Real-Time Uses of Low Level Sound Descriptors as Event Detection Functions Using the Max/MSP Zsa.Descriptors Library

Mikhail Malt[1], Emmanuel Jourdan[1]

[1] Ircam – Institut de recherche et coordination acoustique/musique
1 place Igor Stravinsky, 75004, Paris France

{mikhail.malt,emmanuel.jourdan}@ircam.fr

Abstract. *This paper is a continuation of the research and development we began last year [Malt and Jourdan 2008] on the study and the use of audio descriptors for real-time performance of electroacoustic mix music, and as tools for computer-assisted musical analysis. Our main goal, in this paper, is to propose easy and efficient strategies for event detection in the context of real-time mix music (acoustic and electroacoustic music). We will examine three cases of the use of audio descriptors to build event detection functions: spectral slope, spectral standard variation and the construction of compound descriptors.*

1. Introduction: Event and Onset Detection in Real-Time Electroacoustic Music

Current compositional practice often involves the use of unconventional or extended instrumental techniques (e.g. multiphonics, blowing into an instrument, using key clicks as part of the instrument's vocabulary, etc.). In the context of the mix music itself, we find ourselves in a situation where our main task is no longer a matter of detecting simple pitch onsets, but dealing with considerably more complex event detections in difficult contexts. These new contexts may involve large variations in amplitude by register, considerable variation in terms of the onset of sound, the need to separate a sound from background "noise," and problems with multiple-microphone sound capture of instruments whose patterns of sound radiation are atypical (e.g. the bassoon). In addition, there are also situations where event detection based on a timbre variation (bisbigliando) is more complex and requires specific techniques. Taken as a whole, these factors complicate the task of event detection considerably.

In this paper we propose to speak more in terms of "event detection" than "onset detection" - we are concerned not only with note onset detection, but also any kind of musical events that could be perceived as a discontinuity within a musically static flow, such as sound inflections or noisy playing techniques.

The event detection methods currently available for real-time purposes are mainly based on amplitude, pitch or spectral variation. All of methods are strongly affected by background noise and strongly dependent upon such external factors as microphone type, distance from sound sources and especially room effects.

As onsets, sound inflections and event detection are dependent on a large number of signal parameters variations such as amplitude, spectral brightness, spectral standard variation, spectral slope, spectral onset slope, roll-off point, and spectral envelope. It is important at this point to mention that all these parameters are not fully independent and that some of them are highly correlated. Bearing this in mind, we will propose three experiments in real-time event detection based on the use of audio descriptors.

A large part of our assumptions and hypotheses derive from our own pragmatic observations in building event detectors for different musical contexts, and from our own systematic study of descriptor signals coming from different kinds of musical materials. Our intention in this paper has a more pragmatic and ongoing aspect to it, as well - we are looking forward to planning more systematic experiments to compare and to find the limits of the methods we are going to propose. In the meantime, very good reviews and comparisons of onset detection functions could be found in Bello [Bello and all 2005] and Collins [Collins 2005].

2. Event Detection by Spectral Slope

The spectral slope is an estimation of how quickly the spectrum of an audio signal decreases towards the high frequency range, and is commonly computed using a linear regression on the magnitude spectra. In the recent version of the *Zsa.descriptors* library, we implemented an algorithm which is slightly different from that proposed by Peeters [Peeters 2004, p. 14], but is analytically equivalent. It is based on covariance and variance computation of the frequency bins and squared amplitudes (energy) in an FFT frame.

$$slope[t] = \frac{\mathrm{cov}(f, a^2)}{\mathrm{var}(f)} \quad (1)$$

Where:

a is the linear amplitude vector frame,

f the frequency vector frame

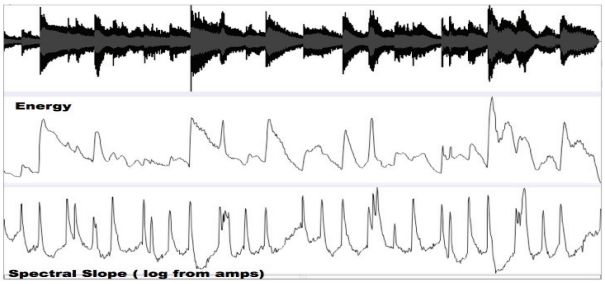


Figure 1. Onset detection function comparison (we have used the slope logarithm for the computation[1]).

What we have observed is that the spectral slope is less affected by amplitude variations and background noise than a signal coming from amplitude onset detection,

[1] This image was done with the *ftm.editor* from FTM library (© Ircam, IMTR team).

and has more well-defined peaks. Using the slope logarithm gives us even more precise peaks (Figure 1).

2.1. Preprocessing Technique

For certain applications we found it useful to introduce an adaptive signal level scaling in order to compensate for any level change (Figure 2). In fact, we used four parameters to control the preprocessing: a reference dB value (mean level to be maintained), a minimum dB to trigger the process (a signal below the trigger threshold will multiply the signal by zero) and a frequency to control a low pass filter to smooth the root mean square amplitude used to control the scaling and the gate. The final signal $S_{adapt}[n]$ will be:

$$S_{adapt}[n] = S[n] * K_{linear} \quad (2)$$

Where:

$S[n]$ is the audio signal,

$A_{rms}(S[n])$ is the Root Mean Square amplitude of $S[n]$,

$A_{rmslp}[n] = aA_{rms}(S[n]) + bA_{rms}(S[n-1])$ is the smoothed $A_{rms}(S[n])$ signal, by a low pass filter,

A_{rms-dB} is the smoothed Root Mean Square amplitude of $S[n]$ in dB,

A_{ref-dB} is a reference dB value, this means, the average dB value we want reach with our input signal,

$A_{\min-dB}$, is the minimum level, in dB to trig the process,

$\Delta_{amp\_ref} = A_{ref-dB} - A_{\min-dB}$, is the is the dB difference between the reference value and the dB minimum reference value,

$\Delta_{amp} = A_{ref-dB} - A_{rms-dB}$ is the dB difference between the reference value and the actual signal amplitude, for

$$K = \begin{cases} 0, & \Delta_{amp} > \Delta_{amp\_ref} \\ \Delta_{amp}, & \Delta_{amp} \le \Delta_{amp\_ref} \end{cases} \quad (3)$$

and $K_{linear} = 10^{\frac{K}{20}}$

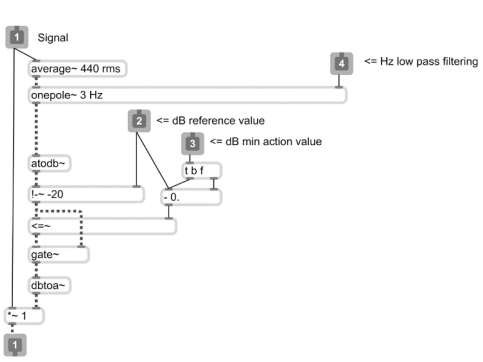
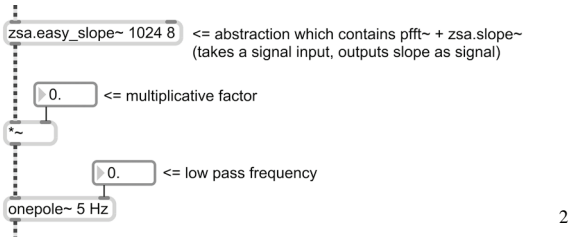


Figure 2. Adaptive signal level scaling, Max/MSP implementation

2.2. Building an Event Detection Function

The detection function was built using the *zsa.slope~* object [Malt and Jourdan 2008] (Figure 3) with a negative multiplicative factor (to turn the data positive) and a low pass filter to smooth out the data.



[2]

Figure 3. Spectral slope detection function

2.3. Peak Selection Technique

We propose a standard process shown in Figure 4 as a peak selection technique: A discrete derivative (we used samples steps for the derivative calculation) with a low pass filter to smooth the signal, a multiplicative stage to constrain the flow to {-1, +1} and threshold detection using standard Max/MSP objects (*thresh~* and *edge~*) with a final stage to control unwanted repetitions (see Figure 5).

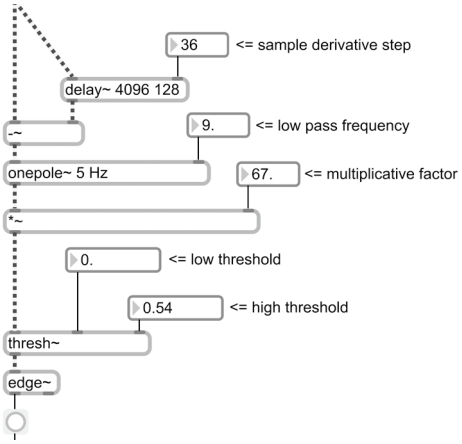


Figure 4. Peak picking technique

2.4. Avoiding Unwanted Repetitions

The avoidance of unwanted repetitions is critical, and can be easily implemented using the *onebang* object in conjunction with a *delay* object (Figure 5).

[2] In this paper we used encapsulated versions of "zsa.*" standard objects. The "zsa.easy_*" and "pfft~zsa.abs_*" abstractions are found in the Max/MSP zsa.descriptors library (http://www.e--j.com/?page_id=83).

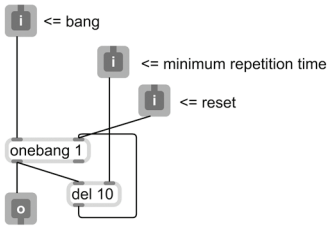


Figure 5. Delayed gate to avoid unwanted repetitions

2.5. Practical Use of Spectral Slope

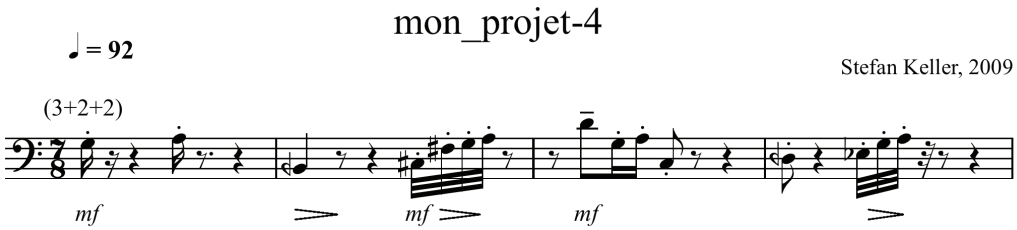


Figure 6. First measures of *mon_projet-4* from Stephan Keller

The sound radiation of the bassoon is heavily dependant on the register being played and therefore usually requires multiple microphones to obtain a satisfactory sound capture for the whole tessitura. Even with such a system, amplitude detection is further strongly affected by the movements of the performer. The spectral slope is correlated with the spectral envelope, and its variation is more related to the variation of the shape than it is related to the amplitude variation. Since we needed a highly compressed signal to obtain a good amplification of the instrument in this specific case, the calculation of the spectral slope clearly gave better results due to the fact that it was less dependant on variations in the amplitude of the signal from the microphone.

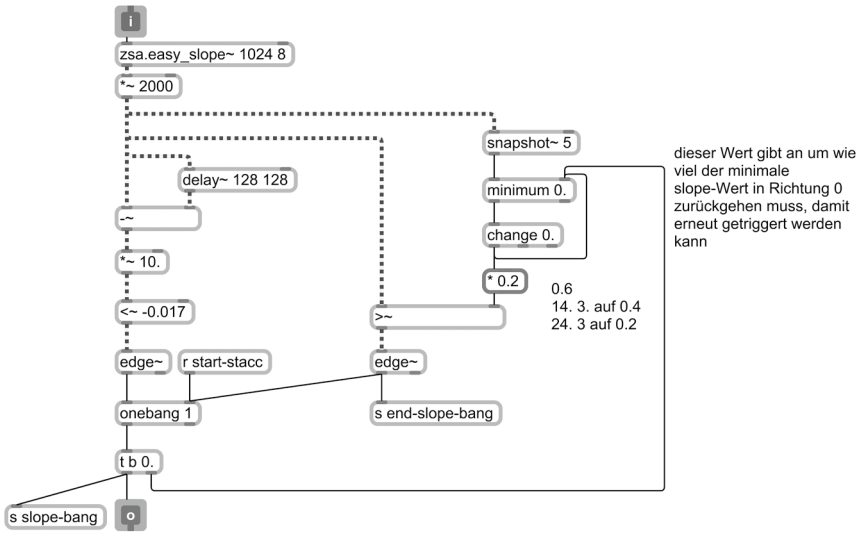


Figure 7. Spectral slope onset detection, implementation, by Stephan Keller

We found that the use of spectral slope also produced similar results when the Italian composer Franscesca Verunelli did some experimentation to detect accordion inflections going from very low dynamics to strongest ones as part of his work in the

Ircam Cursus. As was the case with the bassoon, the use of spectral slope proved to be more robust than a single amplitude follower and less susceptible to background noise.

3. Event Detection by Spectral Standard Deviation

Event detection functions that use spectral standard deviation have showed to be very useful as a technique for detecting noisy events in complex musical situations where different sound materials are mixed together. Any noisy event (e.g. key clicks) can be easily differentiated from more harmonic material; for example, a flute returns standard deviation values around 400-500 Hz (due to the player breath), while a key click returns standard deviation values around 1500-2500. An instrument that produces rich spectral data such as an oboe or violin will return standard deviation values in the1000-1500 Hz. Frequency range. A violin Bartok pizzicato could return values beyond 3000 Hz.

As you would expect, we use the spectral centroid as the first moment of spectra, considered as a frequency distribution, which is related with the weighted frequency mean value. The spectral spread is considered as the second moment - the variance of the mean calculated.

$$v = \frac{\sum_{i=0}^{n-1} (f[i] - \mu)^2 a^2[i]}{\sum_{i=0}^{n-1} a^2[i]} \quad (4)$$

Where:

n is the half of the FFT window size,

i the bin index,

$a[i]$ is the amplitude of the bin i, in the real magnitude spectra of the FFT calculus and

$f[i]$ is the frequency of the bin i. Where:

$$f[i] = i * \frac{sample\ \ rate}{FFT\ \ window\ \ size} \quad (5)$$

and μ is the spectral centroid in Hertz.

3.1. Building an Event Detection Function

We used the spectral standard deviation (the variance square root, gated by a "K" factor) to build our event detection function. The event detection function is defined as:

$$event\_Stdr\_function = \sqrt{v} * K \quad (6)$$

Where the gate factor K is defined as follow:

$$K(A_{rms-dB}) = \begin{cases} 1 \ ; \ A_{rms-dB} \geq A_{\min-dB} \\ 0 \ ; \ A_{rms-dB} < A_{\min-dB} \end{cases} \quad (7)$$

A_{rms-dB} is the rms smoothed signal level in dB and $A_{\min-dB}$ is the threshold dB value.

The use of the K gate is very important to avoid in silent musical passages, where the quick increase of standard deviation may be usually due to noisy flat spectra. The means

of avoiding this problem is implemented in the *level_gate_scaling* subpatch (Figure 8, Figure 9)

Given the nature of the sound material and the presence or absence of interferences, it might be useful to smooth the standard deviation signal with a low pass filter (for situations such as detecting blow playing in woodwinds), but if we have a good sound capture and adequate percussive event detection, it could be avoided.

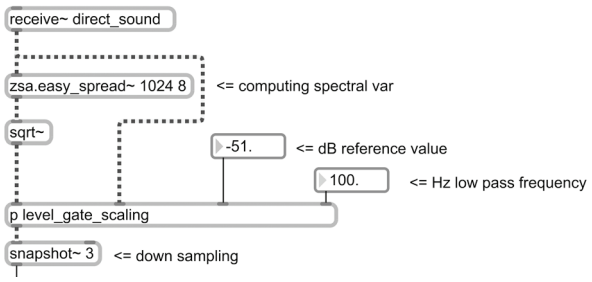


Figure 8. Gated spectral standard deviation, event detection function

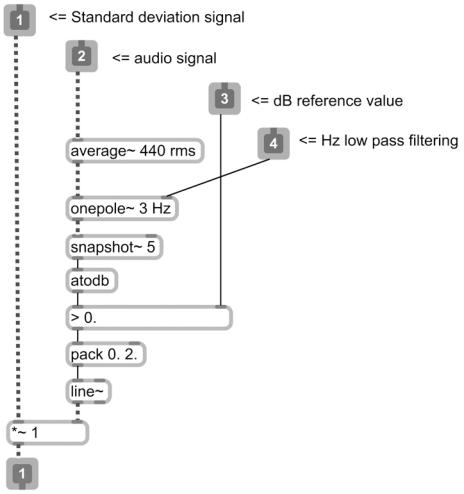


Figure 9. Level gate scaling

3.2. Peak Selection Techniques

For a peak-selection technique we used the same standard technique as for the onset event function, which uses the spectral slope (Figure 4, Figure 5).

3.3. The Practical Use of Spectral Standard Variation

The Italian composer Danielle Ghisi used this technique successfully within the context of the piece "*Comment pouvez vous lire à present ? Il fait nuit*." (for Alto Sax and real-time electronics) created at Ircam's Espace de projection in March, 2009. The technique was used to detect key clicks in the last section of his piece in measures 90 to 94 (see Figure 10) using the spectral standard variation to trigger various real-time processes.

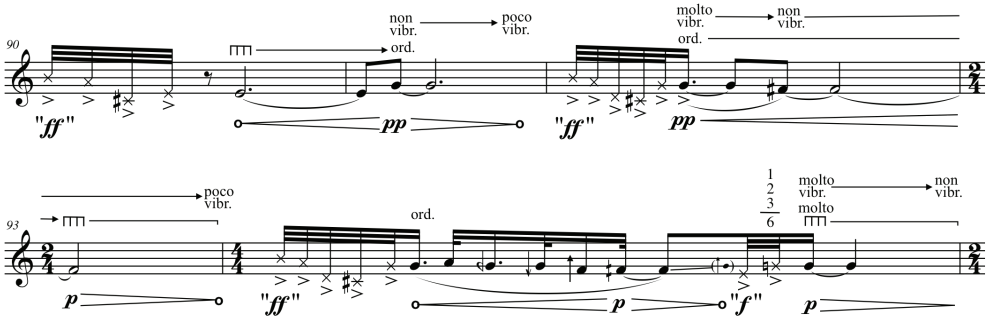


Figure 10. *Comment pouvez vous lire à present ? Il fait nuit,* measures 90 to 94, from Danielle Ghisi

Figure 11 shows the spectral standard deviation signal, the derivative and the onset detection of the first gesture (four key clicks followed by an E4-G4 in crescendo-decrescendo) of measure 90. The difference between the standard deviation from key clicks (around 1500-2000 Hz) and for the E4-G4 (around 150 Hz) is quite obvious and shows the advantage of this technique to detect noisy events.

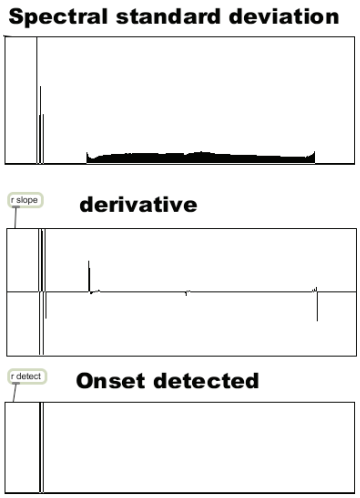


Figure 11. Visualization of spectral standard deviation event detection function, the discrete derivative and the event detection

4. Event Detection by Compound Function

The next experiment in event detection involves the use of a function based on a compound descriptor. In spoken sounds, we have observed that the fricative consonants tend to have a high centroid and a high spectral standard deviation.

$$D[n] = \left(\left(\frac{\mu}{c1} \right) * \left(\frac{\sigma}{c2} \right) * K \right)^2 \quad (8)$$

where:

μ is the spectral centroid,

$c1$ is a constant set in order to normalize the μ value,

σ is the spectral standard deviation,

$c2$ is a constant set in order to normalize the σ value and

K is defined as

$$K(A_{rms-dB}) = \begin{cases} 1 \; ; \; A_{rms-dB} \geq A_{\min-dB} \\ 0 \; ; \; A_{rms-dB} < A_{\min-dB} \end{cases} \quad (9)$$

The K variable is defined in order to avoid side effects in silent passages (see item 3.1).

The main expression (7) maximizes the spectral centroid and standard variation.

4.1. Building an Event Detection Function

The expression (6) is implemented as it is shown in Figure 12, using a *pfft~* object as the core of the process, with two objects sharing the same FFT and energy calculation (Figure 13). To limit the noise impact, low pass filters smoothed the two descriptors signals.

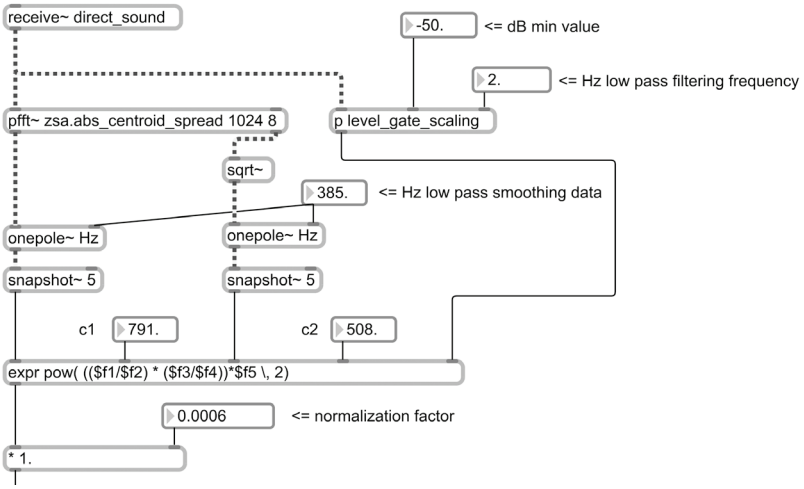


Figure 12. Compound detection function with centroid and spectral standard deviation

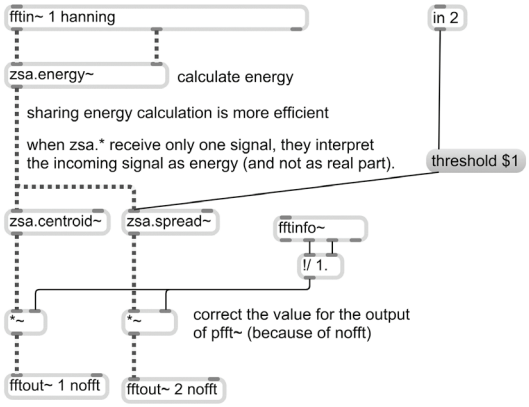


Figure 13. Patcher detail from *pfft~* object from Figure 12

Figure 14 shows the K variable implementation, where we used an RMS level smoothed by a low pass filter.

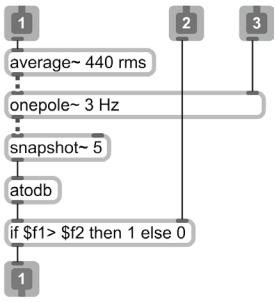


Figure 14: Patcher implementing the K variable

4.2. Peak Selection Techniques

In this experiment, we used peak selection technique with an adaptive threshold (Figure 15).

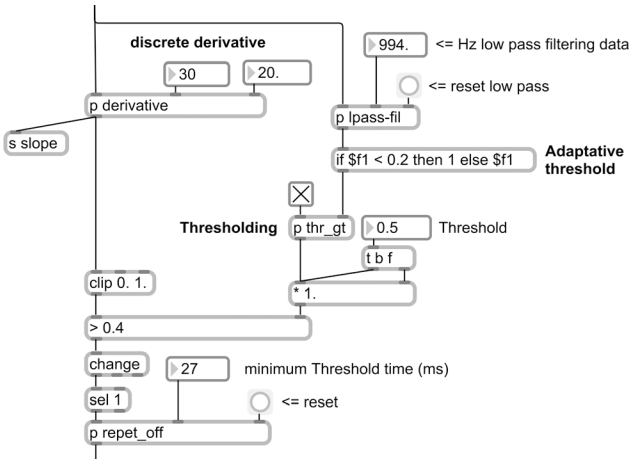


Figure 15. Adaptive threshold

First, the data flow coming from the detection function is derived (in a discrete way, see Figure 16), returning the function $D'[n]$.

$$D'[n] = \frac{\Delta D[n]}{\Delta n} = \frac{D[n+m] - D[n]}{m} \quad (10)$$

Where:

$D[n]$ is the event function at index n, and m is a window size.

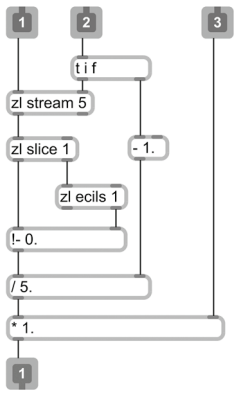


Figure 16. Discrete derivative from expression (10)

The result passes by a threshold process where the reference threshold value δ_{ref} is weighted by a smoothed version of the detection function. We used a low pass filter (Figure 17). The adaptive threshold took the shape of expression (9)

$$\delta[n] = \delta_{ref} * \left(aD[n] + bD[n-1]\right) \quad (11)$$

Where:

$\delta[n]$ is the adaptive threshold at time n,

δ_{ref} is the reference threshold value,

$D[n]$ is the descriptor value from expression (7) at time n, and

a,b are the coefficients for the low pass filter.

Our final detection function $\sigma[n]$ will take values of zero or one.

$$\sigma[n] = \begin{cases} \sigma[n] = 1; D'[n] \geq \delta[n] \\ \sigma[n] = 0; D'[n] < \delta[n] \end{cases} \quad (12)$$

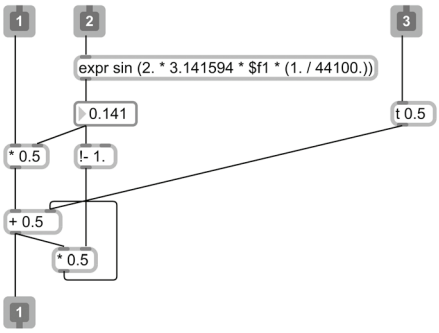


Figure 17. Low pass filter from expression (11)

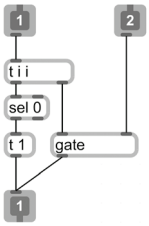


Figure 18. Gate to choose between adaptive or static threshold

5. Conclusions and Perspectives

The use of sound descriptors appears to be a useful and advantageous tool for musical event detection in real-time music. The methods we described were used as pragmatic alternatives to event detection in complex cases where the usual techniques for event detection did not perform sufficiently well. In those cases, the use of low-level spectral audio descriptors proved to be satisfactory and robust. These promising results strengthen our motivation to continue our research in this direction.

We are looking forward to improve the number of *Zsa.descriptors* modules, to add high-level descriptors, and develop to systematic means to compare the efficiency of the various event detection functions for real-time process.

6. Acknowledgments

We would like to thank Gregory Taylor and David Coll for their valuable remarks, comments and suggestions.

7. References

G. Peeters, A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Cuidado projet report, Institut de Recherche et de Coordination Acoustique Musique (Ircam), 2004.

M. Malt and E. Jourdan, "Zsa.Descriptors: a library for real-time descriptors analysis", in 5th Sound and Music Computing Conference, Berlin, Allemagne, 31th july to August 3rd, 2008, p. 134-137. See http://www.e--j.com/?page_id=83.

X. Rodet and F. Jaillet, "Detection and modeling of fast attack transients," in Proc. Int. Computer Music Conf., Havana, Cuba, 2001, pp. 30–33.

I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals," in Proc. 14th Int. Conf. Digit. Signal Process. (DSP2002), vol. 2, Santorini, Greece, Jul. 2002, pp. 967–970.

D. Schwarz, Data-Driven Concatenative Sound Synthesis, PhD Thesis in Acoustics, Computer Science, Signal Processing Applied to Music, Université Paris 6 - Pierre et Marie Curie, January 20, 2004.

J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler, "A Tutorial on Onset Detection in Music Signals", in IEEE Transactions on Speech and Audio Processing, Volume 13, Issue 5, Sept. 2005, p. 1035 – 1047.

N. Collins, "A Comparison of Sound Onset Detection Algorithms with Emphasis on Psychoacoustically Motivated Detection Functions". Proceedings of AES118 Convention, 2005.

A. Klapuri, "Sound Onset Detection by Applying Psychoacoustic Knowledge", in Proc. IEEE Conf. Acoustics, Speech and Signal Proceesing (ICASSP,'99), 1999.

X. Rodet and F. Jaillet, "Detection and modeling of fast attack transients", in ICMC'01, La Habana, Cuba, Sept. 2001.

L. O. Nunes, P. A. A. Esquef, and L. W. P. Biscainho, "Evaluation of Threshold-Based Algorithms for Detection of Spectral Peaks in Audio," in Proc. 5th AES-Brazil Conference, São Paulo, Brazil, May 2007, pp. 66-73.