Detection and modeling of fast attack transients

Abstract

Introduction

The term *attack transient* does not have a precise definition. It corresponds to the beginning of notes produced by an instrument. *Attacks*, as they are called here, are zones of short duration (about a few ms) and fast variation of the sound signal with an abrupt increase in short-time energy distributed on the whole spectrum and noticeable in the high frequencies since energy is usually concentrated in the low ones.

There are many motivations for attack detetcion and modeling. Improving general analysis techniques is one more. Classical additive analysis, for instance, of a sound does not preserve the spectral richness of attacks in the resynthesized sound. Moreover, a visible pre-echo appears right before the attacks of the resynthesized sound since, when a window extends on an attack, sinusoids are detected at a time where they are not yet present in the signal. Therefore, detection of attacks is also necessary to ensure a synchronization of analysis windows in order to forbid them to overlap on attacks.

Review of detection methods

[Ser89] 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 [Gou] (by use of the Prony method) and in [Verma97] (sinusoidal modeling applied to the Cosine Transform).

Some objectives

The attack detection and modeling method developed in this research:

- Should not use additive analysis results, in order to be usable for other purposes (segmentation, instrument recognition, etc.).
- Should succeed in every type of sound (particularly polyphonic sounds) with a good time accuracy.
- Should be simple to use: analysis parameters as much adjusted as possible automatically.

Detection of attacks based on a time-frequency representation

According to the previous objectives, the Short-Time Fourier Transform (STFT) has been chosen, in particular because of the small load of calculation allowed by FFT. The Fast Wavelet Transform could have been used as well. 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.

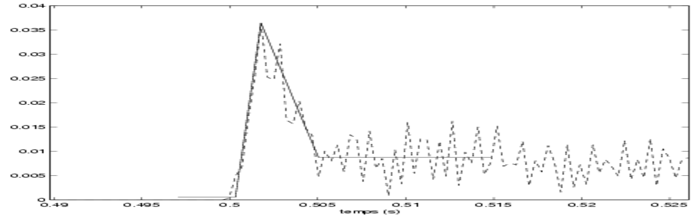

Fig. 1. $|X_f(k)|$ showing an energy peak in observation window W_m in frequency band f

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 signal $|X_f(k)|$ in which short duration peaks are looked for. The signal $|X_f(k)|$ is then 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).

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 calculation 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 edges of the triangle are easily estimated.

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):

$$I_{f,m} = \frac{(M - M_b) + (M - M_a)}{M_b + M_a}$$

$I_{f,m}$ (Fig. 2) takes large values when there is a large peak in the window W_m. The center of gravity of the triangle is chosen as the precise instant of the attack.

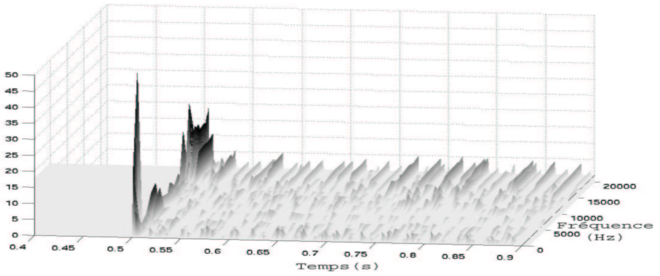

Fig. 2. Observation function $I_{f,m}$ of peaks in the STFT $|X_f\,(k)|$ for a note of a percussive instrument

Selection of aggregates and final decision

A threshold T_d is then applied to $I_{f,m}$ leading to a thresholded observation function $J_{f,m}$ (Fig. 3).

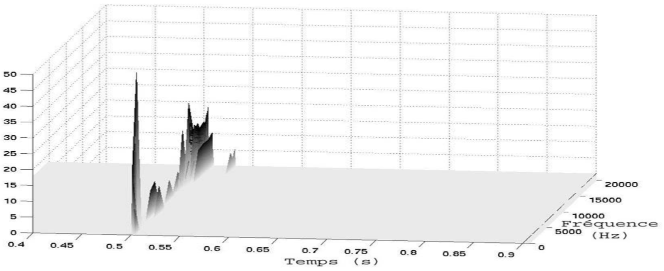

Fig. 3. Thresholded observation function $J_{f,m}$ of peaks in the STFT $|X_f\,(k)|$ for a note of a percussive instrument

Non-zero values of $J_{f,m}$ indicate peaks in the STFT. Then the areas of the STFT in which several peaks appear at close temporal and frequential positions are aggregated as one attack.

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

Data base and choice of parameter values

A data base of 70 recordings of various types has been constituted to test the detection algorithm. It is not large enough for statistically significant results, but a larger data base requires much time for hand labeling 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.

Weighting according to Frequency

The positions of the non-zero values of $J_{f,m}$ are compared to attacks marks placed by hand. For a given frequency, 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 rate of good detection and false alarm are calculated for each frequency and for various threshold values. It appears that medium frequency bands give a more reliable information than others and the lowest frequency bands cause a great number of false detection. Weights according to frequency are adjusted accordingly.

Time-frequency representation and reconstruction of attack transients

For each detected attack, the aggregate, i.e. a subset of the STFT, is the *time-frequency representation* of the attack. The *complex* STFT is used here in order to exactly reconstruct the attack signal. For example, the reconstructed attack signal can then be subtracted from the original signal to remove the attack in a recording. This time-frequency representation is a STFT which is null everywhere, except in the aggregate where it is equal to the original STFT. Consequently, the method of [Griffin and Lim] is applied to compute an optimal reconstructed signal.

Adjustment of the size of reconstructed attacks

When reconstruction is done from the aggregates formed during detection, reconstructed attacks are of very short duration. Effectively, to avoid spurious detection, the detection threshold T_d is rather high. Therefore, the reconstruction aggregate is defined with a reconstruction threshold $T_r < T_d$. Adjustment of T_r allows user control of the size of the reconstructed attack. A supplementary improvement uses the Resonance Modeling analysis technique [Potard86] to better estimate resonance modes which constitute the attack signal.

Implementation and graphical use interface

A detection, modeling and reconstruction program 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 attacks' instants and reconstructed attacks are stored in an SDIF file using the *marks* type and the *time domain samples* or the *STFT* type. This program has been applied to some of the sounds of the Sound Analysis and Synthesis Panel at ICMC2000, largely improving resynthesis. Other examples will be proposed at the conference. For instance, a performance of Indian Sarod (strings) and Tabla (percussion) has been analyzed. Tabla attacks where all correctly detected and modeled to isolate the tabla part.

Perspectives

A large data base of sounds and a systematic study of transients would provide *a priori information* (probability distributions) of attacks and permit to improve parameter values, to improve the shape of the approximation function and to optimize weightings according to energy and frequency bands.

References

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

[Goo97] Michael Goodwin Proc. of the Int. Conf. on ASSP 1997.

[Gou84] P. Goupillaud et al. Geoexploration, 1984.

[Gou00] Fabien Gouyon et al. ICMC 2000

[Gri99] Rémi Gribonval PhD thesis, Université Paris IX, September 1999.

[Gro] A. Grossmann, M. Holschneider, R. Kronland-Martinet, and J. Morlet. Detection of abrupt changes in sound signal with the help of wavelet transforms.

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

[Lev98] Scott Nathan Levine. PhD thesis, Stanford University, December 1998.

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

[Ser89] Xavier Serra. PhD thesis, CCRMA October 1989.

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

[Griffin and Lim] IEEE Trans. Speech & Sign. Processing v. 32, 1984, pp 236-243.

[Potard86] Y. Potard et al. ICMC 1986

[Verma97] ICMC 1997