On the Use of Phase and Energy for Musical Onset Detection in the Complex Domain

Juan P. Bello, Chris Duxbury, Mike Davies, and Mark Sandler, *Senior Member, IEEE*

***Abstract*—We present a study on the combined use of energy and phase information for the detection of onsets in musical signals. The resulting method improves upon both energy-based and phase-based approaches. The detection function, generated from the analysis of the signal in the complex frequency domain is sharp at the position of onsets and smooth everywhere else. Results on a database of recordings show high detection rates for low rates of errors. The approach is more robust than its predecessors both theoretically and practically.**

***Index Terms*—Attack transients, audio, complex domain, energy, music, onset detection, phase.**

I. Introduction

TEMPORAL segmentation of audio into note events is useful for a range of audio analysis, editing and synthesis applications. Examples may include automatic transcription, content analysis and nonlinear time-scale modification and pitch-shifting. The segmentation task is especially difficult for complex mixtures including both percussive and nonpercussive onsets. In musical signals, let us assume notes to be events defined by the temporal concatenation of an attack transient: a short and unpredictable segment characterized by fast changes in the intensity, pitch or timbre of the sound [1]; followed by the steady-state of the signal, where it is stationary, thus easily predictable. An onset can be defined as the instant when the attack transient begins, thus marking the beginning of the note. Typically, note onset detection schemes use energy-based approaches, often involving frequency weighting [2]. In recent years, this has been extended to include subband schemes such as [3], [4].

Here, we depart from the basic theory of energy-based onset detection, extending our analysis to include phase information, then combining both methods in a complex domain approach that improves experimental results while providing a more robust theoretical framework.

II. Energy-Based Onset Detection

Usually, the introduction of a new note leads to an increase in the energy of the signal. In the case of strong percussive note attacks, such as drums, this increase in energy will be very sharp. For this reason, energy has proved to be a useful, straightforward, and efficient metric by which to detect percussive transients, and therefore certain types of note onset. The local energy of a frame of the signal $s(m)$ is defined as

$$E(m) = \sum_{n=(m-1)h}^{mh} |s(n)|^2 \tag{1}$$

where h is the hop size, m the hop number and n is the summation variable. Taking the first difference of $E(m)$ produces a detection function from which peaks may be picked to find onset locations. This is one of the simplest approaches to note onset detection. The idea can be extended to consider frames of an FFT.

Let us consider a time-domain signal $s(mh)$, whose STFT is given by

$$S_k(m) = \sum_{n=-\infty}^{\infty} s(n)w(mh-n)e^{-j2\pi nk/N} \tag{2}$$

where $k = 0, 1, \ldots, N-1$ is the frequency bin index and $w(n)$ is a finite-length sliding window. It follows that the magnitude difference between consecutive FFT frames is then

$$\delta S = \sum_{k=1}^{N} |S_k(m)| - |S_k(m-1)|. \tag{3}$$

This measure, known as the spectral difference, can be used to build an effective onset detection function (an implementation based on this can be found in [5]). Energy-based algorithms are fast and easy to implement, however their effectiveness decreases when dealing with nonpercussive signals and when transient energy overlaps in complex mixtures. Energy bursts related to transient information are more noticeable at higher frequencies as the "tonal" energy is usually concentrated at lower frequencies, masking the effect of these variations on the signal content.

III. Phase-Based Onset Detection

The Short-Time Fourier Transform of the signal, $S_k(m)$, can also be defined in terms of a group of sinusoidal oscillators with time-varying amplitudes $|S_k(m)|$ and phases $\varphi_k(m)$ (with unwrapped phases denoted as $\tilde{\varphi}_k(m)$). During the steady-state part of the signal these oscillators will tend to have constant

Manuscript received August 8, 2003; revised October 20, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiang-Gen Xia.

The authors are with the Department of Electronic Engineering, Queen Mary University of London, London E1 4NS, U.K. (e-mail: juan.bello-correa@elec.qmul.ac.uk; christopher.duxbury@elec.qmul.ac.uk; mike.davies@elec.qmul.ac.uk; mark.sandler@elec.qmul.ac.uk).

Digital Object Identifier 10.1109/LSP.2004.827951

1070-9908/04$20.00 © 2004 IEEE

frequencies. Therefore, the difference between two consecutive unwrapped phase values must remain constant between frames. For the k^{th} oscillator[1] that is

$$\Delta\tilde{\varphi}(m) = \tilde{\varphi}_e(m) - \tilde{\varphi}(m-1) = \tilde{\varphi}(m-1) - \tilde{\varphi}(m-2) \quad (4)$$

where $\tilde{\varphi}_e(m)$ is the estimated unwrapped phase for the current frame. Ideally, a target phase can be defined as

$$\tilde{\varphi}_t(m) = \tilde{\varphi}(m-1) + \Omega_k h \quad (5)$$

where Ω_k is the frequency of the k^{th} sinusoid. However, only synthesized sounds in artificially-controlled conditions behave like this. For real sounds we expect a deviation phase to be added to the target in order to generate the estimated unwrapped phase. This deviation can be calculated as

$$\tilde{\varphi}_d(m) = \text{princarg}[\tilde{\varphi}(m) - \tilde{\varphi}_t(m)] \quad (6)$$

where the function princarg maps the phase to the $[-\pi, \pi]$ range. By calculating the estimated unwrapped phase as

$$\tilde{\varphi}_e(m) = \tilde{\varphi}_t(m) + \tilde{\varphi}_d(m) \quad (7)$$

and substituting (5), (6), and (7) in (4), we can obtain values for the difference of unwrapped phase values. Collecting all terms into the right-hand side of (4), we can obtain an expression for the phase deviation between target and the real phase values in a given frame

$$d_\varphi = \text{princarg}[\tilde{\varphi}(m) - 2\tilde{\varphi}(m-1) + \tilde{\varphi}(m-2)]. \quad (8)$$

d_φ will tend to zero if the current phase value is close to the estimated value and will deviate significantly from zero otherwise. The latter is the case for most oscillators during attack transients.

Let us extend this analysis to the distribution of phase deviations for all oscillators within one analysis frame. Let us call $f_m(d_{\varphi,k})$ the probability density function of our data set on a particular frame m. During the steady-state part of the signal most values are expected to be concentrated around zero, creating a sharp distribution. On the other hand, during attack transients, the corresponding distributions will be flat and wide. In [6], these observations are quantified by calculating the interquartile range and the kurtosis coefficient of the distribution. Here, we propose measuring the frame-by-frame spread of the distribution as

$$\eta_p(m) = \text{mean}(f_m(|d_{\varphi,k}|)). \quad (9)$$

Measuring $\eta_p(m)$ is a fast and reliable approach to generating a detection function. Phase-based onset detection offers an alternative to common energy-based methods, overcoming detection constraints for soft onsets. However, the method is susceptible to phase distortion and to the variations introduced by the phase of noisy components (usually related to low-energy values which are therefore ignored).

IV. Detection of Onsets in the Complex Domain

There are a number of reasons that justify combining phase and energy information for onset detection: while energy-based approaches favor strong percussive onsets, phase-based approaches emphasize soft, "tonal" onsets; the two methods are more reliable at opposite ends of the frequency axis; the information they gather behaves in a similar statistical manner. A first attempt to combine these approaches was presented in [7]. Then, measures of spread for both distributions were simply multiplied, compensating for instabilities in either approach and producing sharper peaks for detected onsets. However, this analysis does not imply a fully combined approach where energy and phase information is simultaneously analyzed. This can only be achieved in the complex domain as will be explained in the following.

For locally steady state regions in audio signals, we can assume that frequency and amplitude values remain approximately constant. In the Sections II and III it has been shown that by inspecting changes in either frequency and amplitude, onset transients can be located. However, by predicting values in the complex domain, the effect of both variables can be considered. Let us assume that, in its polar form, the target value for the k^{th} bin of the STFT is given by

$$\hat{S}_k(m) = \hat{R}_k(m)e^{j\hat{\phi}_k(m)} \quad (10)$$

where the target amplitude $\hat{R}_k(m)$ corresponds to the magnitude of the previous frame $|S_k(m-1)|$, and the target phase $\hat{\phi}_k(m)$ can be calculated as the sum of the previous phase and the phase difference between preceding frames

$$\hat{\phi}_k(m) = \text{princarg}[2\tilde{\varphi}_k(m-1) - \tilde{\varphi}_k(m-2)]. \quad (11)$$

We may then consider the measured value in the complex domain from the STFT

$$S_k(m) = R_k(m)e^{\phi_k(m)} \quad (12)$$

where R_k and ϕ_k are the magnitude and phase of the current STFT frame. By measuring the Euclidean distance between target and current vectors in the complex space, as shown in Fig. 1(a), we can then quantify the stationarity for the k^{th} bin as

$$\Gamma_k(m) = \left\{ \left[\Re(\hat{S}_k(m)) - \Re(S_k(m))\right]^2 + \left[\Im(\hat{S}_k(m)) - \Im(S_k(m))\right]^2 \right\}^{1/2}. \quad (13)$$

Summing these stationarity measures across all k, we can construct a frame-by-frame detection function as

$$\eta(m) = \sum_{k=1}^{K} \Gamma_k(m). \quad (14)$$

Equation (14) can be simplified by mapping $\hat{S}_k(m)$ onto the real axis (forcing $\hat{\phi}_k(m) = 0$), such that

$$\hat{S}_k(m) = \hat{R}_k(m) = R_k(m-1). \quad (15)$$

This implies rotating the phasors, as shown in Fig. 1(b), so that $S_k(m)$ can be represented using the phase deviation (8)

$$S_k(m) = R_k(m)e^{jd_{\varphi k}(m)}. \quad (16)$$

Let us consider the difference between this complex domain prediction approach and the basic amplitude difference measure in (3). This can be rewritten as

$$\delta S_k(m) = \hat{R}_k(m) - R_k(m). \quad (17)$$

[1]For clarity, the subindex k is not included in the equations of Section III.

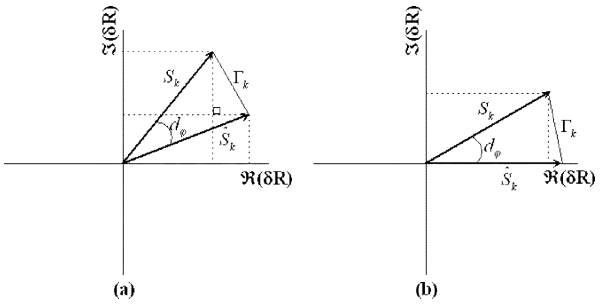

Fig. 1. Phasor diagram in the complex domain showing the phase deviation between target and current vector, and the Euclidean distance between them: (a) normal diagram and (b) rotated diagram.

With the mapping onto the real axis of $\hat{S}_k(m)$, (13) becomes

$$\begin{aligned}\Gamma_k(m) &= \left\{\left[\hat{R}_k(m) - \Re(S_k(m))\right]^2 + \Im(S_k(m))^2\right\}^{1/2} \\ &= \left\{\left[\hat{R}_k(m) - R_k(m)\cos(d_{\varphi k}(m))\right]^2 \right. \\ &\qquad \left. + \left[R_k(m)\sin(d_{\varphi k}(m))\right]^2\right\}^{1/2} \\ &= \left\{\hat{R}_k^2(m) - 2\hat{R}_k(m)R_k(m)\cos(d_{\varphi k}(m)) \right. \\ &\qquad \left. + R_k^2(m)\sin^2(d_{\varphi k}(m))R_k^2(m)\cos^2(d_{\varphi k}(m))\right\}^{1/2} \\ &= \left\{\hat{R}_k(m)^2 + R_k(m)^2 \right. \\ &\qquad \left. -2\hat{R}_k(m)R_k(m)\cos(d_{\varphi k}(m))\right\}^{1/2}. \end{aligned} \tag{18}$$

For the case of $d_{\varphi k}(m) = 0$

$$\begin{aligned}\Gamma_k &= \left\{\hat{R}_k^2(m) + R_k^2(m) - 2\hat{R}_k R_k\right\}^{1/2} \\ &= \hat{R}_k(m) - R_k(m). \end{aligned} \tag{19}$$

Therefore $\Gamma_k(m)$ is only equal to $\delta S_k(m)$ when $d_{\varphi k}(m)$ is equal to zero, e.g., when the phase prediction is "good." In that case, only the energy difference is being taken into account. In the case of $d_{\varphi k}(m) \neq 0$, the phase deviation from the prediction is taken into account. $\eta(m)$ constitutes an adequate detection function showing sharp peaks at points of low stationarity. Fig. 2 depicts the detection function for a section of a guitar signal. The figure also gives examples of phase and amplitude used individually. The complex domain approach is clearly less noisy, therefore simplifying the task of peak-picking and allowing a more robust detection.

V. Quantitative Analysis

A. Peak-Picking

To enhance the selection of peaks in the detection functions, the median filter is used to obtain an adaptive threshold curve $\delta_t(m)$. This is calculated as the weighted median of an H-length section of the detection function around the corresponding frame, such that

$$\delta_t(m) = \delta + \lambda \,\text{median}\, \eta(k_m), k_m \in \left[m - \frac{H}{2}, m + \frac{H}{2}\right]. \tag{20}$$

δ and λ are constant values, however while the latter is only a scaling factor, variations of the former largely affect the good and false detections ratio. Reference [8] demonstrated the effectiveness of the median filter for the thresholding of peaks in detection functions generated from music.

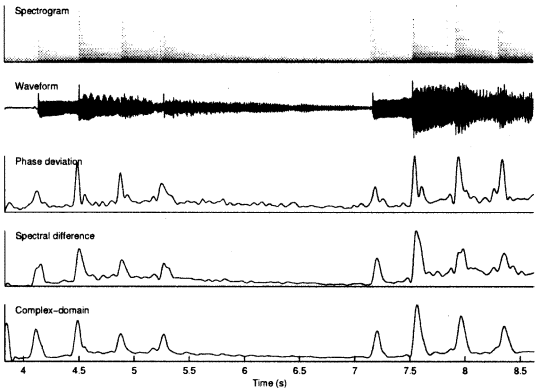

Fig. 2. Spectrogram of a music signal (upper) and onset detection functions using phase-based (upper-middle), energy-based (lower-middle) and the proposed complex-domain (bottom) approaches.

B. Onset Results

Experimental results compare the three presented approaches to onset detection: the spread of the distributions of spectral differences and phase deviations, and the complex-domain approach. The spread measure is used as it was found to be the most efficient and effective as shown in [7]. The experiments were performed on a 1065-onsets database of hand-labeled music segments.

For all methods, Fig. 3 displays the percentage of good detections versus the percentage of false positives for $\lambda = 1$ and different δ values (scaled by -10^2). Better performance shifts the curve up and leftwards. The complex-domain approach outperforms the other two methods. Its curve's optimal point[2] $(\{90.2, 5.0\}_{\delta=5.65})$ is above those of the spectral difference $(\{83.0, 4.1\}_{\delta=8.21})$ and the phase deviation $(\{81.8, 5.6\}_{\delta=4.27})$. It reaches the highest values of correct detections and is only outperformed by the energy method at the bottom of the plot. For low detection rates the spectral difference outperforms the phase deviation, that presents high rates of false positives.

Table I shows results according to onset types: pitched nonpercussive (e.g., bowed strings), pitched percussive (e.g., piano), nonpitched percussive (e.g., drums) and complex mixtures (e.g., pop music). Results are obtained by using the optimal δ value (scaled by -10^2) for each method in each case (from the corresponding performance curve). Results support that, for most cases, combining energy and phase information outperforms either approach alone. The only exceptions being firstly, the PNP case where the phase method performs best,

[2]The point representing the fewest errors for a given δ by being closest to 100% correct detections and 0% false positives

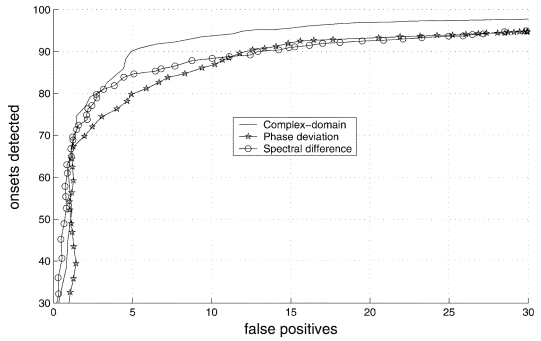

Fig. 3. Percentage of onset detections versus percentage of false positives for different values of δ (the peak-picking threshold): using complex-domain, phase deviation, and spectral difference detection methods.

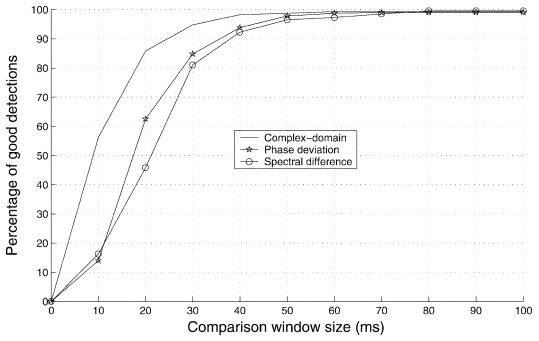

Fig. 4. Percentage of good detections for different lengths of the comparison window (ms): using complex-domain, phase-based and spectral difference detection methods. δ values correspond to the PP case in Table I.

TABLE I
ONSET DETECTION RESULTS FOR: PITCHED NON-PERCUSSIVE (PNP), PITCHED PERCUSSIVE(PP), NON-PITCHED PERCUSSIVE (NPP) SOUNDS AND COMPLEX MIXTURES (CMIX). COLUMNS SHOW THE CORRESPONDING δ VALUES (SCALED BY -10^2), AND THE PERCENTAGES OF CORRECT DETECTIONS (OK) AND FALSE POSITIVES (FP) PER METHOD

	PNP			PP		
METHOD	δ	**OK**	**FP**	δ	**OK**	**FP**
Spectral difference	4.58	87.1	8.6	9.62	94.9	1.6
Phase-based	0.06	95.7	4.3	6.62	95.5	0.3
Complex-domain	4.58	92.5	8.8	5.95	98.8	2.6
	NPP			CMIX		
METHOD	δ	**OK**	**FP**	δ	**OK**	**FP**
Spectral difference	1.50	81.6	5.5	7.02	81.2	10.7
Phase-based	1.16	80.7	5.5	2.33	80.1	24.7
Complex-domain	0.34	94.3	5.6	5.79	84.1	9.3

as quantifying only tonal changes is best for music with soft onsets; and secondly, the PP case where the phase-based algorithm returns less false positives than the complex-domain (at a higher total error rate). Spectral difference and phase-based methods are prone to under and over-detections (due mostly to amplitude modulations and overlapping for the first, and to phase distortion and frequency modulations for the second) especially when dealing with complex mixtures.

Fig. 4 shows how the complex-domain approach also provides better time localization for onsets. It shows percentages of good detections for different comparison windows (between target and detected events) on a database of acoustic recordings of MIDI-generated piano music (thus minimizing the error introduced by hand-labeling). The optimal δ values for pitched-percussive music were used. It supports quantitatively the argument made (Section IV) regarding the sharpness of the different detection functions. These results demonstrate that the theoretical robustness of the complex-domain approach implies also a practical advantage over the other methods.

Finally, Table II analyzes computational expense of the algorithms. Computational cost is calculated using the quantity of FLOPS (floating point operations) per frame (averaged across different executions), where the values have been normalized such that the computational cost of performing the FFT alone is set to 1. For all three methods, the increase in computation over the basic FFT algorithm is small enough to allow real-time implementation. The most expensive algorithm, the complex-domain method, is only 1.32 times the cost of the phase-based algorithm, the fastest of them all.

TABLE II
COMPUTATIONAL COST PER FRAME FOR EACH METHOD NORMALIZED TO THE LOAD OF THE FFT ALGORITHM

METHOD	**COMPUTATIONAL COST**
Spectral difference	1.90
Phase-based	1.80
Complex-domain	2.38

VI. CONCLUSIONS

Energy-based onset detection schemes perform well for pitched and nonpitched music with significant percussive content. On the other hand, phase-based onset detection approaches provide better results for strongly pitched signals (even for "softer" onsets), while being less robust to distortions in the frequency content and to noise. In the complex domain, both phase and amplitude information work together, offering a generally more robust onset detection scheme. Therefore, the presented theory for the complex domain approach to onset detection is not just an evolution of the spectral difference and the phase-based approaches, but a more general framework, in which the others are particular cases. From the practical point of view, it is straightforward to implement while remaining computationally cheap. Additionally, it proves effective for a large range of audio signals, as experimental results corroborate.

REFERENCES

[1] D. Moelants and C. Rampazzo, "KANSEI—The technology of emotion," in *A Computer System for the Automatic Detection of Perceptual Onsets in a Musical Signal*. Genova, Italy: AIMI-DIST, 1997, pp. 141–146.

[2] P. Masri, "Computer Modeling of Sound for Transformation and Synthesis of Musical Signals," Ph.D. dissertation, Univ. Bristol, Bristol, U.K., 1996.

[3] X. Rodet and F. Jaillet, "Detection and modeling of fast attack transients," in *Proc. Int. Computer Music Conf.*, 2001.

[4] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. IEEE Conf. Acoustics, Speech and Signal Proceesing (ICASSP,'99)*, 1999.

[5] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proc. 5th Int. Conf. Digital Audio Effects (DAFX,'02)*, Hamburg, Germany, 2002.

[6] J. P. Bello and M. Sandler, "Phase-based note onset detection for music signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP-03)*, 2003.

[7] C. Duxbury, J. P. Bello, M. Davies, and M. Sandler, "A combined phase and amplitude based approach to onset detection for audio segmentation," in *Proc. 4th Eur. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS-03)*, London, U.K., 2003.

[8] I. Kauppinen, "Methods for detecting impulsive noise in speech and audio signals," in *Proc. DSP-2002*, July 2002.