Improved Modelling of Attack Transients in Music Analysis-Resynthesis

Paul Masri, Andrew Bateman
Digital Music Research Group, University of Bristol
5.11 Merchant Venturers Building, Woodland Road, Bristol BS8 1UB, U.K.
Tel: +44 117 954-5203, Fax: +44 117 925-5265, email: Paul.Masri@bristol.ac.uk

Abstract

Current music analysis-resynthesis models represent sounds through a set of features, which are extracted from a time-frequency representation. So that each time-frame can present a good approximation to the instantaneous spectrum, it is necessary to analyse the waveform in short segments. This is achieved with a window function whose position is advanced by a fixed amount between frames. When the window encompasses a transient event, such as the percussive onset of a note, it contains information both before and after the event. These partially-correlated spectra often become confused during analysis and cause audible 'diffusion' upon resynthesis.

This paper presents a simple, novel technique to avoid the problem, by synchronising the analysis window to transient events. Event locations are identified by observing short-term changes in the spectrum. Thereafter the position of the analysis window is constrained, to prevent it capturing the signal both sides of an event simultaneously. This method, which has been automated, yields an improvement that is clearly audible, particularly for percussive sounds which retain their 'crispness'.

1. Introduction

It has long been known that the onset of a note, the *attack*, plays an important role in our perception of timbre [pp.9-12, Grey, 1975]. In traditional instruments, it is the phase during which resonances are building up, but before the steady state condition of standing waves has been established. Where the attack is short, such as for the trumpet, there are many rapid changes, so that it can sound like a noise burst. For this reason, the attack transient is difficult to study.

It is not surprising therefore, that attacks are not well understood and are not well represented within analysis-resynthesis models. This paper focuses on finding a solution in the context of the popular Deterministic Plus Stochastic model. This model was developed by Xavier Serra[1990], based upon the Sinusoidal model of McAulay and Quatieri[1986]. The presented work was implemented within the authors' own system, from which other model developments have also been forthcoming [Masri & Bateman, 1994, 1995].

1.1 Traditional Spectral Analysis and The Problem Caused by Attack Transients

The first step in the analysis process is the time-frequency representation, which is calculated using the Short Time Fourier Transform (STFT). For each frame, a small portion of the time domain waveform is isolated, by application of an analysis window, and spectral estimation is computed using the Fast Fourier Transform (FFT). Between frames, the analysis window is advanced by a fixed amount, called the *hop-distance*.

During the deterministic analysis, the primary goal is to detect and locate the *partials*, the instantaneously sinusoidal elements that compose the harmonic structure of a pitched sound. For good frequency resolution, it is necessary to have a long analysis window. In opposition to this, for good time resolution, the window must be short. The practical constraint of separating partials in the frequency domain necessarily favours good frequency resolution.

The central assumption when using the FFT is that of stationarity - the waveform is assumed to be truly periodic within the analysis window. Small deviations from this generate a small, but tolerable, amount of distortion.

When the analysis window contains a percussive note onset, there is a dramatic change in the waveform. The introduction of the new sound element is unconnected to the preceding sound signal and cannot in any way be predicted from it. Therefore the waveform preceding the note onset is largely uncorrelated with the waveform following its initiation.[1]

This far-from-stationary situation causes much distortion in the FFT spectrum and often affects the subsequent feature extraction processes adversely. This is demonstrated in Figure 1 where the frames prior to an attack transient are similar; also the frames following its onset are similar to each other; the central frame spanning both regions, though, is a 'confused' average of both spectra.

Within the deterministic analysis, the partials are identified from peaks in each spectrum and continuous trajectories are formed by linking peaks between frames. A frame which contains a percussive attack or other transient event often contains a large burst of energy throughout the spectrum. In such a case, the peak detection algorithm may fail to detect peaks, resulting in a momentary drop-out upon synthesis. Alternatively, the rapid changes may introduce multiple spurious

[1]The sound preceding the note onset may continue, so the spectra may not be *totally* uncorrelated.

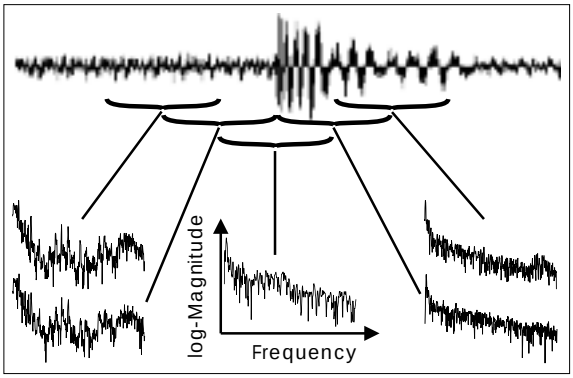

Figure 1 - Spectra surrounding an attack

peaks. Some of these will be discarded at the peak linking stage and some will be erroneously linked, leading to artifacts upon synthesis.

In some cases, where the transient event does not dominate the signal, peak linking may be largely successful. However, upon synthesis, the partial trajectories are smoothly interpolated between frames and the transient events, which were highly localised in time, become diffused or even dissolved completely. The diffusion effect is reinforced by the stochastic aspect of the model: during analysis each spectral envelope represents an averaged snapshot of the spectrum; upon synthesis, there is smooth cross-fading between frames. See Figure 2.

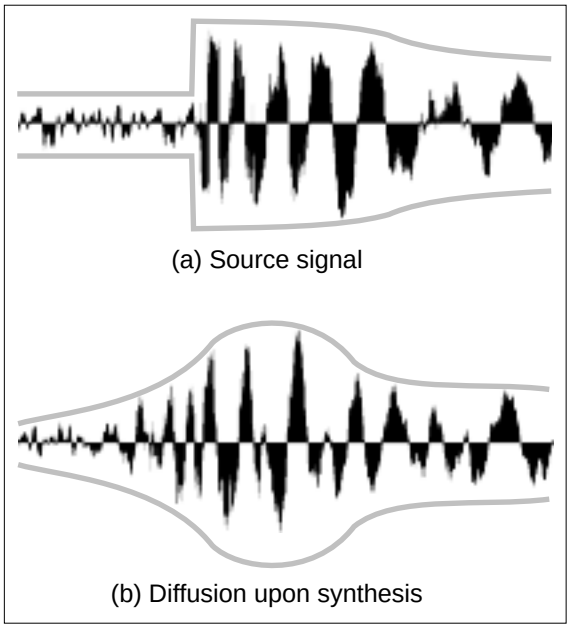

Figure 2 - Smoothed attack causing 'diffusion'

The solution presented in this paper detects transient events and avoids the above problems, by synchronising the analysis and synthesis processes to those event locations.

2. Detection of Transient Events

The detection method was designed to recognise two signal properties associated with a sharp attack: the suddenness of the signal change and the increase in energy. A frequency domain method was chosen because of its ability to reveal both changes in overall energy and the energy concentration in frequency. The frequency location of energy is important because the sudden change to the signal will cause phase discontinuities; in the frequency spectrum this appears as high frequency energy.

Naturally the time resolution for detecting transient events must be smaller than that of the main analysis STFT, if any advantage is to be gained. This necessitates a reduction in frequency resolution, but fine frequency resolution is not an issue here; only the broad spectral envelope is required. The following parameters were found to be useful:-

Window-length =2.9ms (128 samples @ 44.1kHz),
Hop-distance =1.5ms (64 samples @ 44.1kHz),
Hamming window function,
No zero padding.

The hop-distance is set to half the window-length, the maximum value that ensures each transient event will appear toward the centre of the window in at least one frame.

2.1 Detection Function

The energy function is calculated as the sum of the magnitude squared of each frequency bin (in the specified range):

$$E = \sum_{k=2}^{N/2+1} \left\{ \left| X(k) \right|^2 \right\} \qquad (1)$$

where E is the energy function for the current frame,
N is the FFT array length
(so $N/2 + 1$ corresponds to the frequency $F_s/2$,
F_s is the sample rate),
$X(k)$ is the kth bin of the FFT.

The function to measure high frequency content was arbitrarily set to a weighted energy function, linearly biased toward the higher frequencies:

$$HFC = \sum_{k=2}^{N/2+1} \left\{ \left| X(k) \right|^2 \cdot k \right\} \qquad (2)$$

where HFC is the High Frequency Content function for the current frame,
other symbols as defined above.

In both cases the lowest two bins are discarded, to avoid unwanted bias from DC or low frequency components.

The condition for detection combines the results from each pair of consecutive frames thus:

$$\frac{HFC_r}{HFC_{r-1}} \cdot \frac{HFC_r}{E_r} > T_D \qquad (3)$$

where subscript r denotes current frame (equals latter of two in detection function),
subscript r-1 denotes the previous frame,
T_D is the threshold, above which a hit is detected.

(Note that HFC_{r-1} and E_r are constrained to have a minimum value of one, to avoid the potential 'Divide by zero' computation error.)

The detection function is the product of the rise in high frequency energy between the two frames and the normalised high frequency content for the current frame.

For attacks whose growth is slightly slower, but whose onset is nevertheless sudden, the detection function could be triggered on more than one frame. To avoid multiple detections, the algorithm is given a parameter for the minimum closeness of two hits. In practice, setting this to 2 frames is adequate for the majority of sounds (i.e. only disallowing consecutive hits).

2.2 Temporal Resolution and Region Definition

The accuracy, in time, of the detection process is equal to the hop-distance. The above values of STFT parameters give a resolution of 1.5ms, which compares favourably to the accepted resolution of the ear, which is 2-10ms.

The *transient event boundary* - the point of onset of the transient - is stored as the start of the second analysis window of the detection pair. It is the second window that contains the event, so this placement ensures that the whole of the attack is located after that point. In this way, any errors in the deterministic analysis, caused by inaccuracy within this process, are confined to the frame containing the percussive onset, where the suddenness of the attack should dominate perceptually.

The detection process is carried out as a pre-analysis scan, and its results are compiled as a *region list*, where the region boundaries correspond to the detected transient event boundaries, as shown in Figure 3.

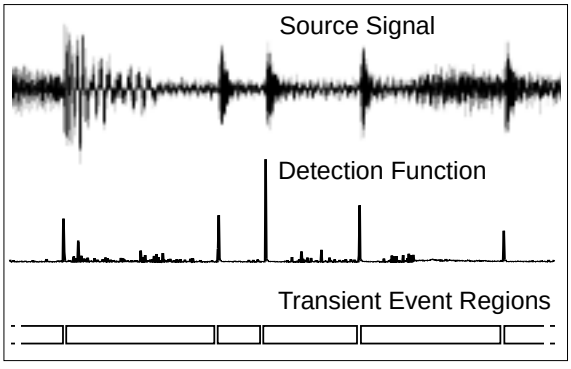

Figure 3 - Event detection and region generation

3. Synchronised Analysis

At the start of each region, the analysis window is positioned with its trailing end at the first sample. Thereafter analysis proceeds, as normal, the window advancing by the hop-distance for each new frame.

The first frame whose window touches or crosses the region boundary is 'snapped' so that its leading edge coincides with the last sample of the region. Naturally, this means that the final hop-distance is reduced.

3.1 Extrapolation Towards Region Boundaries

The data for each frame notionally corresponds to the instantaneous situation at the centre of the frame. Therefore the first half of the first frame and the last half of the last frame in each region are undefined. For simplicity, the data in these frames are extrapolated outward to the respective region boundaries.

This makes the implicit assumption that the waveform is slow changing in these zones. Whereas this may be accurate at the end of a region, we already know that there are rapid changes at the start of a region. Despite this, the STFT provides no way of gaining extra detail.

4. Synchronised Synthesis

Upon synthesis, both deterministic and stochastic, the extrapolated spectra are synthesised beyond the region ends by a short amount. This provides sufficient excess waveform to allow a crossfade between regions. See Figure 4 below.

The crossfade length has been set to 5.8ms (256 samples @ 44.1kHz sample rate), where it is short enough to reproduce the suddenness of the original attack, but not so short that an audible click (termed a 'glitch') is produced.

5. Results

5.1 Performance

Figure 4 shows that the proposed method is able to retain the suddenness of an attack transient. The success of the method is further confirmed through

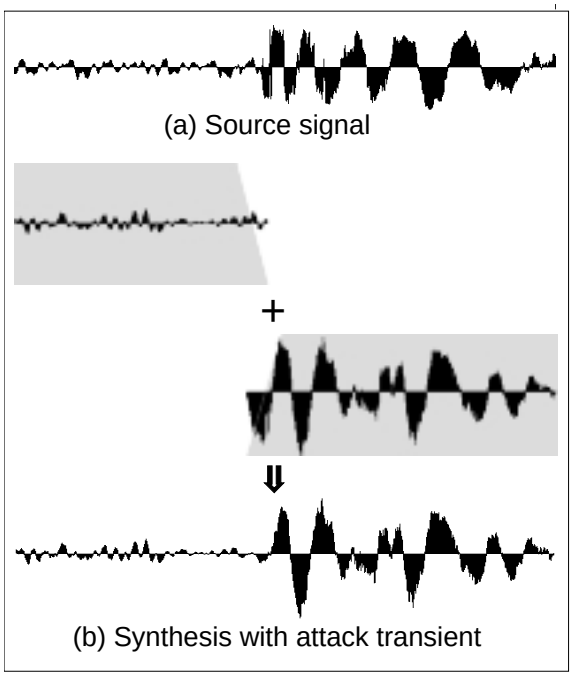

Figure 4 - Crossfade at a region boundary

listening tests. Sounds synthesised by the new method retain the crispness of the original sound.

The performance is good even where the spectral evolution has been simplified. (Some spectral detail is inevitably lost following a transient event boundary, as an inevitable consequence of using the STFT for analysis.) It would appear that the abrupt rise in volume and the sudden change in spectral content are the most prevalent perceptual features.

Some degradation is noticeable for sounds with greater 'depth', such as a booming bass drum or a snare with reverberation, where some of that power is lost. This is probably due to the rapidity of change following the transient onset, the necessary detail of which, the analysis is still unable to extract.

One area where the modification is unable to help is the rapid succession of transient events, such as from a drum roll, or drums with gated reverb. In these cases, the pre-analysis scan is often still able to detect the event boundaries, but the events are too close to one another to apply the synchronisation. That is, the resultant regions are shorter than a single window-length for the main analysis FFT's.

5.2 Cost

The computational cost of the pre-analysis scan is low, when compared to the analysis itself. The FFT's are much shorter - and thereby disproportionately faster[2] - and the hop-distance of half a frame is a reduction in the overlap of analysis windows between frames.

In addition to the pre-analysis scan, some changes have been made to the model structure, but these have an impact only once per transient event, and the impact is minimal.

6. Conclusions

The popular Deterministic Plus Stochastic model makes the assumption that all waveform changes are gradual, an unavoidable consequence of the STFT's limited time-frequency resolution. Consequently, the model fails to capture transient events adequately and audible diffusion or complete drop-outs result.

In this paper a method has been presented that extends the model to incorporate percussive attacks and other transient events. The principle behind the method is that the spectra before and after an attack transient should be treated as different, the change happening instantaneously at the transient onset.

The results have proven successful in preserving the abrupt amplitude change in the waveform and the crispness, perceptually, of attack transients. This is however the first implementation of this technique, and improvements are possible. Two areas of future work are proposed.

The time domain envelope at the start of each region could be captured and imposed on the synthesised output, to improve the 'sense of depth' for certain sounds.

Analysis of closely-spaced attacks may require a change to the time-frequency representation. Higher order spectra (e.g. Wigner-Ville, bispectrum, etc.) do not suffer the same time-frequency resolution restrictions of the linear transforms (e.g. Fourier, Wavelet) [Boashash, 1990][Cohen, 1989]. However they possess their own quirks, and a detailed investigation is needed before they can be applied to music analysis-resynthesis. Further discussion on this subject is soon to be published in [Masri, *to be published*].

Acknowledgements

The authors wish to express their gratitude to **Soundscape Digital Technology** for funding this research and providing essential equipment. Thanks are also due to the **Centre for Communications Research** (University of Bristol) and its director, Prof. Joe McGeehan, for supporting the research with generous access to facilities and the expertise of its staff.

References

[Boashash, 1990] B. Boashash. "Time frequency Signal Analysis" (ch.9) in *Advances in Spectral Analysis.* Ed. S.Haykin. 1990. Vol.1; pp.418-517. Publ. Prentice Hall (Englewood Cliffs, NJ, USA).

[Cohen, 1989] L. Cohen. "Time-frequency Distributions - A Review" in *Proceedings of the IEEE.* 1989. Vol.77:7, pp.941-981.

[Grey, 1975] J.M. Grey. *An exploration of musical timbre using computer-based techniques for analysis, synthesis and perceptual scaling.* Ph.D. dissertation, 1975. Stanford University.

[Masri & Bateman, 1994] P. Masri, A. Bateman. "Partial Domain Synthesis of Music" in *Digital Signal Processing Conference Proceedings (DSP UK 94).* 1994. Vol.1.

[Masri & Bateman, 1995] P. Masri, A. Bateman. "Identification of Nonstationary Audio Signals Using the FFT, with Application to Analysis-based Synthesis of Sound" in *IEE Audio Engineering Colloquium Digest.* 1995. pp.11/1-11/6.

[Masri, *to be published*] P. Masri. *Computer modelling of Sound for Transformation and Synthesis of Musical Signals.* Ph.D. dissertation, due for submission in Summer 1996. University of Bristol.

[McAulay & Quatieri, 1986] R.J. McAulay, T.F. Quatieri. "Speech Analysis/Synthesis based on a Sinusoidal Representation" in *IEEE Transactions on Acoustics, Speech and Signal Processing.* 1986. Vol.34:4, pp.744-754.

[Serra,1990] X. Serra. *A system for sound analysis/ transformation/synthesis based on a deterministic plus stochastic decomposition.* Ph.D. dissertation, 1990. Stanford University.

[2] The processing required for an FFT is related to N, the length of the FFT array, by the factor NlogN. Thus a reduction in array length yields more than a proportional advantage in computation speed.