Transient Detection with Absolute Discrete Group Delay

Volker Gnann and Martin Spiertz
Institute of Communications Engineering
RWTH Aachen University
Aachen, Germany
E-Mail: {gnann, spiertz}@ient.rwth-aachen.de

***Abstract*—This paper presents a new transient detection algorithm which uses the average absolute discrete group delay as a measure for the transient characteristic of sound. It shows that a bell-shaped window function performs a high-pass effect to the spectral coefficients, leading to a concentration of group delay values in the π-area for steady-state signals. This concentration is violated if a transient occurs. From this phenomenon, we derive a new transient detection method, improve it by a maximum order-filter, and show that it works well on percussive and tonal-percussive sounds.**

I. Introduction

A *transient* is an audio signal portion which changes quickly and in a non-predictable way [1]. The detection of transients is important for applications like onset detection or the determination of the best time/frequency resolution trade-off. Beside others, one class of transient detection algorithms is based on the short-time Fourier transform (STFT). These algorithms usually analyze the evolution of the STFT magnitudes or the deviation of phase information across the time axis.

Time-frequency representations like the STFT allow a signal analysis in the time and frequency domain. However, they lead to a resolution trade-off in both domains: A high temporal resolution leads to a low frequency resolution and vice versa. Within one time frame, the temporal behavior is encoded in the phase information. Phase-locking vocoders [4] exploit this fact to yield a more time-accurate result in transient regions. This leads to the idea of analyzing the phase information *across the frequency axis* to detect transients. Therefore, we use the group delay $\Delta\varphi/\Delta f$ as an auxiliary measure.

Similar measures have already been developed for glottal pulse extraction [5] and for beat detection [3]. These approaches search for zero crossings in the derivative of the *unwrapped* phase over frequency. In contrast, our algorithm works directly on the phase difference restricted to the $[-\pi, \cdots, +\pi]$ range. Instead of searching for zero crossings, it uses the absolute value's average of the group delay. As we show, for common window functions, this value approximates π for steady-state signals, but not for transients.

This paper is organized as follows: Section II analyzes the phase coincidences in steady-state conditions and during attack slopes. Furthermore, we introduce the absolute value of the group delay as a measure of phase coincidence, describe the influence of the analysis window shape, and present a transient detection criterion based on these principles. Section IV introduces several modifications of this detection function which lead to improvements. Section V-C illustrates the properties of this detection function with practical examples and an onset-based evaluation. The paper closes with conclusions and an outlook.

II. Fundamentals

The STFT of a given discrete-time signal $x(n)$ is given as

$$X(m,k) = \sum_{n=-\infty}^{\infty} x(n)w(mS-n)e^{-j\frac{2\pi k}{N}n}, \tag{1}$$

where w denotes the analysis window, m the frame index for the STFT, and S the hop size between two analysis frames.

The STFT coefficients $X(m,k)$ are complex numbers and thus can be expressed in polar coordinates as magnitude $|X(m,k)|$ and phase $\varphi_X(m,k)$, where $\varphi_X(m,k) = \arg X(m,k)$ denotes the angular argument of $X(m,k)$. To restrict the range of all phase information to $[-\pi, \cdots, +\pi]$, we use the principle argument function [7]:

$$\mathrm{princ}(\varphi) := ((\varphi+\pi) \bmod (-2\pi)) + \pi. \tag{2}$$

The *group delay* is usually defined as the negative derivative of the unwrapped phase with respect to the frequency, $-d\varphi/df$. In this paper, we use the phase difference between two adjacent frequency bins and restrict the group delay to $[-\pi, \cdots, +\pi]$. Therefore, no phase unwrapping is needed. To distinguish this phase difference from the common group delay, the *discrete group delay* is given by

$$D_X(m,k) = \mathrm{princ}\left(\varphi_X(m,k) - \varphi_X(m,k-1)\right). \tag{3}$$

To determine whether an STFT time frame contains a transient or not, we use the average absolute value of this group delay. The absolute value operator maps the result from the $[-\pi, \cdots, +\pi]$ range to the $[0 \cdots \pi]$ range and thus delivers information whether the phase differences w.r.t. frequency concentrate around the π area or not. In the following, we abbreviate this average absolute discrete group delay $\hat{\mathcal{D}}$ as AAGD:

$$AAGD = \hat{\mathcal{D}}_X(m) = \frac{1}{N}\sum_{k=1}^{N} |D_X(m,k)| \tag{4}$$

III. AAGD with Different Windows

A. Rectangular Windows

In the following, we consider $w(n)$ to be the quasi-symmetric rectangular window, which does not change the phase information except adding a time shift by 1/2 sample.[1]

In steady-state parts of audio signals, the phase $\varphi_X(m,k)$ depends mainly on the previous time frame phase for the same frequency bin, $\varphi_X(m-1,k)$. Phase vocoder theory ([7], pp. 246) models the STFT output as a sequence of a variable-frequency sinusoid oscillator bank and a filter bank. For each frequency band k, let the assumed oscillator frequency at a given time index m be $f_k(mS)$. Then, we can calculate the actual phase approximately as

$$\varphi_X(m,K) = \varphi_X(0,K) + \int_0^{mS} 2\pi f_k(\tau)d\tau \tag{5}$$

[1] As STFT are usually implemented with the FFT, an even window size is preferable. On the other hand, it is not possible to create an even-sized, time-discrete, symmetric window function that includes the y axis as a sample index. For that reason, the window FFT will always include a linear phase term representing a 1/2-sample time shift. We neglect this shift in the following.

We can determine the instantaneous frequency f_k by exploiting the phase difference of the two preceding time frames. f_s denotes the sampling frequency:

$$f_k = f_S \cdot \left(\frac{k}{N} + \frac{\varphi_X(m,k) - \varphi_X(m-1,k)}{2\pi} \right) \tag{6}$$

Since f_k depends on the time index m, we can assume from the phase vocoder model that the phases within one time frame are *independent*. To confirm this assumption, we measured the distributions of phase differences within time frames in steady-state context for some signals. The results are given in Fig. 1 (solid lines). They show an approximately equal distribution in the case of noise. In the case of tonal instruments, the phase distribution has a primary weight in the area around zero.

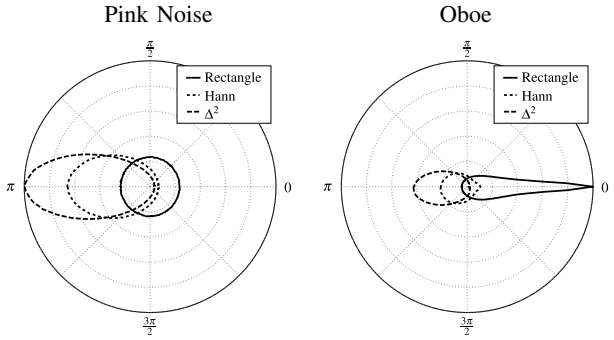

Fig. 1. Distribution of phase differences for pink noise and for an oboe signal. Each signal is windowed with a rectangle, a Hann, and a squared triangle window, respectively. Note that the Hann and the squared triangle window tend to transform the phase differences to the π area.

B. Bell-Curved Window Shapes

Fig. 1 also contains the distribution of the wrapped phase difference for the Hann and the squared triangle window (see also Fig. 2). We can observe that the phase difference distributions for these windows are emphasized around π. This effect can be described best using the convolution theorem. Let a time frame m be given. The signal $x(n)$ is multiplied in the time domain with $w(n)$. This corresponds to a convolution in the frequency domain:

$$y(n) = x(n) \cdot w(n) \circ\!\!-\!\!\bullet\, Y(k) = X(k) * W(k) \tag{7}$$

In the following, we treat the convolution in the frequency domain as FIR "filtering". Since the "filter" works on the frequency domain,

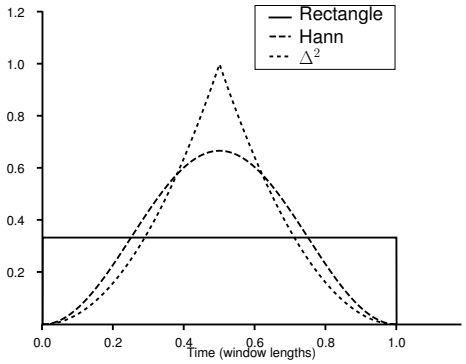

Fig. 2. Rectangle, Hann, and squared triangle window. The windows are normalized such that they contain the same area.

TABLE I
MAX.-AVG. RATIO $\mathcal{P}$ (SEE SECTION III-B1) AND AAGD $\hat{\mathcal{D}}$ FOR RECTANGLE, HANN AND SQUARED TRIANGLE WINDOWS. THE AAGD WAS MEASURED WITH PINK NOISE AND WITH A PIANO SIGNAL.

Window w	$\mathcal{P}$	$\hat{\mathcal{D}}$ Pink noise	$\hat{\mathcal{D}}$ filter Pink noise	$\hat{\mathcal{D}}$ Piano	$\hat{\mathcal{D}}$ filter Piano
Rectangle	1.000	1.568	2.614	0.491	0.935
Hann	2.001	2.301	2.956	2.302	2.961
Squared Triangle	2.997	2.489	3.002	2.514	3.007

we call this process *meta-filtering*, and the spectral domain this filter works on *meta-frequency domain*. The impulse response of the meta-filter is given by the DFT coefficients of the window.

An important property of these meta-filter impulse responses is their high-pass character. To prove this, we show first that the meta-frequency domain is actually the time domain.

As shown in [2], the inverse DFT is equivalent to the DFT of the time-reversed signal:

$$\mathrm{IDFT}_n\left\{X_k\right\} = \sum_{k=0}^{N-1} X_k e^{j\frac{2\pi}{N}kn} = \mathrm{DFT}_{-n}\left\{X_k\right\}. \tag{8}$$

We can follow that

$$\underbrace{\mathrm{DFT}_n\left\{\mathrm{DFT}_k\left\{x_n\right\}\right\}}_{\text{meta-spectral domain}} = \mathrm{IDFT}_{-n}\left\{\mathrm{DFT}_k\left\{x_n\right\}\right\} = \underbrace{x_{-n}}_{\text{time domain}} \tag{9}$$

Since all common window functions are symmetric and considered periodic in DFT analysis, we can assume that $w[-n] = w[n]$. As a result, the meta-filter transfer function is simply the window function in the time domain. For better understanding, we will nevertheless call the time domain "meta-spectral domain" when we refer to the domain of the spectrum of the spectrum.

Most common window functions are symmetric, bell-shaped, and have their maximum at the center. In the meta-spectral domain, this corresponds to the normalized meta-frequency π, yielding a high-pass character for the meta-filter.

In other words: A high peak in the center of the window corresponds to a strong amplification of high-meta-frequency content. Strong high-meta-frequency content, on the other hand, indicates a strong oscillation of the Fourier coefficients along the frequency axis. This holds especially for the meta-spectral rate $f_g = f_s/2$, which corresponds to the center of the window. A strong oscillation at this Nyquist rate yields in alternating phases for neighboring Fourier coefficients. The wrapped group delay $D(n,k)$ and the AAGD approach π (180 degrees) in this case.

1) Measuring the high-pass property: A simple method to measure the meta-high-pass property of a window is the maximum-average ratio $\mathcal{P}$ of the window:

$$\mathcal{P} := \frac{\max_{n=1}^{N} w[n]}{\sum_{n=1}^{N} w[n]/N} \tag{10}$$

For some important windows, Table I gives the maximum-average ratio $\mathcal{P}$ and the AAGD $\hat{\mathcal{D}}$ for pink noise and for a piano signal.

2) Digital Null: A special issue occurs if an STFT coefficient has zero magnitude, because in this case the phase is arbitrary. This problem mainly occurs in signals that consist of zeros. We can avoid this problem by adding some white noise — much below the quantization level — to the signal. This introduces only very small signal changes, but practically ensures that the STFT does not contain zeros.

C. Phase coincidences at attack slopes

The key to understand the behavior during attack slopes is *the envelope of the windowed signal*, not the window function itself. In the steady-state case we can assume that the envelope of the unwindowed signal does not have major changes during the window length. The window function introduces an envelope with a maximum at the center, leading to the high-pass property as described in Section III-B and to a higher probability for the absolute group delay $|D(n,k)|$ to approach π.

During attack slopes, the resulting envelope is influenced *both* by the window function and the attack envelope. The maximum amplitude is not at the center anymore. For that reason, the meta-frequency high-pass effect does not occur, so the AAGD has *not* the tendency towards π as it has in the steady-state case. For this reason, we can employ $\hat{\mathcal{D}}(n)$ as transient detector when we choose an appropriate window function.

The choice of the window function is a classical trade-off: The steeper the bell curve, the stronger is the high-pass effect, the steeper the envelope of the signal must be to count as transient. Hence, steep windows like the squared triangle window are more robust against noise, but less sensitive against transients than e.g. the Hann window.

IV. Modifications

The AAGD as proposed in Section II has the disadvantage that it contains a rather high level of noise. When we apply this measure for onset detection, we must additionally consider that the onset characteristics of different instruments take place in different frequency ranges [4]. For that reason, following postprocessing steps are suitable to improve the measure:

- For denoising, the absolute wrapped group delay $|D(n,k)|$ in AAGD calculation is replaced with the output $\tilde{D}(n,k)$ of a maximum-order filter with the filter order m:

$$\tilde{D}(n,k) = \max_{i\in\mathbb{Z}, -\frac{m}{2}<i<\frac{m}{2}} |D(n,k+1)| \tag{11}$$

  An order of 5 ($i \in \{-2,-1,0,1,2\}$) has empirically shown to be a good choice.
- It is possible to adjust the algorithm to a certain frequency range or a combination of frequency ranges by restricting the calculation of the average to the desired range. For onset detection, the optimal range depends on the instrument. A good range for our piano example is $0 \cdots \frac{f_g}{8}$ (see Section V-A).
- Some instruments (e.g. cembalo) make additional noise when a note event *finishes*. This noise leads to phase coincidences also at the end of the note event. We can suppress this effect by adding more noise to the signal to detect; the noise level should be higher than the note-end noise, but of course lower than the onset.

V. Practical Examples

We show the effectiveness of this transient detection method on two examples, the piano, and, in more details, the castanet onset detection. As shown on the latter case, one onset leads to *two* peaks on the AAGD.

A. Detection of Piano Onsets

In Fig. 3, the beginning of Beethoven's sonata op. 90 is depicted. We chose an STFT frame size of 2048 samples and a hop size between adjacent frames of 512 samples at a sampling rate of 48 kHz. The recording is the left channel from Track 39 of the EBU-SQAM collection [6]. In all graphs except the one notated as "full range", the frequency range was limited to $[0; \frac{f_g}{8}]$.

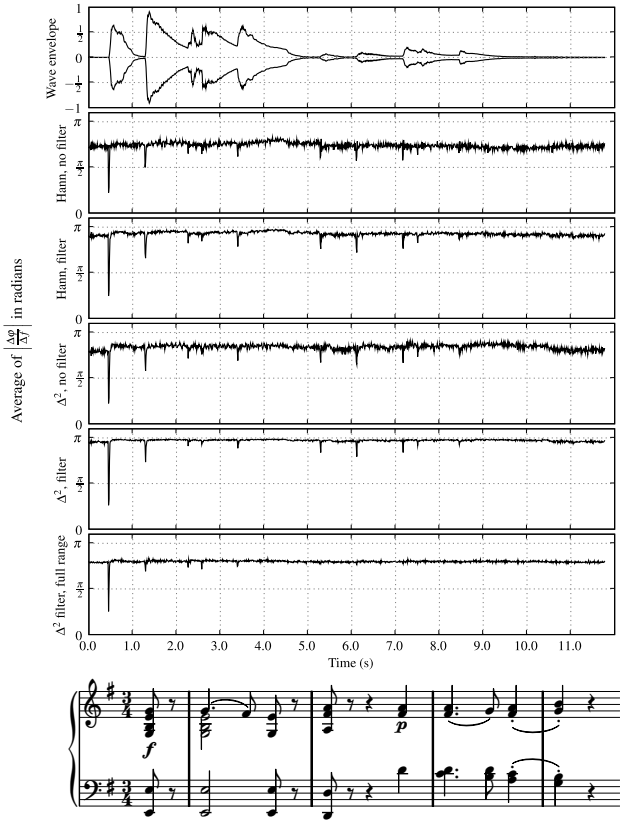

Fig. 3. AAGD for squared triangle and Hann window, with and without maximum order filter (filter order=5) for a piano signal (score below diagrams). In all AAGD plots except "full range", the frequency range is restricted from 0 to 3000 Hz ($f_g/8$). The full range plot has been generated using 3-oder maximum filtering. See Section V-A for details.

The example illustrates the effects of different window functions and postprocessing steps. The AAGD is higher for the squared triangle window (higher peak-to-average ratio) than for the Hann window. Additionally, it is obvious that employing a maximum order-filter additionally rises the AAGD and reduces the variance. These results are confirmed by Table I. However, we run into a trade-off since our goal is a high separability between transients and steady-state parts. The higher the variance is, the more likely transients are detected, but the higher is the difficulty to get the detected transients out of the AAGD deviation noise. A broader evaluation which helps to answer this question is presented in Section V-C.

The importance of the chosen frequency range is illustrated by the comparison of the lowest two graphs. They are generated with a squared triangle window and filter, but the upper one works with a restricted frequency range ($[0; \frac{f_g}{8}]$, the lower one with the full range $[0; f_g]$. The softer onsets are not detectable in the full-range graph.

B. A Single Castanet Onset In Detail

Fig. 4 presents a castanet clap in detail. The recording is from Track 27 of the EBU-SQAM collection [6]. The STFT frame size is also 2048 samples with 48 kHz sampling frequency, but the hop size is set to one sample. The left part of the figure illustrates the waveform envelope, the middle part shows $\hat{\mathcal{D}}(n)$, where n goes downside. The most important points are marked with the numbers 1–4. For all points, the windowed waveform of the signal around the given point is illustrated on the right side of the figure.

As can be noticed, the onset leads to *two* drops in the AAGD (Markers 1 and 3), where the first one becomes even lower than the

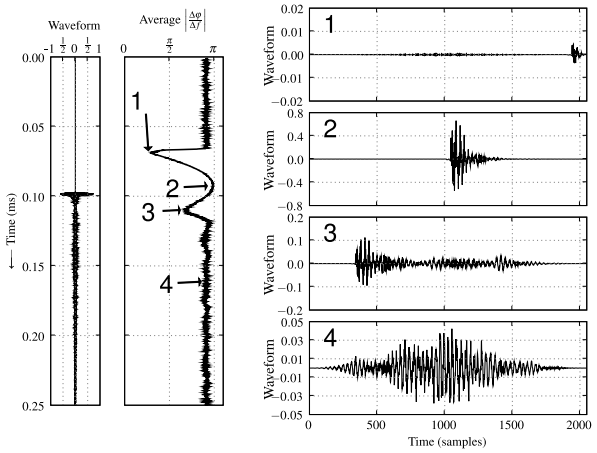

Fig. 4. AAGD of a castanet clap (left) and corresponding waveform at four characteristic points (right) using the Δ^2 window. See Section V-B for details.

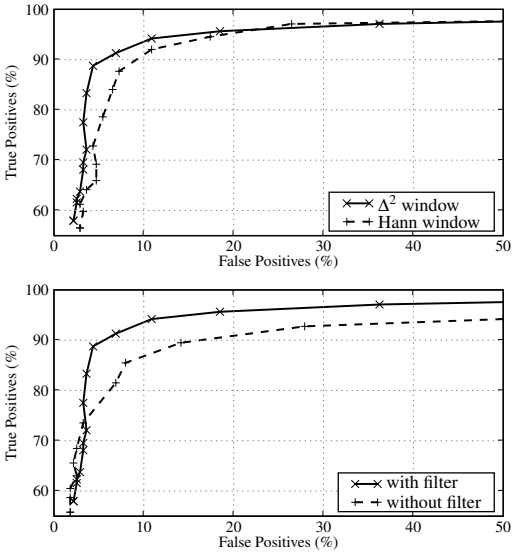

Fig. 5. Transient detection results in percentages of detected transients vs false detections. The default configuration (solid lines) is: Squared rectangle window, maximum filter (order=5), full range.

second one. Between them (Marker 2), the AAGD becomes *higher* than the average. This can be easily explained with the assumption that the AAGD corresponds to the energy concentration in the center of the window. For Marker 2, this is obvious, but it explains even the asymmetry between Marker 1 and 3. Since for most percussive signals the onset time is lower than the release time, the asymmetry for energy concentration is greater at onset time, because an audio signal has usually lower energy before the onset than after. Marker 4 shows an average steady-state example. The window function leads to an energy concentration in the center of the window. However, the concentration is weaker as within the onset (Marker 2).

The comparison between Marker 2 and 4 shows the basic difference between our approach and the group-delay-based calculation of the center of gravity [4]: Our algorithm does not only take the position (=center of gravity) into account, but also interprets the spread of the windowed signal over the time axis. Therefore Marker 2 denotes a higher peak than Marker 4.

C. Evaluation

In order to demonstrate that the proposed algorithm actually works as onset detection, we have annotated manually a subset of percussive and pitched-percussive instruments from the EBU-SQAM collection. The subset contains 276 onsets in 43 files. The transient detection is based on a threshold θ; each frame yielding an $\hat{\mathcal{D}}(n)$ value below θ is considered as transient. We call a detected transient *true positive* if its position is 0 to 50 ms *before* a hand-labeled onset, *false positive* otherwise. The tolerance is the same as in [1], the movement to before the onset is due to the asymmetry of the detection function described in Section V-B. To determine θ, we calculated the mean μ and the standard deviation σ of the $\hat{\mathcal{D}}(n)$ values. The threshold θ is now given by

$$\theta = \mu - \lambda\sigma, \tag{12}$$

where the parameter λ denotes how many standard deviations an AAGD must be off the mean value μ to be detected as transient. The lower λ, the more sensitive and the less robust is the transient detector. As FFT window size we have chosen 2048 samples. To mask the noise at the end of note onsets, white, uniformly distributed noise with a peak level of -34 dB was added. In contrast to Section V-A, we computed the $\hat{\mathcal{D}}(n)$ values over the whole frequency range due to the variety of investigated instruments.

The results are presented in Fig. 5. To record one curve, we varied λ between 1 and 4 to get different operating points. The optimal result is the upper left point of the graph, i.e. 100% true positives at 0% false positives. We can see that the transient detection principally works. The best result (88.8% true vs 4.3% false positives) we got with the squared triangle window and maximum-order filtering with a filter order of 5. Both steps lead to significant improvements.

VI. Conclusions

In this paper, we have demonstrated that the *wrapped* discrete group delay has a distribution which depends on the STFT window function and on the steady-state/transient behavior of the signal. Bell-curved window functions concentrate the AAGD around π in the case of steady-state signals, but not for transients. This phenomenon can be exploited by the creation of a threshold-based transient detector. The detection results depend on the chosen window and a proper denoising-filter method: The squared triangle in combination with a maximum-order filter delivered good results.

References

[1] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M.B. Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(9):1035–1047, September 2005.

[2] P. Duhambel, B. Piron, and J.M. Etcheto. On Computing the Inverse DFT. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(2):285–286, February 1988.

[3] A. Holzapfel and Y. Stylianou. Beat tracking using group delay based onset detection. In *Proc. Int. Conf. on Music Information Retrieval ISMIR '08*, pages 653–658, 2008.

[4] A. Röbel. Transient detection and preservation in the phase vocoder. In *Proc. International Computer Music Conference* , pages 247–250, 2003.

[5] R. Smits and B. Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *IEEE Transactions on Speech and Audio Processing*, 3(5):325–333, September 1995.

[6] European Broadcasting Union. Sound Quality Assessment Material, Tech 3253, 1988.

[7] U. Zölzer. *DAFX — Digital Audio Effects*. John Wiley & Sons, New York, NY, USA, 2002.