Detection and modeling of fast attack transients

Xavier Rodet and Florent Jaillet[1]

Ircam, 1 place I. Stravinsky, 75004, Paris
email: rod@ircam.fr, florent.jaillet@genesis.ac

Abstract

Attack transients or simply attacks, are zones of short duration and fast variation of the sound signal short-time spectrum such as at the attack of percussion instruments. There are many motivations for attack detection and modeling; improving general analysis techniques is one of them. A method and a program have been developed for detection, modeling and reconstruction of attacks. It is based on the detection of energy peaks appearing simultaneously in several frequency bands of a time-frequency representation. This program has been tuned and tested on a data base of percussive sounds and mixtures.

1. Introduction

The term *attack transient*, or simply *attack* does not have a precise definition. It corresponds to the beginning of notes produced by an instrument. Fast *attacks* are zones of short duration (a few ms) and fast variation of the signal short-time spectrum with an abrupt increase in energy particularly noticeable in high frequencies since energy is usually concentrated in the low ones. There are many motivations for attack detection and modeling. One for instance is the improvement of general analysis methods. For instance, *classical* sinusoidal additive analysis-synthesis [Se89] based on peaks in the Short Time Fourier Transform usually does not preserve the sharpness of attacks. This is due to the use of a finite length window in the spectral estimation. Moreover, a pre-echo may appear right before the attack of the synthetic sound: when a window extends over an attack, sinusoids are detected at a time where they are not yet present in the signal. This can be partly alleviated by some more refined techniques such as *reallocation* [Fitz00]. But detection of attacks can still improve general analysis methods. For example it may allow a precise positioning of analysis windows with respect to attacks or avoid them to overlap on attacks. In other musical applications, detected and modeled attacks can be used for further musical processing. Attack times can, for instance, trigger a synthesis algorithm so that the percussive rhythm found in a recording is used in a new musical phrase.

2. Review of detection methods

[Se89] proposes to preserve separately the original attacks and to substitute them for the corresponding parts in the resynthesized sound. [Mas96] present three detection methods but no model. [Lev98] detects the abrupt variation of the energy of the signal. These two authors also propose the use of wavelets for detection. In [Kro87] only the principle of wavelets based detection is described. This has been implemented by [Daud99] but no numeric evaluation is given. [Gri99] proposes High Resolution Matching Pursuit adapted to attack detection and representation. However computational load is high. No modeling of the attacks is usually proposed but in [Tho00] (detection by use of the Prony method) and in [Ver97] (sinusoidal modeling applied to the Cosine Transform).

3. Detection of attacks based on a time-frequency representation

The attack detection and modeling method developed in this research is also based on the following requirements:

- It should not use additive analysis results, in order to be usable for other purposes (segmentation, instrument recognition, etc.).
- It should succeed in every type of sound (particularly polyphonic sounds) and with good time accuracy.
- It should be simple to use: analysis parameters should as much as possible be adjusted automatically.
- It should be tested on a data base of sounds including polyphonic mixtures of percussive and non-percussive sounds.

A good way to design such an algorithm is to start with some time-frequency or time-scale representation. In this

[1] Now with GENESIS, Batiment Beltram, Domaine du Petit Arbois BP69, 13545 Aix en Provence Cedex 04, FRANCE

research, the Short-Time Fourier Transform (STFT) has been chosen, in particular because of the small computational load of allowed by FFT. Another motivation was to facilitate research and shorten development duration, therefore allowing more time for tuning and testing. However, other representations, such as a Wavelet Transform (implemented in a fast algorithm) could have been used as well [Kro87]. It even seems that a Wavelet Transform would lead to better results because of better time resolution at high frequency and better stability at low frequency. Let us call $|X(k, f)|$ the magnitude of the STFT at sample k and frequency f. It is computed from the sound signal on a window of size N and with a step size S.

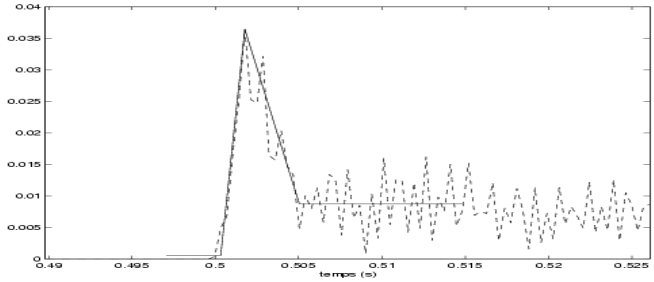

Fig. 1. $|X_f(k)|$ showing an energy peak in observation window W_m in frequency band f.

3.1 Construction of the observation function

For the goal of detection, the definition of an *attack* adopted in this work is *an area of short duration of the STFT in which marked energy peaks appear in several frequency bands*. Examining the energy in one frequency band, i.e. fixing f to some value leads to a monodimensional signal $|X_f(k)|$ in which short duration peaks are looked for. The signal $|X_f(k)|$ is studied in observation windows W_m of length K at locations m. A peak is supposed to occur in an observation window when $|X_f(k)|$ shows a *triangular* shape with a high maximum above prior and post *plateaus* (Fig. 1). Typical value for K is 18 ms. Therefore, in window W_m, the next step is to approximate $|X_f(k)|$ by such a triangular function and to measure its height. To keep computational load low, we avoid classical optimal estimation. Instead, the maximum of a possible peak is said to be the maximum value M of $|X_f(k)|$ in W_m. and the edges of the triangle are easily estimated linearly by using mean square error minimisation.

Calling M_b (respectively M_a) the mean of $|X_f(k)|$ in the window W_m *before* (respectively *after*) the triangle, an indicator function is computed as (except for some special cases): $I_{f,m} = \big((M - M_b) + (M - M_a)\big) / (M_b + M_a)$

$I_{f,m}$ (Fig. 2) takes large values when there is a large peak in the window W_m. For sake of simplicity, the center of gravity of the triangle is chosen as the precise instant of the attack. It would be interesting to use better estimates of the precise perceptual time of the attack as studied in various psychoacoustic works such as [Gor87]. But the main goal of this research is detection and modeling of transients, for which the center of gravity of the triangle is precise enough since its duration is only a few ms.

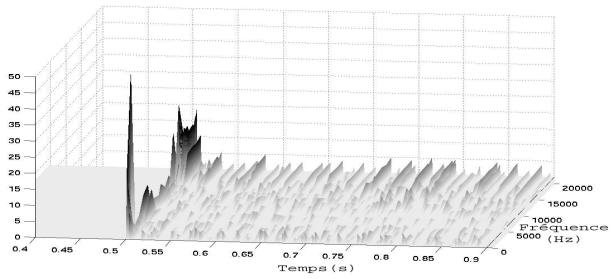

Fig. 2. Observation function $I_{f,m}$ of peaks in the STFT $|X_f(k)|$ for a note of a percussive instrument.

3.2 Selection of aggregates and final decision

A threshold T_o is applied to $I_{f,m}$ leading to a thresholded observation $J_{f,m}=I_{f,m}$ if $I_{f,m} > T_o$, 0 otherwise (Fig. 3). Non-zero values of $J_{f,m}$ indicate peaks in the STFT.

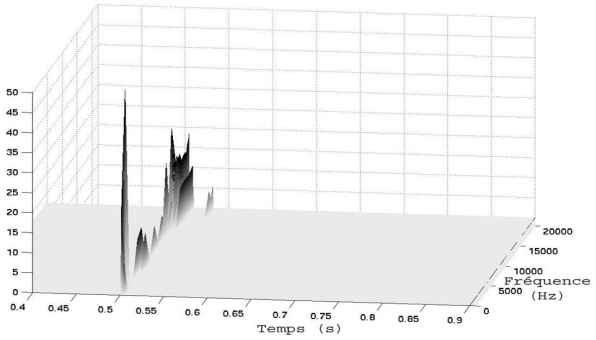

Fig. 3. Thresholded observation function $J_{f,m}$ of peaks in the STFT $|X_f(k)|$ for a note of a percussive instrument.

Then the areas of the STFT in which several peaks appear at close temporal positions are aggregated as one attack. Let (f,m) be a point of the STFT where $J_{f,m}$ takes a non-zero value. Since the location of such a peak is not precisely known, we define an uncertainty interval $I(f,m) = [k\text{-}p(f,m), k+p(f,m)]$ where $p(f,m) = p_{max}(1 - \exp(\beta J_{f,m}))$ where p_{max} and β are parameters. Then two points (f_1,m_1) and (f_2,m_2) are aggregated if their intervals overlap or if they are aggregated with a common other point (transitivity):

$\mathrm{Agr}((f_1,m_1),(f_2,m_2))$ if: $I(f_1,m_1) \cap I(f_2,m_2) \neq \varnothing$ or

$\exists\,(f_3,m_3)$ s.t. $\mathrm{Agr}((f_1,m_1),(f_3,m_3))$ and $\mathrm{Agr}((f_2,m_2),(f_3,m_3))$

An aggregate A_l is any set of points which are aggregated according to the previous definition.

$A_l = \{ (f,m)$ s.t. for each $(f_1,m_1) \in A_l$, $\mathrm{Agr}((f,m),(f_1,m_1))\}$.

The weight of an aggregate is the sum of the values $J_{f,m}$ over the aggregate. Only the aggregates the weight of which is higher than a given threshold T_a, are preserved and considered to be detected attacks.

3.3 Data base and choice of parameter values

To tune parameter values and to test the detection algorithm, a data base of 75 recordings of various types has been built and each of the 390 attacks has been hand marked with a time tag independant of frequency. There are 17 recordings where 305 attacks are mixed with sustained sounds and 58 recordings where attacks are isolated. This data base is not large enough for statistically significant results, but a larger data base would require much time for hand marking of attacks. Optimal parameter values have been found rather dependent on the type of sounds (polyphonic or not, clear/soft attacks...). The tests however permitted to determine ranges for the parameters allowing good results.

3.4 Weighting according to Frequency

In order to evaluate the reliability of various frequency channels, the positions of the non-zero values of the thresholded observation function $J_{f,m}$ are compared to attack marks placed by hand. For each channel center frequency *f*, a non-zero value occurring within less than 10 ms of an attack mark is considered as a good detection. Otherwise, it is considered as a false alarm. The number of good detections and false alarms is calculated for each frequency and for various threshold values. As an example, figure 4 displays the number of good detections and false alarms, versus the various channel center frequencies from 0 to 22 kHz. It appears that channels in the frequency bands from 9 to 20 kHz give a more reliable information $J_{f,m}$ than others

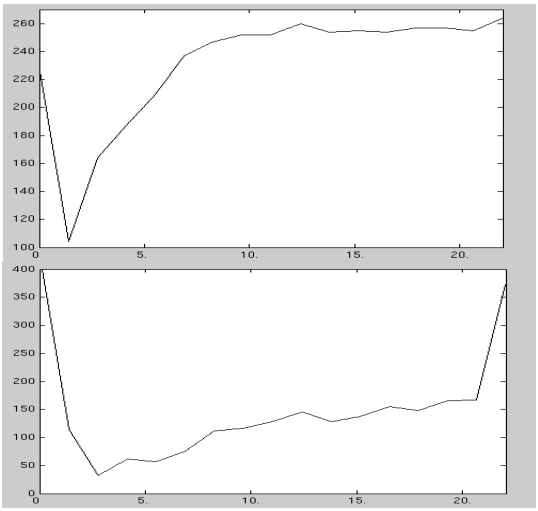

Fig. 4. Number of good detection (top) and false alarm (bottom) versus channel center frequency (kHz).

and that the lowest frequency bands cause a great number of false detection. It would be interesting to study the reliability of frequency bands according to the analysis window size N, and according to the low frequency stationary content, but this has not been done yet. As a result of the reliability measurement, the various frequency bands have been weighted according to their reliability. An evaluation of the detection algorithm has been done on the 17 recordings (305 attacks) where attacks are mixed with sustained sounds. The other recordings have been disregarded since it was relatively easy to choose parameters so as to have no errors in an isolated context. On the contrary, detection of attacks in a multiphonic context

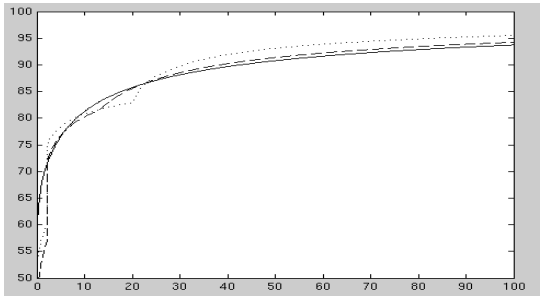

Fig. 5 Percentage of good detection versus percentage of false alarm for various threshold values T_2 and for p_{max} = 5 (-), 10 (--) and 15 (...).

can be extremely difficult. Note also that the evaluation was done before frequency weights were introduced. Figure 5 displays the percentage of good detection versus the percentage of false alarm for various threshold values T_2 from 5 to 130 and for p_{ma} = 5, 10 and 15. The figure 5 shows that, for instance, 80% good detection can be obtained at the cost of some 10% false alarm. Considering the fact that these recordings contain a variety of mutiphonic contexts and that, therefore, some attacks are covered by sustained parts of the sound, the results shown in figure 5 are promissing. Further more, it is the guaranteed minimum performance since the only non-fixed parameter is then the proportion of good detection versus false alarms. On the contrary, adjusting the different parameters according to the type of sound and recording context leads to better results. Also, for additive synthesis applications, only the most prominent attacks need to be detected. Finally, there is no doubt that the algorithm can be significantly improved on several points. For instance, it is necessary to better understand when and why false alarms occur. Among possible improvements, lets us quote the frequency weights and the use of multi-resolution analysis which is expected to improve the reliability of the algorithm in low frequency bands particularly.

3.5 Time-frequency representation and reconstruction of attack transients

For each detected attack, the values of the STFT X in the aggregate A_l, i.e. a subset of the STFT, are considered as the *time-frequency representation* of the attack. The *complex* STFT is used here in order to exactly reconstruct the attack signal $a_l(n)$. The reconstructed attack signal can then be subtracted from the original signal to remove the attack in a recording, or it can be added to the additive synthetic signal to improve the sharpness of attacks. This time-frequency representation is a an ensemble of STFT values which are null everywhere, except in the aggregate where it is equal to the original STFT. In order to compute an optimally reconstructed signal $a_l(n)$, it is thus necessary to use a reconstruction algorithm, such as the method of [Gri99] which is applied here.

3.6 Adjustment of the size of reconstructed attacks

When reconstruction is done from the aggregates formed during detection (*detection aggregates*), reconstructed attacks are of very short duration. Effectively, to avoid spurious detection, the observation threshold T_o is rather high. Therefore, the *reconstruction aggregate* is defined with a reconstruction threshold $T_r < T_o$. Adjustment of T_r allows user control of the size of the reconstructed attack. A supplementary improvement could use a modeling technique better adapted to modeling some of the short time resonance which follows the attack. For instance, the Resonance Modeling analysis technique [Pot86] could be used to better extract the resonance modes following the attack itself.

3.7 Implementation and graphical user interface

A detection, modeling and reconstruction program, named *TransAn* has been implemented. Its GUI facilitates usage according to user needs. It allows visualization of STFTs (sonagram), observation functions, aggregates, detected attacks and original and reconstructed sound signals. It also allows the user to adjust parameter values according to sound or visual results. Detected attack instants and reconstructed attacks are stored in an SDIF file using the *marks* type and the *time domain samples* or the *STFT* type (http://www.ircam.fr/anasyn/sdif/standard/types-main.html).
This program has been applied to some of the sounds of the Sound Analysis and Synthesis Panel at ICMC2000, significantely improving additive resynthesis. Other examples will be demonstrated at the conference. For instance, a performance of Indian Sarod (strings) and Tabla (percussion) has been analyzed. Tabla attacks have all been correctly detected, modeled and resynthesized.

4. Perspectives

The detection, modeling and reconstruction program *TransAn* appears very useful for various musical applications. Used as a complement to sinusoidal additive analysis, it improves the sharpnes of attacks. It can also be used to detect and extract attack transients which can be used in further musical processing, for instance as the excitation of resonating filters [Pot86]. A large data base of sounds and a systematic statistical study of transients would provide *a priori information* (probability distributions) of attacks and permit to optimize parameter values, to improve the shape of the approximation function and to optimize weightings according to energy and frequency bands.

5. References

[Bas86] Basseville M. and A. Benveniste. Detection of abrupt changes in signals and dynamical systems. Springer, Berlin, 1986.

[Goo97] Goodwin M. Matching Pursuit with damped sinusoids, Proc. Int. Conf. on ASSP 1997. 37, 41, 42, 160.

[Tho00] Thornburg H., F. Gouyon. A flexible analysis/synthesis method for transients, ICMC 2000, Berlin Aug. 2000, pp. 400-403.

[Fitz00] Fitz K. Transient preservation under transformaion in an additive sound model, Proc. Int. Computer Music Conf. 2000, Berlin, Augt. 2000, pp. 392-395.

[Gor87] Gordon J. The perceptual attack time of musical tones, J. Acoust. Soc. am. 82(1), July 1987, pp. 88-106.

[Gri99] Gribonval R., Ph. Depalle, X. Rodet, E. Bacry and S. Mallat. Sound signal decomposition using a high resolution matching pursuit, Proc. Int. Computer Music Conf. (ICMC'96), Hong-Kong, Aug. 1996.

[Kro87] Kronland-Martinet R., J. Morlet, and A. Grossmann, Analysis of sound patterns through wavelet transforms. in International Journal of Pattern Recognition and Artificial Intelligence, 1987.

[Lev98] Levine S. N. PhD thesis, Stanford University, December 1998.

[Mas96] Masri P. PhD thesis, University of Bristol, December 1996.

[Se89] Serra X. A system for sound analysis, transformation and synthesis based on a deterministic plus stochastic decomposition, PhD thesis, Stanford University, Oct. 1989.

[Daud99] L. Daudet. PhD thesis, Université de Marseille, December 2000.

[Grif84] Griffin D. and J. Lim. Signal estimation from modified short time Fourier transform, IEEE Trans. Speech & Sign. Processing v. 32, 1984, pp 236-243.

[Pot86] Potard Y. et al. Experimenting with models of resonance, ICMC 1986, The Hague, Oct. 1986, pp. 269-274.

[Ver97] Verma T.S., S.N. Levine, T.N.Y. Meng, Transient modeling synthesis, ICMC 1997, Thessaloniki, Grece, Sept. 1997, pp. 164-167.