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 all existing algorithms the new approach is not based on fixing the time dilation parameter to one during transient segments and works locally in frequency such that stationary parts of the signal will not be affected. For transient detection we propose a new algorithm that is especially adapted for phase vocoder applications because its detection criterion has a direct connection to the phase spectrum and estimates the quality of the transformed signal. 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 detected regions and forcing the time stretching factor to be one during the transient regions 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 compensation in non transient regions to achieve the overall 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 mechanisms classifies transients at the level of spectral peaks and the treatment of the individual transients peaks in the phase vocoder is conceptually simple and nevertheless 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 phase spectrum 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 investigate into the relations between different transient detection criteria and in section **5** we summarize the results and discuss the improvements obtained for processing attack transients in the phase vocoder.

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 obtained with the phase vocoder have poor quality. Time stretching attack transients with the phase vocoder results in less severe cases in softening of the perceived attack. In more severe cases a complete change of the sound characteristics may take place.

Based on the theory of time-frequency distributions a clear understanding of the phase evolution for attack transients can be derived. We denote the Fourier spec-

trum of the signal after application of the analysis window h centered at time position t_m to be

$$S_h(w, t_m) = A(w, t_m)e^{j\phi(w,t_m)}. \quad (1)$$

Here w is the frequency in rad and $A(w, .)$ and $\phi(w, .)$ are the amplitude and phase spectrum respectively. As shown in (Cohen 1995) the center of gravity (COG) of the instantaneous energy of the signal $s(t)$ defined as

$$\bar{t} = \frac{\int ts(t)^2 dt}{\int s(t)^2 dt}, \quad (2)$$

can be calculated by means of

$$\bar{t} = \frac{\int -\frac{\partial\phi(w,m)}{\partial w} A(w, m)^2 dw}{\int A(w, m)^2 dw}. \quad (3)$$

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 where the integration has to be replaced by summations and the differentiation with respect to frequency is understood to be performed using the properly interpolated DFT spectrum. Note that the differentiation of the phase with respect to frequency yields the time reassignment operator which can be calculated efficiently by means of a DFT of the signal using a modified analysis window (Auger and Flandrin 1995).

We now consider a single peak in a STFT frame at position t_m and investigate the COG of the signal related to the isolated peak. Due to spectral overlap between different partials we calculate the COG of the peak using only the spectral range in between the two amplitude minima surrounding the spectral peak. By virtue of the amplitude weighting taking place in eq. (3) the error made due to neglecting part of the signal spectrum will be small.
Without loss of generality we assume that the time origin is in the center of the analysis window. From the theory of Fourier analysis we know that for stationary partials the phase will be constant and, therefore, the COG is 0. For a partial that is starting inside the analysis window the COG is located to the right of the window center and the average phase derivative is negative. The absolute value of the phase slope is constantly decreasing when the window is moving over the partial. The phase slope itself and its change with the window position is violating the assumption made for the phase manipulations in the phase vocoder and this is the main reason for the artifacts produced with the phase vocoder when processing attack transients.

Because the phase slope can be used as an indicator for how well the phase vocoder will do when it is processing the transient we consider the COG an informative candidate for transient attack detection. To understand the phase properties of attack transients we have investigated into a simple attack transient model - a linear ramp with saturation. In fig. 1 the decrease

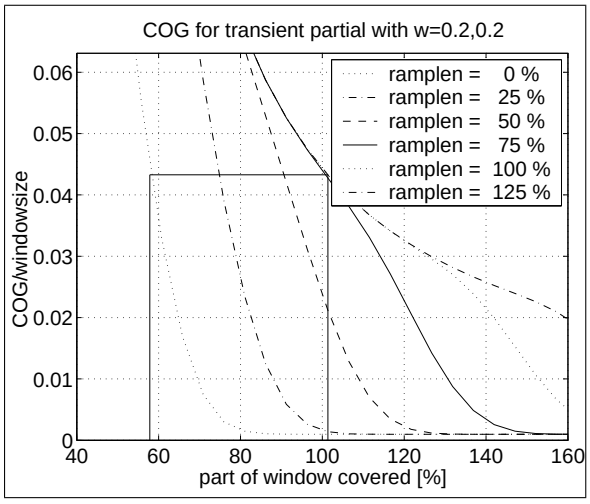

Figure 1: Center of gravity of partial energy as a function of transient position under the analysis window for transient partials with fixed frequency and different length of linear ramp (in percent of window size). The selected threshold C_e (see text) is marked.

of the COG (and the magnitude of the phase slope) with the analysis window moving over attack transients with different ramp length is shown. 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.

Based on the results displayed in fig. 1 we may derive 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 a threshold C_e we assume to be in a situation where the attack is not yet sufficiently covering the window such that the phase vocoder phase treatment should not be applied. Because we are in front of an attack we may, however, without perceivable consequences reuse the frequency and amplitude values estimated in the same bin in the previous frame for phase vocoder processing. As threshold C_e we use a displacement of the COG of 4.4% of the signal window for which we conclude from fig. 1 that all types of transients would at least cover 60% of the window. If the COG falls below the threshold we know that phase processing in the phase vocoder becomes sufficiently reliable. 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 properly reproduces the attack transient for the spectral peak. It has been verified experimentally that for the given threshold the phase manipulation of the phase vocoder will only produce minor artifacts. Note that due to taking the average of the group delay over the peak the relations displayed in fig. 1 will remain nearly unchanged if the partial under investigation is not stationary but has a linear frequency evolution.

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. This however is not a fundamental difference because as shown in section **4** there exists a close relation between the COG and the energy derivative with respect to time. All but the last of these algorithms work with rather low frequency resolution classifying frequency bands, only.

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. Moreover in the case of noise or dense partials, where dense here is related to the frequency resolution of the analysis window, amplitude modulation with a modulation rate in the order of the window length may result which with non zero probability may match the transient criterion.

First we consider the problem of noise and amplitude modulation. To prevent false detection of modulations as transients we need to increase the information about the signal properties by means of statistical description. We divide the spectrum into frequency bands of fixed size (for the experiments presented later a size of 3000Hz has been used) and for each band estimate a statistical model that describes the probability of a transient peak using a short history of frames and compare this probability with the number of transient peaks in the current frame. The statistical model is a simple binomial model describing the probability of a spectral peak to have COG $> C_s$. For transient detection with the statistical model we use a threshold $C_s > C_e$ because in the beginning of a transient the COG should be larger than in the end. The number of events N of the statistical process is determined by the maximum number of peaks that may be contained in the frequency band. This is simply the frequency band width divided by peak band width according to the window length. To obtain a robust detection criterion we consider a safety margin when estimating the transient peak probabilities. Using the formula for the variance of a binomial distribution with transient peak probability p

$$\sigma^2 = p(1-p)N$$

we want to select the transient probability such that it is consistent with the number of observed transient hits n within the range of G times the standard deviation of the mean value pN. Therefore, for p we require

$$n = pN \pm G\sigma = pN \pm G\sqrt{p(1-p)N}.$$

where the plus and minus sign are used to estimate the transient probability for the current frame and frame history respectively. Solving for p we obtain

$$p = \frac{G^2N + 2nN \mp G\sqrt{N(G^2N + 4nN - 4n^2)}}{2N(G^2+N)}.$$

An attack transient is detected if in any of the frequency bands the transient probability in the current frame is larger than the transient probability in the frame history. 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 C_e are collected into a set of transient bins. Bins stay in the set even if their COG falls below the threshold. The attack is finished when 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. The transient collection ensures that all parts of the same attack are reinitialized in the same frame such that no attack disintegration will take place.

While the attack transient detector described here has been especially developed to work in connection with the phase vocoder it can also be used as a tool for transient detection. To evaluate its performance we have compared the results obtained for a small data base of polyphonic and monophonic sounds with the results obtained with the algorithm presented in (Rodet and Jaillet 2001). The data base has already been used in the same paper. It contains a set of 17 hand labeled sound signals with a total of 305 attack transients. For the following experiments 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 8 frames. The safety margin factor G has been fixed to 3.

There are two user selectable parameters for the transient detector. The first one is the analysis window size. With respect to this parameter there may exist contradicting demands because on one hand tonal attack transients will be detected more reliable if the partials of subsequent notes are individually resolved. On the other hand for sequences of broad band attack transients the individual attacks will be detected only if the distance between two attacks is less than the window size. For the evaluation we have selected the window sizes of 35 and 45ms, which are reasonable parameters for processing the sounds with the phase vocoder. The second parameter is the transient start threshold C_s. Depicted in fig. 2 are the relations between good versus bad transient detections for the all sounds in the data base and for C_s ranging from 5% up to 27% of the window size. A transient is considered correctly detected if the hand labeled transient is no further then 10ms away from the region detected as transient by means of the algorithm. All other detections are counted as false. In fig. 2 good and false detections are expressed in % of the number of true transients. When decreasing C_s from 0.27 the number of correct detections increases

and saturates with nearly 99% correct detections for $C_s = 0.08$. Comparing the results we conclude that for the new algorithm the number of false detections to accept to achieve a certain level of correct detections is considerably lower for all interesting cases.

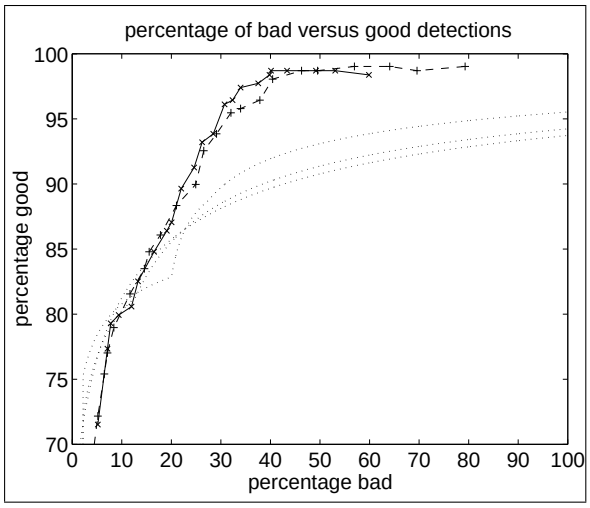

Figure 2: Comparison of relation between correct and false transients. The proposed algorithm with window size 45ms (solid) and 35ms (dashed) is compared to the results presented in (Rodet and Jaillet 2001) (dotted) .

4 Relation to other algorithms

As stated earlier transient detection algorithms are usually making their decisions based on the time evolution of the signal energy. In the following we show that the COG is closely related to the change of energy with time. From the theory of reassignment we know that the group delay is equal to

$$-\frac{\partial}{\partial w}\phi(w,n_0) = -\operatorname{real}\frac{\overline{S_h(w,n_0)}S_{hT}(w,n_0)}{|S_h(w,n_0)|^2}$$

where S_{hT} is the Fourier transform of the signal segment using the analysis window multiplied with a time ramp having its origin in the center of the window. For the derivative of the energy of the spectral bin with respect to window position and normalized by its spectral energy we obtain

$$\frac{\partial|S_h(w,n_0)|^2}{|S_h(w,n_0)|^2\partial n_0} = -2\operatorname{real}(\frac{\overline{S_h(w,n_0)}S_{dh}(w,m)}{|S_h(w,m)|^2}$$

which besides a constant factor 2 uses the derivative of the analysis window with respect to time dh instead of hT for the modified Fourier transform. dh and hT having qualitatively the same form both criteria should yield similar result. As shown above the COG has the advantage to have a clear relation to the phase evolution of transient peaks.

5 Summary and Results

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 well suited to be used in the phase vocoder and, moreover, has been shown to outperform a previously published transient detection algorithm.

Processing attack transients in the phase vocoder with the proposed algorithm results in striking improvements of attack quality. As an example we have time stretched a piece of polyphonic music with castanet sounds by more then a factor 2 without affecting the sound characteristics of the castanet transients. Further research will address the problem to further increase the reliability of the transient detection and automatically adapt the size of the analysis window to the content of the signal to reduce the impact of the contradicting demands concerning the window size during stationary tonal segments and segments with dense sequences of attack transients.

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.

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.