Acústica 2008
20 - 22 de Outubro , Coimbra, Portugal
Universidade de Coimbra

KNOWLEDGE-BASED ONSET DETECTION IN MUSICAL APPLICATIONS

Norberto Degara-Quintela, Antonio Pena, Manuel Sobreira-Seoane, Soledad Torres-Guijarro

Universidade de Vigo
Departamento de Teoría do Sinal e Comunicacións
ndegara@gts.tsc.uvigo.es, apena@gts.tsc.uvigo.es , msobre@gts.tsc.uvigo.es, marisol@gts.tsc.uvigo.es

Abstract
Onset detection is useful in a number of applications for audio signals. The goal of this paper is to present a combination of techniques using a blackboard system, an approach taken from expert systems. The proposed system defines a global sound source separation system that allows for the integration of multiple audio analysis methods. In this particular case, experts integrate onset detection algorithms that use their own analysis strategy. A voting scheme of combination of algorithms is presented.

Keywords: onset detection, blackboard systems, redundant approach.

1 Introduction

A sound signal can be modeled as an event-based phenomenon and the importance of identifying these events is clear. Actually, automatic detection of events in audio signals is of interest in a number of applications including sound source separation, music information retrieval and automatic music transcription.

Finding automatically the starting time (onset time) of these audio events is a difficult process. Audio events are typically related to short bursts of energy in time (percussive onsets), local changes in the spectral content of the signal (pitched onsets) or a complex combination of others. Onset detection methods are usually based on the calculation of signal features that manifest the presence of transients in the signal using either signal analysis techniques or probabilistic models. Hence, a single method is only appropriate for those signals that show the specific features the method was designed to detect. In addition, a precise fine-tuning of the analysis parameters of each method is usually needed in order to accurately determine the presence of onsets and this parameter selection is strongly signal-dependent. Consequently, the performance of current detection methods is highly dependent on the nature of the signal to be analyzed as shown by Bello et al. in [1].

It is not expected that a single method will perform accurately for strongly nonstationary signals and audio signals are intrinsically variable in nature. In this sense the development of appropriate tools to analyze nonstationary signals is of great interest. Instead of designing a very complex algorithm, a promising development known as the redundant approach lies in the combination of a number of different and simple techniques. In fact, this is most likely the way human perception works [2]. It

seems that human perception uses different processing principles for the same purpose, and when one of them fails perhaps another succeeds.

There have been several attempts of combining signal cues from different detection functions to provide with a more accurate estimation of onsets. A hybrid scheme that considers energy changes in high-frequency bands and spectral changes in lower bands is proposed in [3]. In [4] note onsets are classified into hard and soft onsets, using an energy-based detection function for hard onsets and a pitch-based detection for soft onsets. This simple combination method obtained the first place in the last audio onset evaluation task of the Music Information Retrieval Evaluation eXchange (MIREX).

A simple alternative for combining different algorithms is to use a voting mechanism as shown in [5]. In this approach several tempo induction algorithms are compared and a redundant tempo induction approach is introduced. Algorithms are ordered in a list and each algorithm gets one vote from all the algorithms that agree with its estimate. The estimate of the algorithm with the largest number of votes is selected as the output. Another interesting way to face the problem of combining a number of algorithms is to adopt a mixture-of-experts architecture, a machine learning approach [6]. Different onset detection algorithms can be specialized on a restricted set of signals by doing a fine-tuning of the algorithm analysis parameters.

In a redundant approach, where many algorithms are combined to accomplish the same goal and interact to adapt its behavior to the input signal, the architecture plays an essential role. In that respect, blackboard modeling, an approach taken from expert systems, has been successfully applied in the field of audio processing to relevant applications such as speech understanding systems [7], computational auditory scene analysis [8] and polyphonic music transcription [9]. In a blackboard model, experts communicate using a common database allowing to pursue multiple lines of reasoning at the same time and to adapt the strategies to a particular problem context [10]. Most of the applications where a blackboard has been adopted use top-down reasoning, however the parameters of the signal processing analysis are fixed. On the contrary, Klassner [11] proposed an adaptive selection of the parameters of a time-frequency analysis based on predictions from previous analysis and source models. Nevertheless, the line of analysis is unique, meaning that all the algorithms use the same frame length and overlap values.

In [12], Goto presents a multi-agent architecture where agents maintain a separate segmentation scheme but interaction between agents is managed directly by a control method.

The goal of this paper is to present a combination of techniques using a blackboard-agent system. In order to be able to follow multiple lines of signal analysis at the same time a blackboard-agent model is defined. Each system communicates as an agent and process as a blackboard. An agent interacts with other agents to cooperate and adapt to the current situation using an additional agent, the blackboard manager. The proposed system, a development of [13], defines a sound source separation system that allows the integration of multiple audio analysis methods such as beat tracking, multiple fundamental frequency estimation, partial tracking, etc.

In the particular case of transients, agents integrate onset detection algorithms and use their own strategy: each expert has its own instant-time of analysis, frame length, overlap and threshold values, and adapts these parameters interacting with other experts.

A redundant approach is then chosen, what implies diversity in the design and tuning of the different algorithms. A voting scheme of combination of algorithms will be defined.

2 A Blackboard-Agent Architecture

Audio content analysis is a complex problem that involves both signal processing data and high level knowledge. Whereas data-driven analysis comprises a bottom-up flow of information, prediction-driven processing uses sound source models and prior knowledge to create expectations and originate top-down flow of information. Integration of these different processing principles is difficult but necessary when facing a complex problem such as sound source separation.

A redundant approach combines many analysis techniques to accomplish a common goal. Multiple analysis algorithms may interact following both bottom-up and top-down principles going back and forward in time. Hence, the architecture is essential to be able to manage such a number of complex interactions.

A blackboard model, an approach taken from the field of artificial intelligence, has been chosen as the basis for the sound source separation system. The blackboard problem solving model is usually illustrated as a group of experts working cooperativley around a physical blackboard to solve a problem. The experts watch the blackboard, looking for an opportunity to apply their expertise to the developing solution.

A blackboard model is then composed of three components:

1. The blackboard, a global database that contains the hypotheses (data and feasible partial solutions).
2. A set of knowledge sources (KSs), independent entities that represent the problem–solving knowledge. KSs analyze the state of the blackboard and create or modify hypotheses. KSs are independent and communication between KSs only takes place with the creation and modification of hypotheses on the blackboard.
3. A control system that conducts the incremental problem-solving strategy and opportunistically selects actions.

There are several advantages in following a blackboard strategy. First, blackboard modelling provides flexibility in structuring complex problems since it supports incremental development of solutions. It also allows to pursue multiple lines of reasoning concurrently at the same time, adapting its strategies to a particular problem context. This opportunistic approach to problem solving means that the system may select at each moment the best-suited action to reach a final goal, combining both bottom-up and top-down processing. An additional benefit is that blackboard systems are intrinsically extensible. Actually, extending a blackboard system is as simple as defining a new action module. Therefore, blackboard systems provide the platform with the flexibility of incremental design which allows to add new analysis techniques when these are discovered. A description of the encapsulation and software interface of this blackboard architecture is presented in [13].

2.1 Generic Architecture

There are many issues that must be addressed in order to apply blackboard models successfully. In particular, effective control of selecting the actions is critical for this problem solving strategy. As a result, control is a major concern in the formulation of the blackboard architecture. Most of the sound applications where a blackboard has been used do not take full advantage of the blackboard capabilities. The reason seems to be that developing an effective control system is extremely difficult. Therefore, to easy the definition of a control system, the proposed architecture supports the definition

of multiple agents, where each agent is a blackboard and specific control strategies are enclosed into different agents.

Figure 1 shows the generic architecture for the sound separation system. This architecture defines a multi-agent federated system where one manager agent coordinates the interaction between other agents. An agent can be seen as a knowledge source for the manager blackboard. Hence, agents interact with other agents to cooperate and adapt to the current situation using a manager agent.

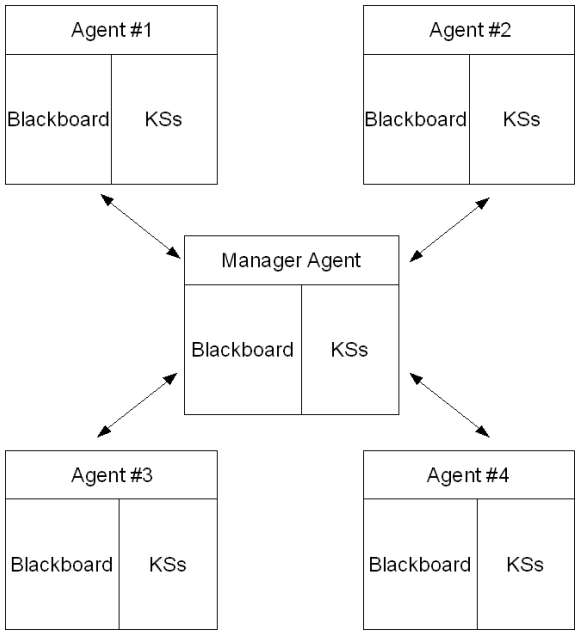

Figure 1 – Generic architecture.

In a sound source separation problem, this system would allow the integration of multiple audio analysis methods such as beat tracking, multiple fundamental frequency estimation, partial tracking and onset detection techniques into several agents. An analysis agent could manage the interaction of all these methods. In addition, classification and synthesis methods could be defined by other agents. A separation manager agent could control the action of the analysis, classification and synthesis subsystems.

2.2 Blackboard Architecture

In a redundant approach multiple techniques for the same objective act together. These techniques usually include the execution of a set of knowledge sources defining a line of reasoning (LOR)[1]. In addition, any line of analysis can be split into several sub-lines. Therefore, hypotheses and signal processing data store the identity of the LOR that created or modified that hypothesis. Interpretation knowledge sources analyse the lines defined in a redundant approach, integrating knowledge and potentially creating forward and backwards expectations.

[1] The terms *line of analysis* and *line of reasoning* will be used equivalently.

A key aspect of this architecture is that each LOR includes a set of parameters that define the signal processing analysis. The control system dynamically modifies the response of a LOR by modifing the analysis parameters as a response to distorted data, expectations or changes in the characteristics of the input signal. These parameters are defined by the control parameters of the knowledge sources the line executes. Hence, there are two levels of abstraction for a knowledge source: the generic knowledge source and the knowledge source instance. A KS instance is described by specific values of the control parameters of a generic KS. Examples of these control parameters are the current time index of analysis, frame length and overlap of a segmentation knowledge source or the threshold used by a peak picking method.

Figure 2 shows the generic blackboard architecture. Solid arrow lines indicate lines and sub-lines of reasoning. Dotted arrow lines indicate control decisions that the planner makes. There are two lines of analysis in this example and the line of analysis 1 creates two sub-lines of analysis at the highest level of the blackboard, sub-line 1-1 and sub-line 1-2.

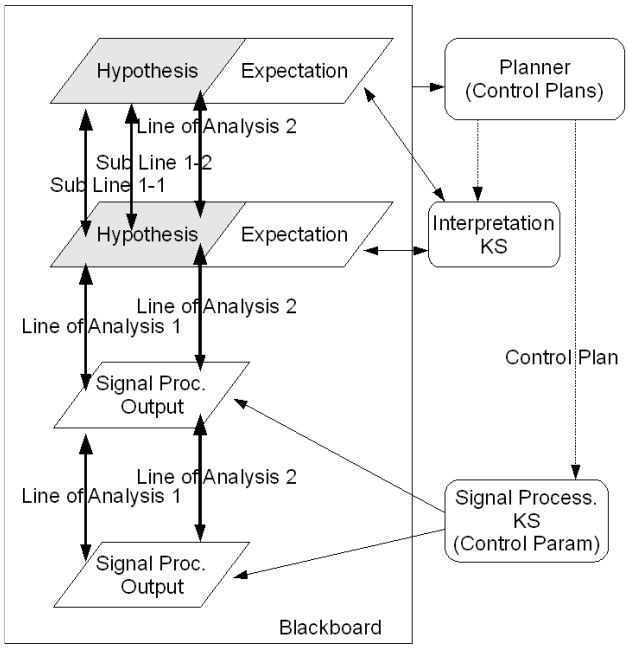

Figure 2 – Blackboard architecture and lines of analysis.

3 Example: A Redundant Approach to Onset Detection

In this section a redundant approach to onset detection is presented. There is a single blackboard that controls several onset detection methods and integrates the information obtained from these methods. This onset detection blackboard could be integrated in a global sound source separation system and managed by an analysis blackboard.

The blackboard system's architecture is shown in Figure 3. Four levels of information can be defined in this simple blackboard: *segments*, *spectra*, *detection functions* and *peaks*. The hypothesis *segments* includes the input signal and the results of segmenting this signal according to the segmentation parameters (instant time, frame length and overlap) defined for each line of reasoning. The hypothesis

spectra contains the magnitude and phase of the spectrum of the segments defined in the *segments* hypothesis according to the *fft* size and window type parameters. The level of information corresponding to *detection functions* incorporates the detection functions corresponding to each of the onset reduction methods implemented. And, finally, the hypothesis *peaks* stores the detected peaks obtained from the detection functions by using a specific threshold and peak-picking algorithm.

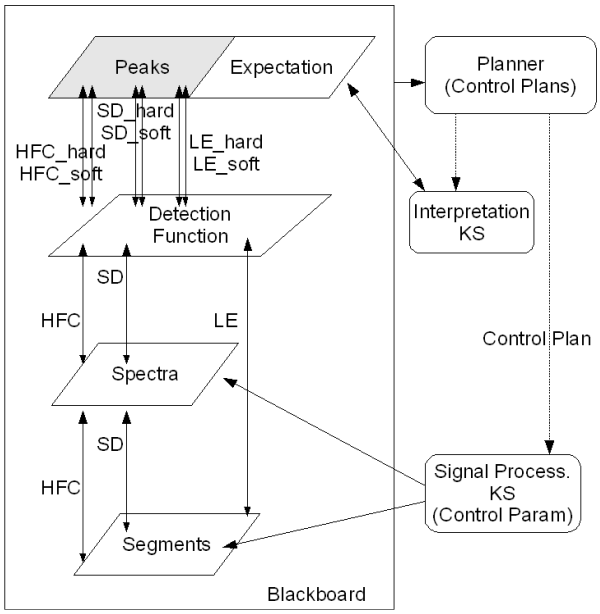

Figure 3 – Redundant onset detection: hypotheses and lines of reasoning.

The onset detection blackboard includes the following signal processing knowledge sources: A *segmentation* knowledge source segments the input signal according to the starting instant time, frame length and overlap. A *phase vocoder* knowledge source calculates the *fft* and instant frequencies. The configuration parameters of this knowledge source are the *fft* size and window type applied to each segment of signal. There are also knowledge sources for the reduction functions. As defined in [1], reduction refers to the process of transforming the audio signal into a detection function that shows the presence of transients in the signal. The following reduction functions have been implemented: High Frequency Content analysis (HFC), Spectral Difference (SD) and Local Energy (LE) calculation. These methods are described in [1]: HFC calculates the spectral energy weighting emphasizing high frequencies because transients are usually clearer in high frequency; SD calculates the energy of the spectral difference of consecutive frames; LE is just an estimation of the energy in time. Last, a peak-picking function identifies the local maxima of the detection function above a defined threshold. This peak picking function adaptively calculates the mean value of the detection function within a time interval. A peak is detected if a local maximum satisfied the following condition:

$$\frac{d(n)}{mean\{d(n-M_l),...,d(n+M_u)\}} \geq \lambda \ . \tag{1}$$

Where $d(n)$ is the detection function, λ is the threshold and M_l and M_u are the lower and upper limits for the mean calculation.

Figure 3 also shows the different lines and sub-lines of reasoning. Three lines of reasoning are defined for the High Frequency Content analysis (*HFC*), the Spectral Differences (*SD*) and the Local Energy (*LE*) calculation. These lines of analysis are a set of knowledge source instances to be executed by the

control system. The parameters of the knowledge sources are adapted by the control system according to the processing plan corresponding to that line of analysis. The reduction algorithms are tuned independently so as each line of reasoning uses different segmentation and analysis parameters. For example, the *HFC* line of reasoning executes the *segmentation* KS with a frame length of 1024 and an overlap of 512 samples. Then, the control system activates the *phase vocoder* KS to calculate the spectrum using an *fft* size of 4096 and a hanning window. The HFC KS reduces the *spectra* hypothesis to calculate a HFC detection function.

As can be seen in the *peaks* hypothesis of figure 3, each of these three lines of analysis creates two sub-lines of reasoning. These sub-lines correspond to the application of a hard and soft threshold to the *peak picking* knowledge source.

Finally, an *integration* knowledge source integrates the peaks detected by the three lines of reasoning and implements a voting mechanism to rate the detected onsets. Hence, if a peak meets the hard threshold the peak is labelled as an onset. On the other hand, if a peak meets the soft threshold by at least two of the lines of analysis the peak is also labelled as an onset.

Figure 4 shows a simple test signal with three onsets located in 0.5, 1 and 1.5 seconds. The test signal starts with 0.5 seconds of normally distributed random samples of power 1. Then, the power of these random samples is doubled. Next, a low frequency tone of 1 *KHz* starts at $t = 1$. And finally, the last part of the signal has a tone of 1 *KHz* and a high frequency tone of 20 *KHz*.

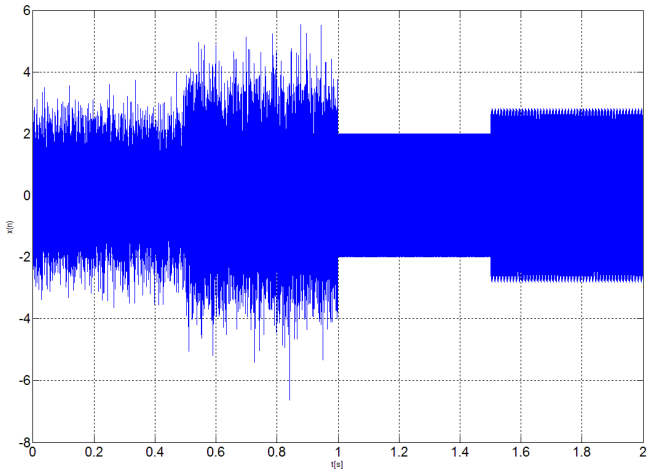

Figure 4 – Onset detection example.

Table 1 summarizes the results of the voting mechanism described above. The first of the onsets is just a change in the energy and the three lines of analysis detect this onset as a soft onset, therefore the onset is correctly detected. The second onset is a transient to a low frequency tone that maintains the energy of previous segments. As we could expect neither the *Local Energy* line of analysis nor the *High Frequency Content* analysis detect the onset. The *Spectral Difference* line of analysis detects this onset because the spectral characteristics change drastically. Finally, the last of the onsets is detected by the LOR that use spectral features because there are both high and low frequency tones. The *Local energy* line does not detect this last onset because the energy does not change from previous segments.

Table 1 – Onset detection results.

	Onset 0.5s	Onset 1s	Onset 1.5s
High Freq. Content	*soft*	*none*	*hard*
Spectral Differences	*soft*	*hard*	*hard*
Local Energy	*soft*	*none*	*none*

The *spectral differences* reduction function detects the three onsets but the number of false positives is high. False positives are reduced by using soft thresholds and combining the information of the three lines of analysis. The detection functions of the three lines of analysis are presented in Appendix A.

4 Conclusions

A blackboard-agent framework, an approach taken from expert systems, has been proposed as a global sound source separation system that allows the integration of multiple and redundant audio analysis methods. The system simplifies the definition of a control mechanism by dividing a complex system in several subsystems. The blackboard architecture defines multiple lines of reasoning. This allows the integration of redundant analysis methods. Finally, an example of a redundant onset detection approach has been presented.

Acknowledgements

This work has been partially financed by the Spanish MEC, ref. TEC2006-13883-C04-02, under the project AnClaS3 “Sound source separation for acoustic measurements”.

References

[1] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, “A tutorial on onset detection in music signals,” *IEEE Transactions on Speech and Audio Processing*, September 2005.

[2] A. Bregman, “Psychological data and computational ASA,” in *Computational auditory scene analysis*, D. Rosenthal and H. Okuno, Eds. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, USA, 1998.

[3] Chris Duxbury, Mark Sandler, and Mike Davies, “A hybrid approach to musical note onset detection,” in *Proc. of the 5th Int. Conference on Digital Audio Effects (DAFx-02)*, Hamburg, Germany, September 26-28, 2002.

[4] Ruohua Zhou and Josh Reiss, “Music onset detection combining energy-based and pitch-based approaches,” in *Third Music Information Retrieval Evaluation eXchange (MIREX)*, Vienna, Austria, September 2007.

[5] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, “An experimental comparison of audio tempo induction algorithms,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 5, pp. 1832–1844, Sept. 2006.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley-Interscience Publication, 2000.

[7] V. Lesser, R. Fennell, L. Erman, and D. Reddy, "Organization of the hearsay II speech understanding system," *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, vol. 23, no. 1, pp. 11–24, Feb 1975.

[8] Daniel P.W. Ellis, Prediction-driven computational auditory scene analysis, Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, 1996.

[9] Juan Pablo Bello, Towards the Automated Analysis of simple polyphonic music: A knowledge-based approach, Ph.D. thesis, King's College London – Queen Mary, University of London, 2003.

[10] Norman Carver and Victor Lesser, "The evolution of blackboard control architectures," Tech. Rep., Amherst, MA, USA, 1992.

[11] Frank Irwin Klassner, Data reprocessing in signal understanding systems, Ph.D. thesis, University of Massachusetts Amherst, 1996.

[12] Yoichi Muraoka Masataka Goto, "Real-time beat tracking for drumless audio signals: chord change detection for musical decisions," *speech Communications*, , no. 27, pp. 311–335, 1999.

[13] A. Pena, N. Degara-Quintela, M. Sobreira-Seoane, and S. Torres-Guijarro, "Anclas3: A blackboard-based cooperative framework for sound separation," in *124th AES Convention*, Amsterdam, The Netherlands, May 2008.

Appendix A

The reduction function obtained for the different lines of reasoning are included bellow. Soft and hard peaks are presented as a red stem plot. As it can be seen in Figure 6 (right), the number of false positives of the soft reduction function is high. This is corrected by introducing a voting scheme.

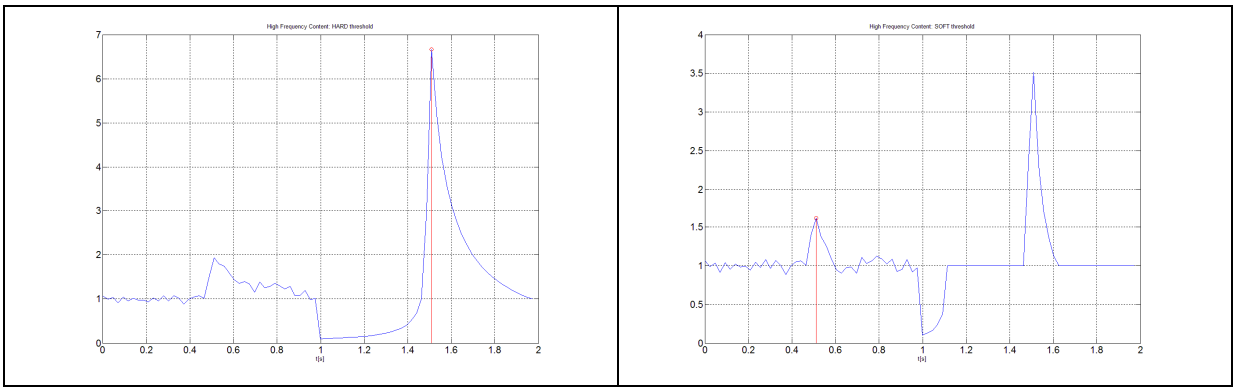

Figure 5 – *High Frequency Content* reduction functions: *hard* (left) and *soft* (right) peaks.

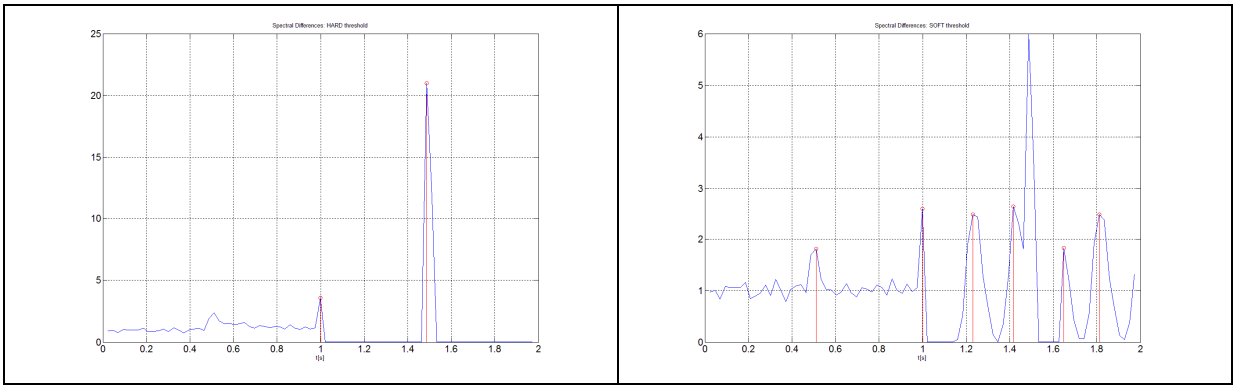

Figure 6 – *Spectral Differences* reduction functions: *hard* (left) and *soft* (right) peaks.

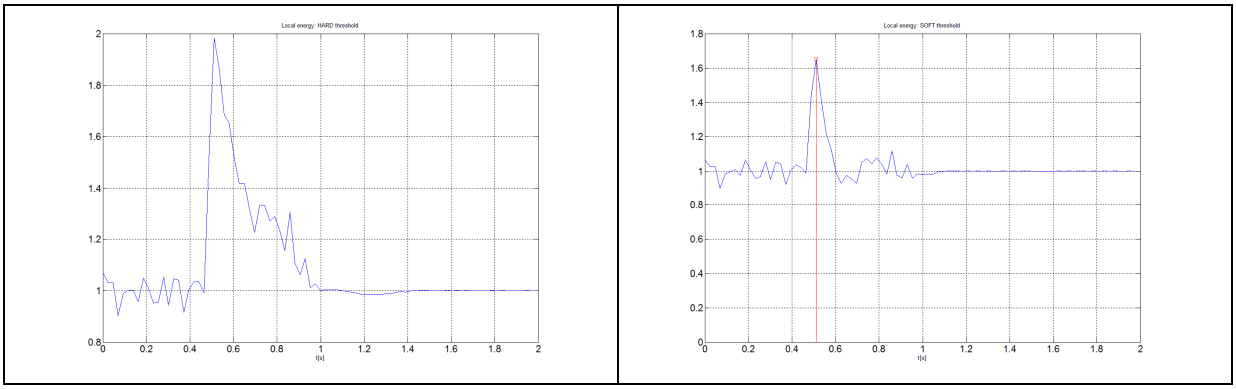

Figure 7 – *Local Energy* reduction functions: *hard* (left) and *soft* (right) peaks.