SEPARATION OF TRANSIENT INFORMATION IN MUSICAL AUDIO USING MULTIRESOLUTION ANALYSIS TECHNIQUES

Chris Duxbury, Mike Davies, Mark Sandler

Department of Electronic Engineering, Queen Mary, University of London,
Mile End Road, London E1 4NS, UK
christopher.duxbury@elec.qmul.ac.uk

ABSTRACT

Whilst musical transients are generally acknowledged as holding much of the perceptual information within musical tones, most research in sound analysis and synthesis tends to focus on the steady state components of signals. A method is presented which separates the noisy transient information from the slowly time varying steady state components of musical audio. Improvements of using adaptive thresholding, and multiresolution analysis methods are then illustrated. It is shown that by analyzing the resulting transient information only, current onset detection algorithms can be improved considerably, especially for those instruments with noisy attack information, such as plucked or struck strings. The idea is then applied to audio processing techniques to enhance or decrease the strength of note attack information. Finally, the transient extraction algorithm (TSS) is applied to time-scaling implementation, where the transient and noise information is analyzed so that only steady state regions are stretched, yielding considerably improved results.

1. BACKGROUND

Much of the research on sound models focuses on the steady state portion of sound, looking at spectral information. Despite this, many texts on psychoacoustics insist that attack characteristics play a large part in human recognition of musical instruments. Transient information in music can also be a good indication of note onset time, implying that good transient recognition is essential for audio to music transcription software. Hence, this work is motivated by a need to obtain a signal model where the transient and steady state (used here to refer to tonal components which are locally steady state) parts are separated, allowing individual analysis of each.

Generally, transient information is considered as the residual once the steady state part of the signal has been analyzed, resynthesized and subtracted [1], [2]. The problem in this approach is that higher-level information about sinusoidal tracks or fundamental plus partials detection is generally needed to yield good results. The aim here is not to develop a sound model which yields improved results in synthesis, but one which performs a good separation of slowly time varying, and quickly time varying parts of the signal for further sound analysis and exploration.

The intention is to look at frequency bins on a frame-by-frame basis, and to decide whether or not they can be considered steady state or transient/noise information. Each bin is selected as belonging to either group. For this purpose, a 'steady-state measure' is needed for each bin. A solution to this was presented in [3], with a good explanation also provided in [4]. It uses phase information between frequency bins in adjacent time frames to facilitate transient/steady-state (TSS) separation. The idea of phase increment is taken from phase vocoder theory, where it is used to calculate the instantaneous frequency.

If we consider a windowed STFT of the signal *x(m):*

$$X(n,k) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j2\pi mk/N} \qquad (1)$$

where n and k represent the hop number and frequency bins respectively. In polar form, this can be expressed as:

$$X(n,k) = \left|X(n,k)\right| \cdot e^{j\phi(n,k)} \qquad (2)$$

For frequency bin k, a 'target phase', ϕ_t, can be calculated using the bin frequency, and unwrapped bin phase of the previous hop:

$$\phi_t(n,k) = \phi(n-1,k) + \omega_k h \qquad (3)$$

where ω_k is the frequency of bin k, and h is the hop size. This target phase represents the perfect case of a steady state sinusoid fitting exactly into a frequency bin. Since this is almost never the case, we have some phase deviation, ϕ_d, given by:

$$\phi_d(n,k) = \phi(n,k) - \phi_t(n,k) \qquad (4)$$

The principle argument (corresponding value in the range $-\pi$:π) of this deviation phase is then used to calculate the unwrapped phase increment per hop, $\Delta\phi_h$:

$$\Delta\phi_h(n,k) = \omega_k h + princ\left[\phi_d(n,k)\right] \qquad (5)$$

The instantaneous frequency, f_i, can then be calculated:

$$f_i(n,k) = \frac{\Delta\phi_h(n,k)}{2\pi h} \cdot f_s \qquad (6)$$

where f_s is the sample frequency. In the case of a single sinusoid with constant frequency, it is expected that the value of instantaneous frequency should be equal between adjacent frames. Similarly, for noisy components, the value for instantaneous frequency, and therefore phase increment, would be expected to vary. Hence, by considering differences in phase increment between adjacent windowed frames, the noisy component of the signal, containing the transient information can be isolated. For the steady state case, we expect:

$$f_i(n,k) \cong f_i(n-1,k) \tag{7}$$

and hence:

$$\Delta\phi_h(n,k) - \Delta\phi_h(n-1,k) \cong 0 \tag{8}$$

for steady state bins. Hence, using (3),(4) and (5), we obtain:

$$\phi(n,k) - 2\phi(n-1,k) + \phi(n-2,k) < T_{ss} \tag{9}$$

with T_{ss} being the *steady state detection threshold*. Similarly, we can use:

$$\phi(n,k) - 2\phi(n-1,k) + \phi(n-2,k) > T_t \tag{10}$$

where T_t is the *transient detection threshold*.

2. MULTIRESOLUTION ANALYSIS

Although the main idea here was computationally efficient and straightforward to implement, problems arise in the quality of the results. Firstly, using a fixed window size throughout leads to either poor time resolution, and undesirable pre-echo effects, or good time resolution, but poor frequency resolution, making accurate thresholding problematical. Also, by using the same threshold across the full frequency spectrum, there is a tendency for too much low frequency component to be selected as steady state and all high frequency component of the signal to be selected as transient (plus noise) information.

A solution to these problems with the previous implementations is to use a frequency selective threshold and/or a frequency dependant window size. However, as this signal model is only intended as a front end to musical analysis programs, it is important that computation is kept minimal. For this reason, the signal is divided up into octave spaced subbands for individual processing, as shown in figure 1. This was implemented using a constant-Q filter bank, to obtain six bandlimited signals from 0-700Hz through to 11-22kHz. The frequency bins of each bandlimited STFT can then be selected as in the previous algorithm to facilitate TSS separation. This is similar in many ways to wavelet based analysis techniques, [5].

The filterbank is implemented using a bank of perfect reconstruction filters [6], which is essential if the TSS representations are to be faithful components of the original.

Whilst the subband filterbank approach overcomes the problems associated with frequency dependency in the transient component selection, a second problem arises in the case of sinusoids which have certain frames lying outside the steady state threshold, especially when in the presence of noisy components. This considerably reduces the sound quality of the resulting signals, giving the sound a synthetic, distorted quality. From this, it was decided that an adaptive TSS selection threshold would be preferable.

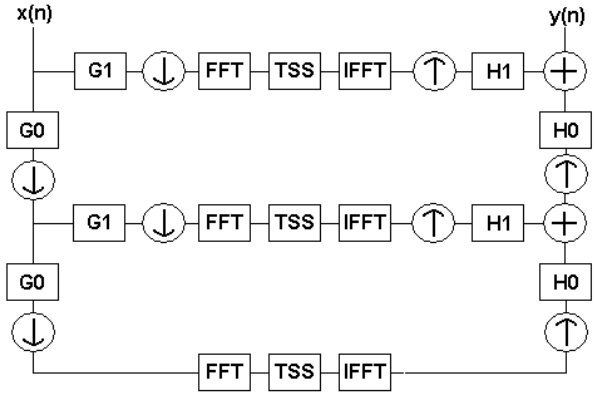

Figure 1. *Octave spaced filter bank implementation. Only two subband levels are shown here, whereas six bands are used in the actual implementation.*

3. ADAPTIVE THRESHOLDING

It is desirable to take into account previous information of the steady state/transient nature of frequency bins when carrying out TSS separation on the current frame. If a bin is slightly outside the steady state threshold, but was within it for these last few frames, it is likely that this frame is still the same sinusoid, which must be continued in order to preserve the quality of the results.

The adaptive threshold algorithm was developed based on this idea, which calculates an individual threshold for each frequency bin based on the nature of the same frequency bin in the previous frame, yielding improved results. If T_{ss} is the fixed threshold set by the user, with *a* and *b* being real numbers, the adaptive threshold, A_{ss}, algorithm is:

$$A_{ss}(n,k) = T_{ss} + \alpha(n,k)T_{ss} + \beta(n,k)T_{ss} \tag{11}$$

where

$$\alpha(n,k) = \begin{cases} 0 & \Delta\phi(n-1,k) - \Delta\phi(n-2,k) < A_{ss}(n-1,k) \\ a & \Delta\phi(n-1,k) - \Delta\phi(n-2,k) > A_{ss}(n-1,k) \end{cases} \tag{12}$$

and

$$\beta(n,k) = \begin{cases} 0 & \Delta\phi(n-2,k) - \Delta\phi(n-3,k) < A_{ss}(n-2,k) \\ b & \Delta\phi(n-2,k) - \Delta\phi(n-3,k) > A_{ss}(n-2,k) \end{cases} \tag{13}$$

Here, α weights by how much the threshold is increased based on previous frame information. This idea is then extended further such that the threshold is increased further if both the previous frames were steady state for that frequency bin. This is the reason for the second weighting function, β, giving weight to the amount both previous frames being steady state in the corresponding bin affects the threshold. The main reason for using weighting functions is to tie the increase in threshold to the original fixed threshold. Hence, the user need only set the value of T_{ss} to obtain

their desired TSS separation. Values for *a* and *b* of 3 and 4 respectively produce good results, regardless of the value of T_{ss}.
Another key motivation for this adaptive algorithm was to keep computation at a minimum. Clearly an intelligent algorithm using all previous information may yield better results, but this would increase computation considerably. In order to allow this calculation without the phase information of the previous three frames being stored, the above can be simplified as follows.
Searching for the case where:

$$\Delta\theta(n-1,k)-\Delta\theta(n-2,k)>A_{ss}(n-1,k) \tag{14}$$

is the equivalent of saying:

$$F_{ss}(k,n-1)>0 \tag{15}$$

where $F_{ss}(k,n)$ is the resultant vector of frequency bins kept from the frame n. This is one place where computation can be minimized.
Secondly, the case of:

$$\Delta\theta(n-2,k)-\Delta\theta(n-3,k)>A_{ss}(n-2,k) \tag{16}$$

can also be calculated using:

$$A_{ss}(n-1,k)\geq T_{ss}+\alpha\cdot T_{ss} \tag{17}$$

since the previous frame must have had an increased threshold at that bin if the frame before that had been selected as steady state in that bin. Hence the function to calculate the adaptive threshold does not require vectors of previous phase information to be stored. It need only be passed the previous adaptive threshold vector and frequency results vector, process it, and overwrite it with the new values.
By sufficiently weighting the previous functions, sinusoids are considerably less likely to be broken, whilst there is little effect on noisy components, particularly at transients.
Note that in order to ensure that that steady state component of the signal contains all the information not in the transient part of the signal, the transient detection thresholds must be set similarly. This then allows a single threshold to be used to separate the signal into transient/steady state components with good results, as in *Figure2*. In the original algorithm, the transient threshold, T_t, had to be set much lower than the steady state threshold, T_{ss}, which left a lot of 'in between' signal information.

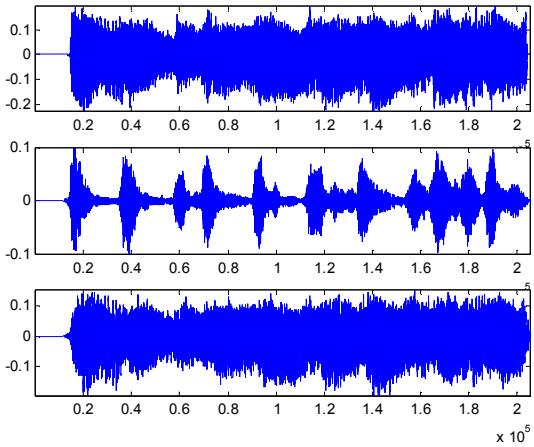

Figure 2. *Separation of a strummed guitar signal (top) into transient/noise (center) and steady state (bottom) parts.*

4. APPLICATIONS

Once the TSS separation has been performed, the two signals obtained contain very different information. The steady state signal can be viewed as carrying the pitch information whereas, for many instruments, the transient signal carries onset information.
A first application therefore is improved onset detection. Current onset detection algorithms tend to use signal energy measures, usually given frequency weighting. An example of this, using high frequency content (HFC) was proposed by [7], where HFC is given as:

$$HFC=\sum_{k=2}^{N/2+1}\left|(X_k)\right|^2\cdot k \tag{18}$$

The onsets are then detected for a given frame, r , using:

$$\frac{HFC_r}{HFC_{r-1}}\cdot\frac{HFC_r}{E_r}>T_D \tag{19}$$

where E_r is the L_2 norm squared energy of the frame. This method proves successful in certain cases, however, for close onsets, especially when a short window is used, it remains difficult to detect onsets.
In using only the transient output from the TSS separation, there is much less data to analyze for the HFC, leading to clearer detection, and results are improved considerably.
High precision onset detection is achieved by doing a short window frame-based energy analysis. Frequency content weighting can be included, but is not necessary in many cases.
This signal can be low pass filtered to yield a smoother signal for detection. By taking the points of maximum gradient from this, we can locate transient regions accurately, usually representing note attacks (see *Figure 3*).

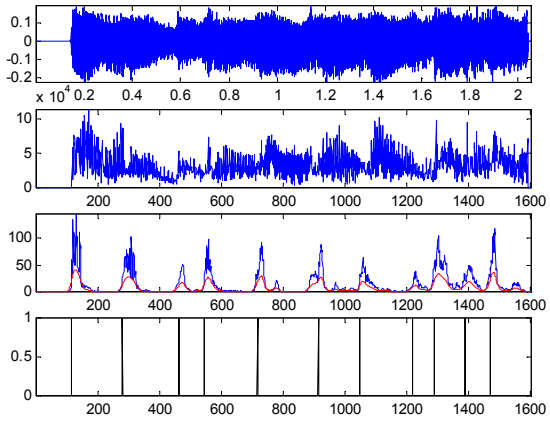

Figure 3. *Comparison between standard HFC content (upper middle) and TSS separated transient only HFC content (lower middle) of strummed guitar signal (top) yielding near-perfect onset detection (bottom).*

A second possible application for this work is in the field of audio effects. Isolating transient components can be used to post process musical signals by enhancing, or decaying attack information or "punchiness". For example, a pre-recorded guitar

track can be post-processed to accentuate guitar attacks, producing a hard "plectrum plucked" sound. If the transient information is kept the same, but the steady state is processed, a guitar can be made to sound like the strings are being muted. The processing involved in these effects requires more than just varying the amplitude of transient or steady state components. If this is the case, too many spectral subtraction artifacts tend to be introduced when either component is increased or decreased by a considerable amount. Much better quality and more perceptually realistic results are obtained by filtering the transient or steady state components individually, as in *Figure 4*.

By high pass filtering the transient components of many instruments, it takes away the heavy attacks, producing the effect of the piece being played more softly. For the opposite effect such as harder playing, more low-pass filtered transient component can be added to the signal.

In the case where the steady state component is high pass filtered, it gives the effect of low resonance, such as muted instruments. Similarly, by adding more low pass steady state component, an instrument can be made to sound more resonant. These basic processing ideas produce perceptually convincing results for much musical audio, and can be combined, leading to powerful post-processing tools. For example, in the case of polyphonic pop music, the percussion can be processed with almost no effect on the quality of the remaining instruments.

Since all this has been set out with the aim of being minimally computational, these processing ideas have a fast computation time. It is our intention to develop a real time application based on this to allow user-friendly post processing of audio in this way. Further experimentation with filters is also intended as future work.

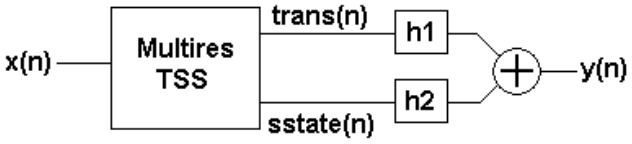

Figure 4. *Percussive post-processing algorithm. Once the signal has been split using the multiresolution TSS separation algorithm, resulting signals are filtered separately to produce hard/soft percussive attacks.*

.

A third application is in time-scaling of audio. When time-scaling using phase vocoder-like spectral methods, considerable improvement can also be found when stretching signals only in regions where there is minimum transient information. An implementation based on this idea using the TSS signal separation presented here has yielded very high quality results, particularly for signals such as pop music, or plucked/struck strings.

5. CONCLUSION

Improvements have been made to a transient/steady state algorithm outlined in [4]. This algorithm used phase vocoder theory to analyze which frequency bins represent steady state information in a frame-by-frame analysis. By exporting this idea to a mutiresolution framework the frequency dependency problems of the previous algorithm were greatly reduced. A second improvement has been shown which calculates a frequency dependant adaptive steady state threshold, based on previous frames. This transient/steady state (TSS) separation model has been applied to yield improvements in several musical applications. High precision onset detection has been illustrated. An audio effects application has also been given which allows post processing of audio to edit the percussive parts of a signal to control hard/soft playing expression parameters. Finally, spectral methods of time-scaling musical audio have shown considerable improvements when using the transient signal output of the TSS separation algorithm as information for which regions should be stretched.

Other filter bank configurations may yield some improved results. For example, an ERB or bark band filter bank may be used in place of the constant Q filter bank, which makes sense from a perceptual point of view in synthesis applications. Some investigation should also be made in the post-processing filters to see what effects can be achieved. It is also intended that the low computational complexity of all the processing involved in both the TSS separation and the post processing application be exploited to yield real-time desktop applications.

6. REFERENCES

[1] Serra, X., *Musical Signal Processing Ch. 3,* Swets and Zeitlinger Publishers, 1997.

[2] Daudet, L., "Transients modelling by pruned wavelet trees", Proc. International Computer Music Conference (ICMC2001), 2001.

[3] Settel, J. and Lippe, C., "Real-time musical applications using the FFT-based resynthesis", Proc. International Computer Music Conference (ICMC94), 1994.

[4] Arfib, D., Keiler, F., Zoelzer, U., *Time-frequency Processing,* to be published in *DAFX: Digital Audio Effects*, 2001.

[5] Mallat, S., *A Wavelet Tour of Signal Processing*, *2nd Edition,* Academic Press, 1998.

[6] Haddad, A., *Multiresolution Signal Decomposition,* Academic Press, 1992.

[7] Masri, P., Bateman, A., "Improved Modelling of Attack Transients in Music Analysis-Resynthesis", Proc. International Computer Music Conference (ICMC96), 1996