Transient detection and preservation in the phase vocoder

Axel Röbel
IRCAM, Analysis-Synthesis Team, France
email: Axel.Roebel@ircam.fr

Abstract

In this paper we propose a new method to reduce phase vocoder artifacts during attack transients. In contrast to existing algorithms the new approach does not enforce the time dilation parameter to one during transient segments and works at the level of spectral peaks such that stationary parts of the signal will not be affected by detected transients. For transient detection we propose a new algorithm that is especially adapted for phase vocoder applications because its detection criterion may be interpreted in terms of transient position and and allows us to chose the optimal location for phase reinitialization. The evaluation of the transient detection shows superior performance compared to a previously published algorithm. Attack transients in sound signals transformed with the new algorithm provide very high quality even if strong dilation is applied to polyphonic signals.

1 Introduction

The phase vocoder (Serra 1997) is widely used for signal transformation. Due to recent advances (Dolson and Laroche 1999) it can be considered a very efficient tool for signal transformation that achieves high quality transformed signals for weakly non stationary signals. Abrupt changes in the amplitude of a signal, however, will usually lead to considerable artifacts and remain a challenge for phase vocoder applications.

The problem has been studied recently (Bonada 2000; Duxbury, Davies, and Sandler 2002) and it has been shown that significant improvements concerning the sound characteristics of transients can be achieved by means of detecting transients, reinitializing the phase for the transient regions and forcing the time stretching factor to be one during the transient such that the phase relations remain unaltered. The transient detection is usually based on energy change criteria in rather broad bands and the phase is reinitialized for all bins in the frequency band detected as transient. For polyphonic signals this will almost certainly destroy the phase coherence of stationary partials passing through the same frequency region. Fixing the delay factor to one in the transient regions requires automatic compensation in non transient regions to achieve the requested stretch factor. For a dense sequence of transients this may be difficult to achieve.

The algorithm proposed in the following article addresses all these issues. The transient detection mechanism classifies transients at the level of spectral peaks allowing us to improve the frequency resolution of the transient processor. We show that it is not necessary to force the stretch factor to one during transient regions if the phase initialization is done when the transient is close to the window center. Despite of this simplification of the algorithm it achieves sufficient phase synchronization to reproduce the transients with subjectively high quality.

In section **2** of this article we will investigate into the problem of processing attack transients with the phase vocoder. Based on the theoretical understanding of the spectral characteristics of transient partials we propose a conceptually simple yet effective transient processing scheme. In section **3** a transient detection algorithm is developed that is especially adapted for the application in the phase vocoder and the performance of the algorithm is evaluated using a small data base of hand labeled sounds. In section **4** we describe the results that have been achieved when processing transient sounds with the new algorithm.

2 Transient processing

The theoretical foundation of signal transformation by means of modifying the short time Fourier transform (STFT) of the signal has been established in (Griffin and Lim 1984). For changing the time evolution of a signal in the STFT domain one assumes that every frame contains a nearly stationary signal in which case the time evolution can be changed by simply repositioning the frames in time. To achieve coherent overlap of adjacent frames during resynthesis the phase of each bin of the discrete Fourier spectra has to be corrected based on an estimation of the frequency of the related partial.

The phase correction can be derived for properly resolved and nearly stationary partials (Serra 1997; Dolson and Laroche 1999). If the amplitude of a signal partial changes abruptly, a situation normally denoted as attack transient, the prerequisites of the phase correction are no longer valid and consequently the results will have bad quality.

The main problem of the phase vocoder when processing attack transients is the fact that the transient signal does not have a predictable relation to the previous frames such that a reinitialization of the phase spectrum is inevitable for proper resynthesis of the transient. After reinitialization of the phase spectrum two further problems require investigation. To understand the impact of these effects we have investigated into a simple attack model that is a linear ramp with saturation. First, as shown later in fig. 1 the phase spectrum of a sinusoid with attack transient is not con-

stant but will have a negative slope that reduces toward zero if the window is moved over the transient. Concerning this problem we found that the evolution of the phase spectrum is nearly linear for each bin of the DFT and, therefore, it does not present any problems for the phase vocoder as long as the time increment between the frames involved, i.e. the frame positions before and after time stretching/compression, is kept smaller than an 8th part of the window size. Second, the amplitude spectrum of a transient sinusoid, i.e. the peak bandwidth, amplitude and side lobe positions, changes with time. We found that the modifications of the amplitude spectrum to be applied to compensate a change of the transient position within the window are quite involved and difficult to model. However, the error that is made due to keeping the amplitude spectrum fix remains small for small frame offsets and if the transient is sufficiently covered by the analysis window.

Concerning the optimal position of the transient for phase initialization there exist two arguments that both require the phase reinitialization to take place if the transient is close to the window center. First, the transient is reproduced without any error only in the frame that gets the phases of the transient bins reinitialized and, therefore, the reinitialization should take place when the impact of the reconstructed transient on the output signal is largest. Second, the reinitialization of the phases will produce the transient at the very same position where it was located in the analysis window. To avoid the need to reposition the transient to properly fit into the transformed time evolution the phase reinitialization should take place when the transient is close to the center of the window.

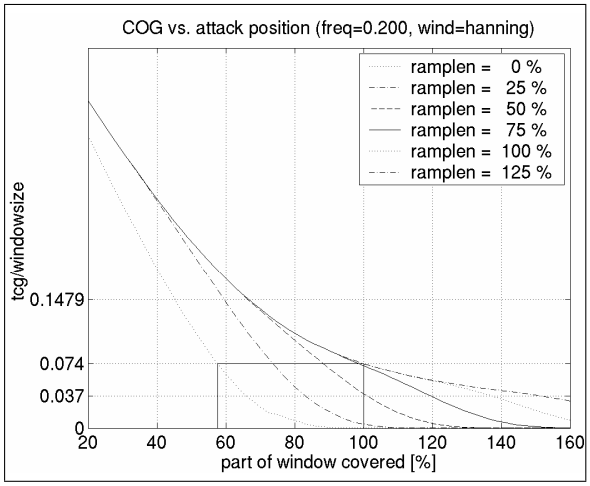

Figure 1: Center of gravity of partial energy as a function of transient position under the analysis (hanning) window for transient partials with fixed frequency and varying length of linear ramp (in percent of window size). The selected threshold C_e (see text) is marked.

The proposed transient preservation is based on the hypothesis that the phase vocoder algorithm will properly handle transient events if the phase of the transient parts of the spectrum are reinitialized when the center of the transient is close to the center of the window. To estimate the transient position we propose to use the center of gravity (COG) of the signal energy related to isolated spectral peaks. The COG of the instantaneous energy of the windowed signal $s_h(t, t_m)$, with the analysis window h centered at time t_m is defined as

$$\bar{t} = \frac{\int t(s_h(t,t_m))^2 dt}{\int (s_h(t,t_m))^2 dt}.$$

As shown in (Cohen 1995) it can be calculated by means of

$$\bar{t} = \frac{\int -\frac{\partial \phi(w,t_m)}{\partial w} A(w,t_m)^2 dw}{\int A(w,t_m)^2 dw}, \qquad (1)$$

where $A(w,.)$ and $\phi(w,.)$ are representing the amplitude and phase spectrum of the windowed signal. The negative phase derivative, called group delay, signifies the contribution of a frequency to this position. While these equations are derived for time continuous signals the same type of relations can be established for the DFT of discrete time signals. Note that the group delay may be obtained by means of the time reassignment operator which can be calculated efficiently by means of a Fourier transform of the signal using a modified analysis window (Auger and Flandrin 1995). To be able to estimate the COG of sinusoidal components in a composite spectrum we calculate the COG of the peak main lobe, only. By virtue of the amplitude weighting taking place in eq. (1) the error made due to neglecting part of the signal spectrum is small. For sinusoids that are too close in frequency to be individually resolved the treatment of individual peaks performs a somewhat arbitrary signal decomposition which nevertheless will correctly detect transients as long as all the sinusoids that are contributing to the same peak are transient.

In fig. 1 the decrease of the COG (the decay of the phase slope) is shown that results if the analysis window moves over attack transients with different ramp length. The ramp length is given in percent of the analysis window length and the window position is given in terms of the part of the window (in percent of the window length) overlapping with the attack transient. By means of comparing the COG evolution for different transient forms and window types (rectangular, triangular, hanning, hamming, and blackman windows) we have found that the centers of all the attack transients are positioned close to the window center if the COG is close to C_e which is the COG of a linear ramp starting exactly on the left side of the analysis window (window position equals 100%). Therefore, C_e is used in the following to determine whether the transient is properly located for phase reinitialization.

Based on the results obtained so far we may present the principles for a new method for treating transients in the phase vocoder. The basic idea is to determine whether a peak is part of an attack transient by means of its COG. If the COG is above C_e we assume to be in a situation where the attack is close to the start of the window such that reinitialization of the phase is still not appropriate. Because we are in front of an attack we may, however, without perceivable consequences reuse the frequency and amplitude values estimated in the in the previous frame for phase vocoder processing.

If the COG falls below C_e we suppose that the attack transient is located close to the center of the analysis window and at this point we reinitialize the phase for the related bins and restart with phase vocoder processing in the next frame. The reinitialization of the phase exactly reproduces the attack transient for the spectral peak. Due to the fact that the previous frames do not contribute to the transient its amplitude will be slightly to low which can be compensated by increasing the amplitude of the reinitialized bins by 50%.

3 Transient detection

There exist many approaches to detect attack transients (Bonada 2000; Duxbury, Davies, and Sandler 2002; Rodet and Jaillet 2001). In contrast to the algorithm proposed here all those methods are based on the energy evolution in frequency bands, where each band is classified in total as transient or not. For the application in the phase vocoder we want to increase the resolution such that spectral peaks are individually classified. Due to space constraints we will give only a short overview over the transient detector and refer to a more complete discussion available in (Röbel 2003). As shown in that article the COG and the energy derivative are qualitatively similar functions.

The basic idea of the proposed transient detection scheme is straightforward. A transient peak is detected whenever the COG of the peak is above a threshold. Two problems prevent the simple use of this rule. First the phase reinitialization of all partials that belong to the same transient has to be synchronized to prevent a disintegration of the perceived attack. Second, in the case of noise or dense partials amplitude modulation may result which may trigger the transient detector for a non transient situation.

To achieve a robust detection of transients for noisy signals we extend the deterministic transient model described above by means of a statistical model that treats the randomly occurring transient events that are due to modulations of dense sinusoids as a background transient process. The stationary background noise should be distinguished from singular events due to a change of sound characteristics or beginning of a new note. To achieve the statistical description we divide the spectrum into frequency bands with equal bandwidth B and for each band estimate a statistical model that describes the probability of a transient peak using a short history of F_h frames. To detect the singular transient events that are related to instrument onsets we compare this probability with the number of transient peaks in the last F_c frames. The statistical model is a simple binomial model describing the probability of a spectral peak to have COG $> KC_e$ with $K \geq 1$. As is shown later an increase in K decreases the sensitivity of the algorithm and is the major means to control the robustness of the detection. A transient is detected if the transient probability in the current frame is larger than the one in the frame history by at least G-times the standard deviation of the estimated probabilities. After having detected an attack transient we want to assemble all the transient peaks into a single event. Until the end of the attack event is detected all peaks that have a COG above KC_e are collected into a set of transient bins. The transient bins stay in the set until the spectral energy of the bins having a COG above C_e in the current frame is smaller than half the spectral energy contained in the set of bins marked as transient. In this case the phases of all bins in the transient set are reinitialized such that all parts of the same attack are reinitialized in the same frame.

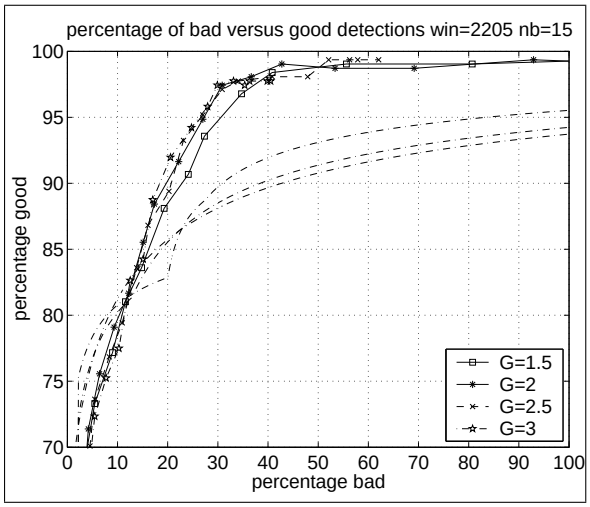

Figure 2: Comparison of the relation between correct and false transients. The new algorithm with window size 50ms and different confidence factors G is compared to the results given in (Rodet and Jaillet 2001) (dash dotted).

To evaluate the performance of the transient detector summarized above we have applied it to a small data base of polyphonic and monophonic sounds introduced in (Rodet and Jaillet 2001) and have compared our results with the results obtained when applying the transient detector presented in the same paper. The database contains a set of 17 hand labeled sound signals with a total of 305 attack transients. For the following experiment the history size to estimate the back ground transient probability has been fixed to contain all frames that are covered by the analysis window. Because the window step is the eights part of the window the history always contains $F_h = 8$ frames and the current transient probability is estimated using the last $F_c = 2$ frames. Increasing the bandwidth B results in increased reliability of the estimated transient probabilities, however, requires more transient bins to significantly change the transient probability. For the experiment below we have used $B = 1500$Hz. The effect of increasing K can be understood as requiring a larger step in energy in the related bins for a transient to be detected. Consequently, the parameter K is a natural means to control the sensitivity of the detection algorithm.

Depicted in fig. 2 are the relations between good and bad transient detections using a window size of 50ms for all the sounds in the data base and for K ranging from 1 to 4.3. For the experimental investigation a transient is considered correctly detected if the hand la-

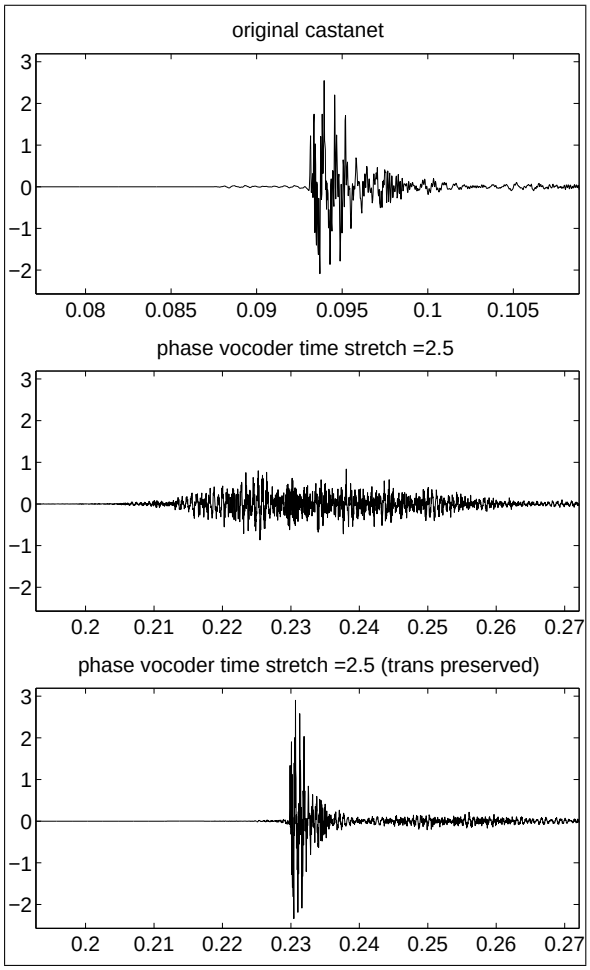

Figure 3: Comparison of original castanet with time stretched version obtained with standard phase vocoder (middle) and new algorithm (bottom) .

beled transient is no further then 10ms away from the region detected as transient by means of the algorithm. Good and false detections are expressed in % of the number of true transients. False and correct detections decrease with increasing K in an approximately exponential relation such that K may be used to control the sensitivity of the detector. Close investigation of the results reveal that the curves for all values of the confidence factor G are nearly superimposed, however, for a larger G a smaller value of K is required to achieve the same result. Therefore, it appears to be sufficient to fix G and provide only K as a user selectable parameter to control the algorithm. From our experiments we conclude that $G = 2$ is a reasonable setting to fix G. Comparing our results to the results obtained in (Rodet and Jaillet 2001) we find that for the new algorithm the number of false detections to accept to achieve a given level of correct detections is considerably lower demonstrating the superior performance.

4 Results

Processing attack transients in the phase vocoder with the proposed algorithm results in significant improvements of attack quality. Therefore, the algorithm has been integrated into AudioSculpt/SuperVP the phase vocoder application of IRCAM. Due to the fact that the algorithm is selectively processing spectral peaks it is well suited for processing multi-phonic sounds, however, the graphical representation of the results would be confusing. Therefore, we have chosen a monophonic castanet sound to demonstrate the performance of the algorithm. The upper part of fig. 3 displays a single castanet sound. In the center the result that has been obtained after time stretching the signal with a standard phase vocoder by a factor of 2.5 is shown. The destruction of the attack event is obvious. In the lower part the same signal has been time stretched by the same factor with transient preservation switched on. The attack is preserved and its sound characteristics are very close to the original attack.

5 Summary

The present article has investigated into the problem of time stretching attack transients with the phase vocoder. We have shown that the group delay of spectral peaks can be used to detect transient peaks and how transient peaks can be preserved during time stretching without fixing the stretch factor to one. The proposed transient detector is especially adapted to be used in the phase vocoder. Nevertheless, it has been shown that it outperforms a previously published algorithm as an independent tool for transient detection.

References

Auger, F. and P. Flandrin (1995). Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. on Signal Processing 43*(5), 1068–1089.

Bonada, J. (2000). Automatic technique in frequency domain for near-lossless time-scale modification of audio. In *Proceedings of the International Computer Music Conference (ICMC)*, pp. 396–399.

Cohen, L. (1995). *Time-frequency analysis*. Signal Processing Series. Prentice Hall.

Dolson, M. and J. Laroche (1999). Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Speech and Audio Processing 7*(3), 323–332.

Duxbury, C., M. Davies, and M. Sandler (2002). Improved time-scaling of musical audio using phase locking at transients. In *112th AES Convention*. Convention Paper 5530.

Griffin, D. and J. Lim (1984). Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing 32*(2), 236–243.

Röbel, A. (2003). A new approach to transient processing in the phase vocoder. In *Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx03)*. To appear.

Rodet, X. and F. Jaillet (2001). Detection and modeling of fast attack transients. In *Proc. Int. Computer Music Conference (ICMC)*, pp. 30–33.

Serra, M.-H. (1997). *Musical signal processing*, Chapter Introducing the phase vocoder, pp. 31–91. Studies on New Music Research. Swets & Zeitlinger B. V.