A new approach to transient processing in the phase vocoder

Axel Roebel

A NEW APPROACH TO TRANSIENT PROCESSING IN THE PHASE VOCODER

Axel Röbel

IRCAM, Analysis-Synthesis Team, France
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 transient preservation algorithms that have been proposed up to now the new approach does not impose any constraints on the time dilation parameter for processing transient segments. By means of an investigation into the spectral properties of attack transients of simple sinusoids we provide new insights into the causes of phase vocoder artifacts and propose a new method for transient preservation as well as a new criterion and a new algorithm for transient detection. Both, the transient detection and the transient processing algorithms are designed to operate on the level of spectral bins which reduces possible artifacts in stationary signal components that are close to the spectral peaks classified as transient. The transient detection criterion has a close relation to the transient position and allows us to find an optimal position for reinitializing the phase spectrum. The evaluation of the transient detector by means of a hand labeled data base demonstrates its superior performance compared to a previously published algorithm. Attack transients in sound signals transformed with the new algorithm achieves high quality even if strong dilation is applied to polyphonic signals.

1. INTRODUCTION

The phase vocoder [1] is widely used for signal transformation. Due to recent advances [2] 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 [3, 4] and it has been shown that significant improvements concerning the sound characteristics of transients can be achieved if the phase relations between transient bins are kept unchanged. In existing algorithms this is accomplished 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. 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 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 simplified due to the fact that there is no need to force the stretch factor to one if the phase initialization is done when the transient is close to the center of the window. Despite this simplifications the algorithm reproduces the transients with subjectively high quality.

In section **2** of this article we investigate into the problem of processing attack transients with the phase vocoder. Based on the theoretical understanding of the spectrum of transient sinusoids we propose a conceptually simple yet effective transient processing scheme. In section **3** a transient detection algorithm is developed that is based on an estimation of the position of the signal energy and is especially adapted for the application in the phase vocoder. In section **4** the performance of the new transient detector is evaluated using a small data base of hand labeled sounds and it is shown that it outperforms a recent algorithm. In section **5** we investigate into the relations between different transient detection criteria and in section **6** 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 [5]. 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 that needs to be applied can be derived for properly resolved and nearly stationary sinusoids [1, 2]. If the amplitude of a sinusoid 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.

To understand the origin of the problems that arise when processing attack transients with the phase vocoder we will investigate into the phase and amplitude spectra of attack transients of a single sinusoid. The attack model that is used in the following is a linear ramp with saturation. The signal is analyzed by means of moving the analysis window over the attack and performing a STFT. Without loss of generality we assume that the time origin is moving and is always in the center of the analysis window.

We denote the Fourier spectrum of the signal $s_h(t, t_m)$ which is the signal $s(t)$ windowed with the analysis window centered at

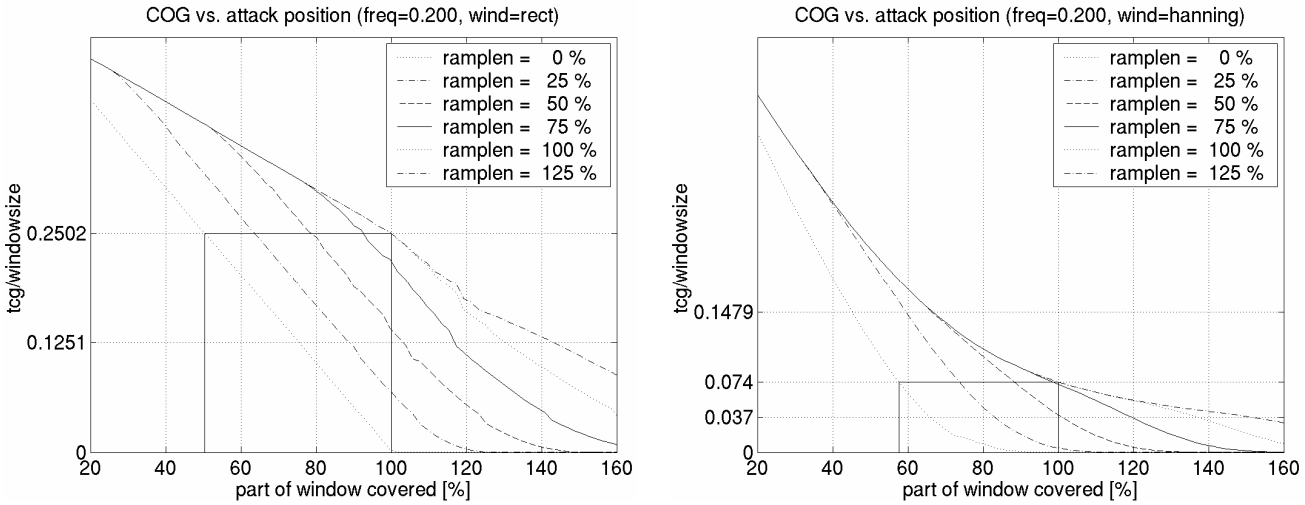

Figure 1: *Center of gravity of partial energy (according to eq. (4)) as a function of transient position under the analysis window for transient partials with fixed frequency $w = 0.2\pi$ and different length of linear ramp (in percent of window size). Window type used is rectangular (left) and hanning (right). The thresholds C_e (see text) indicating proper transient position for phase reinitialization are marked.*

time position t_m, $h(t, t_m)$, 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 [6] the center of gravity (COG) of the instantaneous energy of the windowed signal $s_h(t, t_m)$ defined as

$$t_{cg} = \frac{\int t s_h(t, t_m)^2 dt}{\int s_h(t, t_m)^2 dt}, \quad (2)$$

can be calculated by means of

$$t_{cg} = \frac{\int -\frac{\partial\phi(w,t_m)}{\partial w} A(w, t_m)^2 dw}{\int A(w, t_m)^2 dw}. \quad (3)$$

The negative phase derivative, called group delay, determines the contribution of a frequency to this position. While equations (1) - (3) are derived for time continuous signals the same type of relations can be established for the DFT of discrete time signals where the integrations have to be replaced by summations and the differentiation with respect to frequency is understood to be performed using the properly interpolated DFT spectrum. The origin of the coordinate system for the sample positions has to be chosen consistently when calculating the DFT. Note, that the differentiation of the phase with respect to frequency, the group delay, is equal to the time reassignment operator which is calculated efficiently by means of a Fourier transform (or DFT) of the signal using a modified analysis window [7].

If the analysis window is moved from the left over the attack of the sinusoid the COG is first located to the right of the window center and correspondingly the average phase derivative is negative. Due to the small duration of the windowed signal the band width of the peak is large. Moving the window further to the right results in the COG moving to the left such that the absolute value of the phase slope is decreasing together with the bandwidth of the peak. Finally, the phase slope becomes zero and the peak reaches its minimum bandwidth if the window has completely moved over the attack transient in which case we have reached the stationary part of the sinusoid.

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 if the shape of the transient signal shall be recreated. After reinitialization of the phase spectrum two further problems require investigation. The phase slope itself and its change with the window position as well as the change in peak bandwidth are violating the assumptions that are made for deriving the phase manipulations to be applied when time stretching a signal with the phase vocoder.

First we consider the phase slope that is changing with the position of the transient within the window. In fig. 1 the decrease of the COG (the decay of the phase slope) is shown that results if the analysis window moves over sinusoids having an attack transient of different ramp length. The analysis windows that have been used in fig. 1 are a rectangular (left) and a hanning window (right). The ramp length of the attack phase is given in percent of the analysis window length and the window position is given in terms of the position of the right end of the window relative to the start of the attack. The window position is normalized by window length and expressed in percent of the window length. As shown in fig. 1 the relation between COG (and with it the phase slope) and transient position is nearly linear over a wide range of positions, especially, if the transient is close to the window center. This is advantageous because a linear relation is handled without any error by the standard phase vocoder algorithm. In this case the frequency that will be estimated for the different bins within a single peak would deviate by a constant offset from the frequency of the sinusoid and the time scaling procedure would result in correctly predicted phase spectra. For frame relocation with offsets smaller than an 8th part of the analysis window the error in the phase spectrum is negligible such that the discontinuities are prop-

erly located and coherend overlap and add is guarranteed.

The second problem is due to the fact that the amplitude spectrum, i.e. the peak bandwidth and side lobe positions, are kept unchanged when the frame is repositioned within the phase vocoder. The evolution of the amplitude spectrum depends in a complicated manner on the transient position and on the form of the transient such that it appears to be impossible to adapt the amplitude spectrum according to the new frame position. To be able to quantify the error made by the phase vocoder when relocating transient signals we have studied the normalized root mean squared error (NRMSE) between the signal frame after relocation with the correct signal at the new frame position taking into account a synthesis window eual to the analysis window. We found that the total NRMSE due to phase and amplitude errors decreases with decreasing frame offset during relocation of the frame and with increasing overlap between transient sinusoid and the analysis window. To give an idea of the error range we note that for a step transient the NRMSE is below 25% for transient positions in the center of the window and if the frame relocation offset is smaller than an 8th part of the window size.

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. The first reason concerns the exact reconstruction of the transient during synthesis. Because the transient is reproduced without any error only in the frame that gets the phases of the transient bins reinitialized the reinitialization should take place when the impact of the reconstructed transient on the output signal is largest The second argument is concerned with the error in transient position. Reinitialization of the phases will produce the transient at the very same position where it was located in the analysis window. During synthesis the frame is repositioned to its new location using the frame center as time reference. To avoid the need to reposition the transient to properly fit into the transformed time evolution phase reinitialization should take place when the transient is in the center of the window.

2.1. A new method for transient processing

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. A COG threshold C_e is used to determine when the transient is sufficiently close to the window center to reinitialize the phases of the transient bins. If the COG is above the threshold 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 same bin 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 did 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%.

2.2. Determining transient position

To be able to determine transient positions for sinusoidal components that are part of a conglomerate spectrum we need to modify the estimation of the COG such that it operates local in frequency. This is achieved by means of considering each spectral peak independently and limit the integral in eq. (3) to the frequencies located between the amplitude minimum surrounding each peak. Consequently, the COG is calculated using

$$t_{cg} = \frac{\int_{w_l}^{w_h} -\frac{\partial\phi(w,t_m)}{\partial w} A(w,t_m)^2 dw}{\int_{w_l}^{w_h} A(w,t_m)^2 dw}, \qquad (4)$$

where w_l and w_h are the positions of the amplitude minima below and above the current maximum respectively. Due to the amplitude weighting taking place the difference between eq. (3) and eq. (4) will be small as long as the partial is sufficiently resolved. For sinusoids that are to close in frequency to be individually resolved the treatment of individual peaks performs a somewhat arbitrary signal decomposition which nevertheless will correctly detect transient situations as long as all the sinusoids that are contributing to the same peak are transient.

From fig. 1 we conclude that by means of a simple threshold test for the COG of the peak main lobe we may detect whether the attack is roughly positioned in the center of the analysis window. Note, that in the coordinate system used to express transient position in fig. 1 the optimal transient position varies with the ramp length of the transient. Moreover, the thresholds to apply to obtain a desired transient position depend on the analysis window that is used. By means of comparing the COG evolution for different transient forms and window types (besides the ones shown in fig. 1 we analyzed triangular, hamming, and blackman windows) we have found that the center of the attack transient is properly positioned close to the window center if the COG is close to the COG of a linear ramp starting exactly on the left side of the analysis window (window position equals 100%). Therefore, the COG that is related to a linear ramp just starting at the left side of the analysis window has been selected as threshold C_e that will be used to determine when the phase spectrum of the transient bins should be reinitialized.

3. TRANSIENT DETECTION

There exist many approaches to detect attack transients [3, 8, 4, 9]. 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 and we will show in section **5** that the COG and the energy derivative with respect to time are qualitatively similar functions. As a further difference we note that all but the last of these algorithms work with rather low frequency resolution classifying frequency bands instead of single peaks, 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. Second, in the case of noise or dense partials (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 depending on the window position may result in a COG $> C_e$ triggering the transient detector for a non transient situation.

First we consider the problem of noise and amplitude modulation. The detection of attack transients in case of dense collections of sinusoids is important from a perceptional point of view to correctly handle percussive sounds. The erroneous detection of attack transients in noisy regions, however, should be avoided because the artificial treatment of the phase in the pre-transient regions would result in subjectively perceivable changes of the sound characteristics.

In the following we will 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 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 $> C_s = KC_e$ with $K \geq 1$. As will be shown later in the experimental evaluation of the algorithm an increase in K decreases the sensitivity of the algorithm and is the major means to control the robustness of the detection.

The number of independent events N of the statistical process is determined by the maximum number of peaks that may be contained in a frequency band. This is simply the bandwidth of the frequency bands divided by peak bandwidth according to the length of the analysis window and multiplied by the number of frames, F_c or F_h respectively. A further means to control the robustness of the detection is the confidence level required when testing for a change in transient probability between the frame history and the current frames. Using the formula for the variance for a binomial distribution with transient peak probability p

$$\sigma^2 = p(1-p)N \tag{5}$$

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

$$n = pN \pm G\sigma = pN \pm G\sqrt{p(1-p)N}. \tag{6}$$

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

$$p_c = \frac{G^2 N_c + 2n_c N_c - G\sqrt{N_c(G^2 N_c + 4n_c N_c - 4n_c^2)}}{2N_c(G^2 + N_c)} \tag{7}$$

$$p_h = \frac{G^2 N_h + 2n_h N_h + G\sqrt{N_h(G^2 N_h + 4n_h N_h - 4n_h^2)}}{2N_h(G^2 + N_h)}, \tag{8}$$

where N_x and n_x are the number of independent events and observed transient peaks in the frame history (for $x = h$) and the current frames (for $x = c$), respectively. An attack transient is detected if in any of the frequency bands the transient probability in the current frames p_c is larger than the transient probability in the frame history p_h.

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_s are collected into a set of transient bins. This set is non contracting and 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.

4. EXPERIMENTAL RESULTS

While the attack transient detector described above has been developed especially to work in the phase vocoder it can also be used as a stand alone tool for transient detection. To evaluate its performance we have applied it to a small data base of polyphonic and monophonic sounds introduced in [9] 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 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 $F_h = 8$ frames. For estimating the actual transient probability we have experimented with $F_c = 1$ and $F_c = 2$. Both settings provide similar results with slight advantage for $F_c = 2$ which will be used in the following experiments.

There remain four user selectable parameters for the transient detector. The first one is the analysis window size. With respect to this parameter there exist contradicting demands because on one hand attack transients of sinusoids that mix with stationary sinusoids will not be correctly detected such that frequency resolution should be high and window size large. On the other hand we can not detect more than one broadband attack transient within a single window such that window size should be small. This is a variant of the well known time resolution/frequency resolution trade off for time frequency analysis. For evaluating the transient detector two experimental setups have been used. In the first experiment the same window size of 50ms has been applied to all signals in the database. While the application of the same window size for all sounds is certainly suboptimal, it allows us to compare with the results presented in [9] which have been achieved with a single set of parameters, too. In the second setup which is closer to practical applications we choose for each sound and parameter K the optimal window size out of the set $[35ms, 45ms, 50ms, 55ms]$ by determining the window size that obtains the largest number of correct transient hits.

The second parameter is the threshold factor K. A simple theoretical investigation shows that for the noise free case the maximum COG normalized by the analysis window is 0.5 and for maximum robustness C_s should be close to this value. Due to background noise or preceding notes, however, part of the transient may be covered in real signals such that the maximum value of the observed COG will generally be lower than 0.5. As shown in section **5** the COG has a close relation to the energy derivative and we may understand the parameter C_s (or K) to control the jump in energy that is required for a transient to be detected. Therefore,

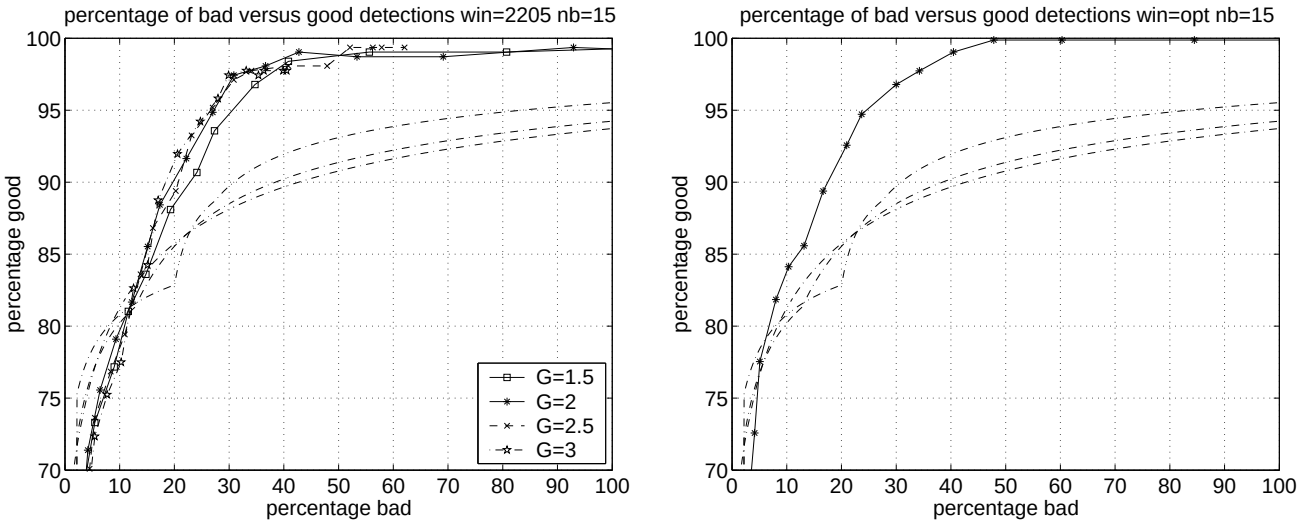

Figure 2: *Comparison of relation between correct and false transients for frequency bands of bandwidth 1500Hz. The proposed algorithm with window size 50ms and different confidence factors G (left) and a fixed confidence factor $G = 2$ together with the optimal window for each sound and threshold C_S (right) is compared to the results presented in [9] (dash dotted).*

the parameter K is a natural means to control the sensitivity of the detection algorithm.

The third parameter is the bandwidth of the frequency bands that are used to obtain the statistical model for background transient activity. By increasing the bandwidth we increase the reliability of the transient probability estimation, however, at the same time we increase the number of bins that have to be affected by a transient event to trigger the transient detector. For the experiments in the following we tried bandwidths ranging from 1100Hz to 4400Hz which may appear to be rather broad, however, given a frequency resolution of about 70-110Hz (hanning window) this results in about 10-60 independent events per frame and band. In our experiments the results did change only weakly with the bandwidth with a slight optimum for a bandwidth of 1500Hz. Therefore, we have fixed the bandwidth to this value for the discussion of the experimental results presented here.

The last parameter is the variance factor G that is used to control the confidence in detecting a change in transient probability. G has been varied in the range $[1.5, 3.5]$.

Depicted in fig. 2 are the relations between good versus bad transient detections for all the sounds in the data base and for K ranging from 1 up to 4.3. For the experimental investigation 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. Good and false detections are expressed in % of the number of true transients. In the left part of fig. 2 we compare the results for four different values of $G = [1.5, 2, 2.5, 3]$. Generally, for $K = 1$ the number of correct detections is close to 100%, however, the number of false detections is much larger than 100% and we are outside the region displayed. Increasing K reduces the number of false detections first with a nearly stable amount of correct detections. Above a certain value of K the number of correct detections falls approximately linearly with the number of false detections. Close investigation of the results shown in fig. 2 reveal that the curves for all values of G that have been used 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. The possibility to exchange K and G is obviously limited because for very large G even switching all bins to transient will not provide sufficient confidence. From the experiments conducted so far it appears that $G = 2$ is a reasonable setting to fix G. In the right part of fig. 2 we used $G = 2$ and searched to find the window that maximizes the number of correct transient hits for each value K and for each sound. As shown in the figure the relation between correct and false detections does improve slightly when using the optimal window.

Comparing the results to the once obtained in [9] 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, which demonstrates its superior performance.

5. RELATION TO OTHER ALGORITHMS

As mentioned above 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

$$-\frac{\partial}{\partial w}\phi(w,t_m) = -\operatorname{real}\frac{\overline{S_h(w,t_m)}S_{hT}(w,t_m)}{|S_h(w,t_m)|^2} \qquad (9)$$

where $S_h(w,.)$ and $S_{hT}(w,.)$ are the Fourier transforms of the signal s using the windows h and h_T centered at position t_m. The window h_T is obtained from the analysis window h by multiplication with a time ramp having its origin in t_m. If we calculate the derivative of the spectral energy $|S_h(w,t_m)|^2$ with respect to window position t_m and normalize the derivative by the spectral

energy we obtain

$$\frac{\partial |S_h(w,t_m)|^2}{|S_h(w,t_m)|^2 \partial t_m} = -2\,\mathrm{real}(\frac{\overline{S_h(w,t_m)} S_{hd}(w,t_m)}{|S_h(w,t_m)|^2} \qquad (10)$$

which besides a constant factor 2 can be derived from eq. (9) by replacing the Fourier transform using the window h_T by a Fourier transform using the window h_d which is the derivative of the analysis window with respect to time. Because h_d and h_T are qualitatively similar functions the group delay eq. (9) and the normalized derivative of the spectral energy eq. (10) will be similar functions as well. As a consequence, it should to be possible to derive a transient detection algorithm with similar performance formulated in terms of the normalized derivative of spectral energy.

6. 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. For the graphical representation of the results, however, we have chosen a monophonic castanet sound to simplify the interpretation of the performance of the algorithm. The upper part of fig. 3 shows the time signal of a single beat within a sequence of castanet sounds. Beneath 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. At the bottom of the figure the same signal has been time stretched by the same factor with transient preservation switched on. The attack is preserved and the sound characteristics of the attack are very close to the original attack.

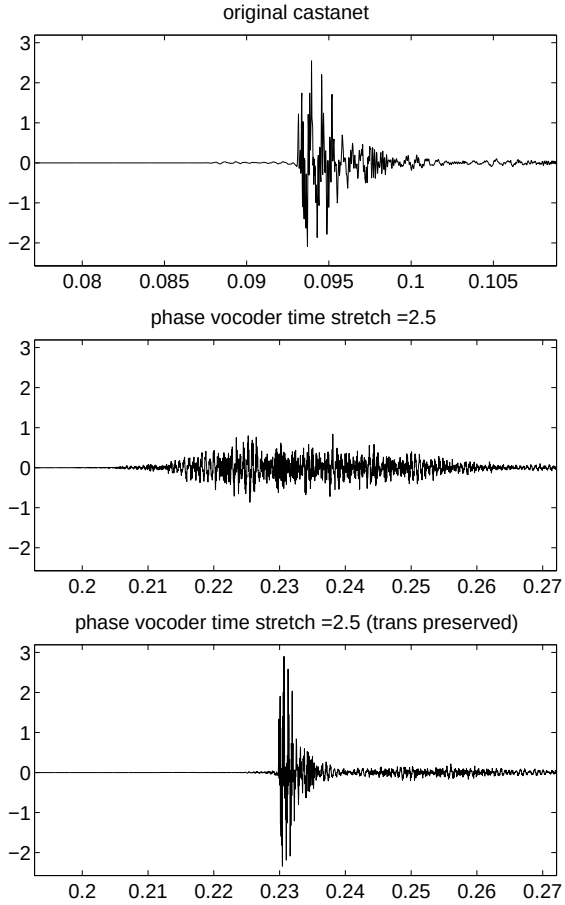

Figure 3: *Comparison of original castanet (top) with time stretched version obtained with standard phase vocoder (center) and new algorithm with K=1.5 and bandwidth=1500Hz (bottom).*

7. 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. Due to the fact that the group delay may be interpreted in terms of transient position the proposed transient detector is especially adapted to be used in the phase vocoder. Moreover, it has been shown that it has a close relation to energy derivate based transient detectors and that it outperforms a previously published algorithm if used as an independent tool for transient detection.

8. REFERENCES

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

[2] M. Dolson and J. Laroche, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

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

[4] C. Duxbury, M. Davies, and M. Sandler, "Improved time-scaling of musical audio using phase locking at transients," in *112th AES Convention*, 2002, Convention Paper 5530.

[5] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[6] L. Cohen, *Time-frequency analysis*, Signal Processing Series. Prentice Hall, 1995.

[7] F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," *IEEE Trans. on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[8] P. Masri and A. Bateman, "Improved modelling of attack transients in music analysis-resynthesis," in *Proceedings of the International Computer Music Conference (ICMC)*, 2000, pp. 100–103.

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