the randomly occurring transient events that are due to background noise or dense sinusoids as a background transient process. The stationary background noise needs to be distinguished from singular events related to a change of sound characteristics or beginning of a new note. The fundamental idea here is, that the average observed number of transient peaks should stay nearly constant as long as now attack transient takes place. In the later case the number of transient peaks should significantly increase, giving an indication about the onset. Because the statistical model should be able to indicate whether the state in a narrow frequency band has changed from non transient to transient, for example if a single sinusoid takes part in an onset, it appears to be favorable to operate the statistical model within frequency bands that will not cover the whole spectrum.

Therefore, to achieve the statistical description we divide the spectrum into overlapping frequency bands with equal bandwidth. For each band a statistical model is estimated that describes the average probability of a transient peak using a short history of F_h frames. To detect the singular transient events that are related to instrument onsets we compare this probability with the number of transient peaks in the next F_c frames. The statistical model is a simple binomial model describing the probability of a spectral peak to have COG $> C_s = KC_e$ with $K \geq 1$. As will be shown later in the experimental evaluation of the algorithm an increase in K decreases the sensitivity of the algorithm and is one of the major means to control the robustness of the detection.

For the estimation of the binomial model the number of independent events N of the statistical process is needed. There exist multiple choices to select this parameter. A first idea would be to use the number of observed peaks in each spectral band. This choice, however, would link the number of peaks to the confidence of the decision and a strong transient that creates a single wide band peak in the spectrum would always be biased by low confidence because the number of observed events becomes small. Therefore, a more sensible means to select N is the average number of peaks that may be contained in a frequency band given the current analysis window. A simple means to obtain N is to divide the bandwidth of the main-lobe of the spectrum of stationary sinusoid by the width of the spectral band and multiply the value by the number of frames, F_c or F_h respectively. A slightly more reasonable procedure would be to estimate the average number of peaks that will be contained in a frequency band if the observed signal is pure white noise. As long as the number of events does not change with time the exact value is not very important because its impact can be adjusted by the confidence level parameter introduced below. Therefore, for the following experiments the simpler method has been used.

A means to control the robustness of the detection is the confidence level required when testing for a change in the transient probability model between the frame history and the current frames. Using the formula for the variance of a binomial distribution with transient peak probability p

$$\sigma^2 = p(1-p)N \tag{5}$$

we want to select the transient probability such that it is consistent with the number of observed transient hits n in the frequency band within the range of G times the standard deviation of the mean value pN. Therefore, for p we require

$$n = pN \pm G\sigma = pN \pm G\sqrt{p(1-p)N}. \tag{6}$$

where the plus and minus sign are used to determine the transient probability for the current frames and frame history, respectively. Solving for p we obtain

$$p_c = \frac{G^2N_c + 2n_cN_c - G\sqrt{N_c(G^2N_c + 4n_cN_c - 4n_c^2)}}{2N_c(G^2 + N_c)} \tag{7}$$

$$p_h = \frac{G^2N_h + 2n_hN_h + G\sqrt{N_h(G^2N_h + 4n_hN_h - 4n_h^2)}}{2N_h(G^2 + N_h)} \tag{8}$$

where N_x and n_x are the number of independent events and observed transient peaks in the frame history (for $x = h$) and the current frames (for $x = c$), respectively. An attack transient is detected if in any of the frequency bands the transient probability in the current frames p_c is larger than the transient probability in the frame history p_h, because this means that there does not exist a single p

that can explain the observed transient peaks in both frame sets with the required confidence.

After having detected an attack transient we want to assemble all the transient peaks into a single event to be able then to make a more precise estimation of the transient position. Until the end of the attack event is detected all peaks that have a COG above C_e are collected into a set of transient bins. This set is non contracting and bins stay in the set even if their COG falls below the threshold. The attack is finished when the spectral energy of the bins having a COG above C_e in the current frame is smaller than half the spectral energy contained in the set of bins marked as transient.

4.1 Determining transient position

An attack transient event can be characterized in time by its start and end time. As start time of the transient we define the time location where the related signal energy becomes detectable. Having classified the spectral frames into transient and non transient bins it is now possible to reproduce the transient signal by means of removing all transient bins from the spectrum and transforming it back into the time domain. As a simple means to estimate a precise location of the end of the transient we search for the maximum absolute amplitude of the transient time signal. For the start of the transient we take the absolute value of the transient time signal before the detected end time and search for a minimum mean squared error representation by means of two line segments, the first one horizontal and the second one with arbitrary slope. The connecting point between these to segments is adapted such that a global minimum error of the line segment representation is obtained and the optimal connecting point is selected as start time of the transient signal. The start time of the transient signal has been used as transient position in the MIREX evaluation.

4.2 Transient energy ratio

As will be shown in the following sections the use of the transient detector as explained above yields good performance for the detection and preservation of transient events during the phase vocoder treatment. The detection parameters C_s and G are adjusted rather sensitive such that recall percentage is close to 100%. The somewhat large number of false positives does not have any negative impact because in nearly all cases the related signal is part of a noise signal and has rather low energy. The algorithm as described so far has been submitted for the MIREX evaluation under the name *Roebel, A. 1*. For the use as onset detector a number of improvements may be applied. The first one would be to require a minimum distance between two detected onsets keeping only the maximally significant onset if there are more within the given time window. This approach has not yet been tested. Another approach that is currently used for the application to signal segmentation is to filter the detected transient with respect to their normalized energy variation *(NEV)*.

As normalized energy variation we define the maximum of the ratio between total signal energy in a transient frame and the transient energy in the same frame, where the maximization is done over the whole duration of the onset. As defined the NEV is bounded between 0 and 1. Usually this filtering is done interactively by the user who can adapt the NEV threshold after the detection process as desired. For the current evaluation the threshold has been optimized using a set of training data and for algorithm *Roebel, A. 2* has been selected to NEV$= 0.35$ such that for the training data the maximum F-measure was achieved.

5 COG and energy slope criteria

As mentioned above transient detection algorithms are usually making their decisions based on the time evolution of the signal energy. In the following we show that the COG is closely related to the change of energy with time. From the theory of reassignment we know that the group delay is equal to

$$-\frac{\partial}{\partial w}\phi(w,t_m) = -\operatorname{real}\frac{\overline{S_h(w,t_m)}S_{hT}(w,t_m)}{|S_h(w,t_m)|^2} \tag{9}$$

where $S_h(w,.)$ and $S_{hT}(w,.)$ are the Fourier transforms of the signal s using the windows h and h_T centered at position t_m. The window h_T is obtained from the analysis window h by multiplication with a time ramp having its origin in t_m. If we calculate the derivative of the spectral energy $|S_h(w,t_m)|^2$ with respect to window position t_m and normalize the derivative by the spectral energy we obtain

$$\frac{\partial |S_h(w,t_m)|^2}{|S_h(w,t_m)|^2 \partial t_m} = -2\operatorname{real}(\frac{\overline{S_h(w,t_m)}S_{hd}(w,t_m)}{|S_h(w,t_m)|^2} \tag{10}$$

which besides a constant factor 2 can be derived from eq. (9) by replacing the Fourier transform using the window h_T by a Fourier transform using the window h_d which is the derivative of the analysis window with respect to time. Because h_d and h_T are qualitatively similar functions the group delay eq. (9) and the normalized derivative of the spectral energy eq. (10) will be similar functions as well.

6 Optimizing parameters

To give an indication of the precision/recall curves that can be obtained with the proposed onset detector we will discuss experimental results that have been performed on a set of training signals. This set has been hand labeled to obtain a ground truth for onset detection. The database contains a set of 17 sound signals with a total of 305 attack transients. For the following experiments the history size to estimate the back ground transient probability has been fixed to contain all frames that are covered by the analysis window. Because the window step is the eights part of the window the history always contains $F_h = 8$ frames. For estimating the actual transient probability we have experimented with $F_c = 1$ and $F_c = 2$. The results obtained with these two settings are qualitatively similar, however, with clearly better results for $F_c = 2$. Therefore, all following experiments will use this value.

There remain four user selectable parameters for the transient detector. The first one is the analysis window size. With respect to this parameter there exist contradicting demands because on one hand attack transients of

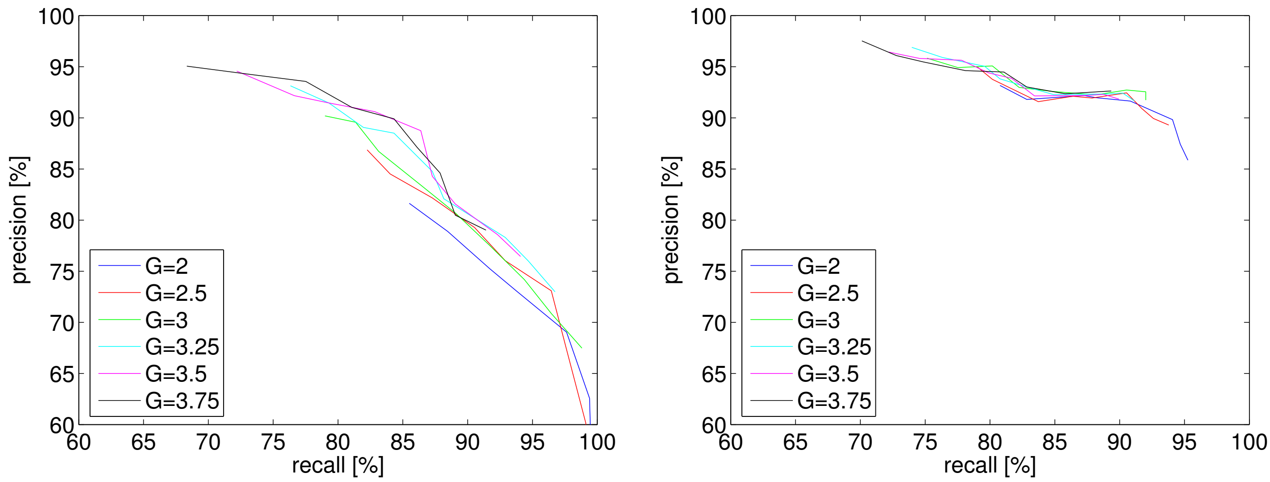

Figure 2: Comparison of relation between recall and precision ratio for transient detection with and without energy level filter. Size of the analysis window has been set to 57ms and the confidence factors G have been set as indicated in the legend. For all curves the transient threshold factor K varies in the range from $1.4 \leq K \leq 3$.

sinusoids that mix with stationary sinusoids will not be correctly detected such that frequency resolution should be high and window size large. On the other hand we can not detect more than one attack transient within a single window such that window size should be small. This is a variant of the well known time resolution/frequency resolution trade off for time frequency analysis. Aiming at the mirex comparative study a fixed window size had to be selected. For polyphonic signals a window of $50ms$ or more is generally preferable. Accordingly a window size of 2500 samples has been selected.

The second parameter is the threshold factor K. A simple theoretical investigation shows that for the noise free case the maximum COG normalized by the analysis window is 0.5 and for maximum robustness C_s should be close to this value. Due to background noise or preceding notes, however, part of the transient may be covered in real signals such that the maximum value of the observed COG will generally be lower than 0.5. As shown in section **5** the COG has a close relation to the energy derivative and we may understand the parameter C_s (or K) to control the jump in energy that is required for a transient to be detected. Therefore, the parameter K is a natural means to control the sensitivity of the detection algorithm.

The third parameter is the bandwidth of the frequency bands that are used to obtain the statistical model for background transient activity. By increasing the bandwidth we increase the reliability of the transient probability estimation, however, at the same time we increase the number of bins that have to be affected by a transient event to trigger the transient detector. The band width of the statistical model is expressed in terms of the number of independent events within the band. During the optimization procedure in preparation for the mirex evaluation we tried a bandwidths covering the range from 7 to 16 events per band. Unfortunately no clear relation between bandwidth and results has been established. The optimal bandwidth varies with the parameters K and G, the relations need further investigation. For the evaluation we simply selected two parameter sets that did work well with the training data. It turned out that for the case with and without NEV filter two different bandwidth were necessary to obtain optimal training performance. Without NEV filter a number of 7 events per band provided optimal performance while for the case with NEV filter 13 events per band was a slightly better selection.

The last parameter is the confidence factor G that is used to control the confidence in detecting a change in transient probability. G has been varied in the range $[2, 3.75]$. Depicted in fig. 2 are the recall and precision ratios obtained for the hand labeled training data set and for K ranging from 1.4 up to 3.

According to the MIREX evaluation procedure a transient has been considered correctly classified whenever the hand labeled transient was not further then 50ms away from the estimated transient start time. All other detections are counted as false. In the left part of fig. 2 the results for the simple transient detector without NEV filter is depicted. In the right part of the figure the same experiment with additional threshold for the NEV=0.35 is displayed. As expected, an increase in K as well as an increase in G increases the precision and reduces recall rate. Therefore an increase in K can be approximately compensated by an appropriate decrease in G. Precision and recall rate vary somewhat stronger if no NEV filter is used. In this case the optimal performance has been achieved for $G = 3.5$ and $K = 2.4$. Enabling the NEV filter reduces the variation of the results for the different parameter settings and at the same time significantly increases the performance. Due to the fact that the NEV filter removes a number of unreliable onsets the optimal transient confidence factor G and the threshold factor K are reduced to $G = 3$ and $K = 1.6$.

7 MIREX evaluation results

In the following section we are going to shortly discuss the results obtained with the two parameter settings during the MIREX onset detection contest.

A first remark concerns the significance of the outcome. Clearly the neural network based algorithm of Lacoste and Eck achieved best overall performance. This, however, is combined with a significantly increased processing time (between 6 to 25 times longer than the proposed algorithms). An important problem with this evaluation lies in the fact that each algorithm was trained on different training data (because no training data was available) and that only a single parameter set was allowed. A certain amount of luck was certainly required to win this competition given the fact that the selected training data may or may not have been similar to what was used in the contest. For our algorithm the precision and recall rates with the selected parameters were very similar, while in the contest they are off by more than 10%. This indicates a suboptimal algorithm working point and, therefore, it can be expected that a simple change of parameters should improve the performance.

Therefore, for further evaluation it is certainly essential to allow parameter variations during the competition, such that each algorithm can be adapted and for each algorithm only the globally best parameter set will be counted.

Considering now the places 3 to 5 it appears that all these algorithms achieve similar performance. The algorithm with NEV filter proposed in the current paper is at position 5 with only 0.16% difference to place 3. It is probably save to say that this difference is not significant. Still there is an important remark in order. The algorithm proposed in this paper has by far the largest number of doubled onsets. From the discussion of the algorithm this appears to be a consequence of the fact that close by transients are explicitly favored. It would be simple to improve the algorithm by means of preventing the number of close by detections, however, the open question especially with respect to the polyphonic and complex reference data would be, whether the ground truth reference is correct. For signal segmentation it is certainly undesirable to have multiple onsets triggered by slightly de-synchronized instruments, however, for signal separation, polyphonic f0 estimation or the present target application - transient preservation in the phase vocoder - the detection of slightly de-synchronized instruments may be a benefit. Besides that it appears also save to say that the NEV filter significantly improves the performance of the proposed algorithm.

Considering the different classes we may say that most of the results could have been more or less expected. A problem with the following comments is the fact that the sound data for the experiments is not available. Therefore, some of the following comments are based on guessing how the sounds may have been.

It is obvious that the proposed algorithm should work very well for impulsive attacks as for the solo bars and bells, drums and solo plucked strings. These signals have low background activity and will provide clear COG offsets for the resolved transient partials. Here the algorithm performs rather good. For the drums the algorithm achieves a very high recall with not as high precision, may be due to the number of doubled detections which is significantly higher than the average. Here again it would be interesting to see, whether the algorithm is really wrong or whether it was able to resolve drum beats that were rather close such that the reference did not consider them as distinct events.

It is by far the best algorithm for solo wind instruments which is to some extend a surprise. Probably the wind instruments also have clearly resolved partials with rather strong attack covered by a lot of wind noise which is a rather favorable environment for the high frequency resolution that the algorithm provides.

Expected from all experiments so far was the weak performance related to sustained strings. Often an onset in a sustained string is just a move in frequency with a nevertheless continuous partial trajectory. For this kind of signal the peak based COG will always stay close to the center of the window and detection cannot be expected. A surprise is the relatively good performance for the solo singing voice.

8 Acknowledgments

We would like to express our gratitude for the enormous amount of work the MIREX team invested into this contest. It would certainly be a very good idea to repeat the evaluation with a somewhat stabilized and improved procedure taking into account the weak points outlined above.

References

F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. on Signal Processing*, 43(5):1068–1089, 1995.

N. Bogaards, A. Röbel, and X. Rodet. Sound analysis and processing with audiosculpt 2. In *Proc. Int. Computer Music Conference (ICMC)*, 2004.

J. Bonada. Automatic technique in frequency domain for near-lossless time-scale modification of audio. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 396–399, 2000.

L. Cohen. *Time-frequency analysis*. Signal Processing Series. Prentice Hall, 1995.

C. Duxbury, M. Davies, and M. Sandler. Improved time-scaling of musical audio using phase locking at transients. In *112th AES Convention*, 2002. Convention Paper 5530.

P. Masri and A. Bateman. Improved modelling of attack transients in music analysis-resynthesis. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 100–103, 1996.

A. Röbel. A new approach to transient processing in the phase vocoder. In *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*, pages 344–349, 2003a.

A. Röbel. Transient detection and preservation in the phase vocoder. In *Proc. Int. Computer Music Conference (ICMC)*, pages 247–250, 2003b.

X. Rodet and F. Jaillet. Detection and modeling of fast attack transients. In *Proc. Int. Computer Music Conference (ICMC)*, pages 30–33, 2001.