Sound Onset Localization and Partial Tracking in Gaussian White Noise

Ludger Selbach and Rolf Wöhmann
Arbeitsgruppe Verteilte Systeme
Technische Informatik VI
TU Halle-Wittenberg
21071 Hamburg, Germany
{ludger, rolf}@tit8.tu-harburg.de

Abstract
In auditory scene analysis the detection of common partial onsets is an important clue for event formation. Our wavelet based approach for onset localization does not suffer from either smearing effects caused by windowing in the time domain or from consistency problems caused by multiband level thresholding. Furthermore the method makes use of the beneficial effect of partial removal by adaptive signal cancelling. The results presented indicate that our approach is a promising basis for multimcomponent acoustic signal separation.

1 System Architecture
The fundamental signals considered in our approach are time limited sinusoidal functions disturbed by Gaussian white noise. Our model for these signals consists of two basic components, an onset part and a steady part. These parts correspond to two types of agents used in the system called the master module and the partial tracker module (PTM) shown in fig. 1 and in fig. 2, respectively. There is exactly one master module in the whole system, while the number of PTMs changes dynamically. The master’s task is to create or delete PTM’s triggered by the detection of an onset or an offset, respectively. The task of the PTM is to track the partials originating from the event. The tracked partial signals \( a(k) \) are fed back to the master and subtracted from the overall input signal \( x(k) \).

All filters used in our system are analytic constant-Q energy-normalized gammatone filters (Selbach et al., 1996). We chose the gammatone filter because of its physiological justification and computational efficiency.

The global architecture of the system resembles the one described in (Nakatani et al., 1995) in many respects. The realization of the system components, however, differs considerably.

1.1 The Master Module
The master module depicted in fig. 1 consists of a wavelet transformation module (WTM), an event detector and a partial tracker module controller.

The WTM is basically a bank of logarithmically spaced analytic constant-Q gammatone filters (Selbach et al., 1996). Its purpose is to preprocess the signal for the event detector and to provide suitable initial states for the tracking filters of the PTMs. As soon as the event detector detects an onset, the PTM controller initiates the proper PTMs with one-to-one copies of the wavelet filters residing closest to the frequencies where the partials are expected. By building the residual between the overall input signal and the tracked partials, the event detector \( \omega \) finds the following events in an almost undisturbed context.

In the current realization the master depends on the previous knowledge of a partial’s frequency localization at its onset and on the knowledge of an event’s status as an onset or offset. This information could be gained from musical notation as done by Schaefer (1996) for expressive music performance analysis. While the use of previous knowledge leads to a simplification in the current state of the implementation, the entry point will still be kept as a higher level knowledge feed after the future removal of this dependency.
Event Detection

Our approach for event detection is based on the fact that homogeneous signals cause a characteristic phase pattern in the time-frequency plane (Soltau et al., 1996). The homogeneity condition is approximately fulfilled in the neighborhoods of sound onsets.

To reduce the computing load for pattern detection a single line of constant phase in the phase pattern being close to the amplitude maximum of the impulse response is selected. Let p(k) denote the sample of the selected phase line in the frequency band with index k ∈ {0..N-1}, if a singularity occurs at n = 0. With \( u(n) \) being the wavelet coefficient in band k at sample n, consider the expression

\[
Y(n) = \sum_{k=0}^{N-1} u(n+k) \tag{1}
\]

Selecting the p(k) close to the maximum of the filter onset responses, causes a prominent peak in Y(n), since the phases sum up constructively along the phase line.

In the presence of noise Y(n) will exhibit spurious peaks which we wish to eliminate. In the following we present an algorithm for event localization in Gaussian white noise. As our wavelet filters are energy normalized we have for each scale of the wavelet transform due to the Wiener-Lee theorem

\[
E(\|u_n\|^2) = \sigma^2, \tag{2}
\]

where \( \sigma^2 \) is the variance of the white noise. If the noise is not only white but also Gaussian and if the real and imaginary filter responses are uncorrelated the modulus exhibits a Rayleigh distribution for which the expectation value is given by (Kummerer, 1992)

\[
E(\|u_n\|) = \frac{\sqrt{2}}{2} \sigma. \tag{3}
\]

Thus for Gaussian white noise the expectation value of the modulus is the same constant for each scale. We make use of this fact for noise floor estimation by fitting a straight line given by

\[
g(k) = a + b \cdot k \tag{4}
\]

to the wavelet transform modulus along the selected phase line via least squares linear regression.

As the ideal case b = 0 will hardly ever be observed, a tolerance must be given for b to deviate from this ideal. For the variance of the modulus \( \sigma_n^2 \) we have with eq.2 and 3

\[
\sigma_n^2 = E \left( \|u_n\| - E(\|u_n\|) \right)^2
\]

\[
= E(\|u_n\|^2) - E^2(\|u_n\|)
\]

\[
= \sigma^2 \left( 1 - \frac{\pi}{4} \right)
\]

\[
\approx \frac{\pi}{4}, \tag{5}
\]

this condition holds for analytic filters

where

\[
m_1 = \frac{1}{N} \sum_{k=0}^{N-1} |u(n+k)| \approx E(\|u_n\|). \tag{6}
\]

Thus if

\[
|\theta| \leq \theta_0 \cdot \sqrt{\frac{2}{4 - \pi}} \tag{7}
\]

with some constant \( \theta_0 > 0 \), we consider the signal as Gaussian white noise. In this case this condition holds, \( m_1 \) is used as threshold for event detection.

With threshold \( \theta_0 \) the complete algorithm for event detection is as follows:

1. Set \( Z(n) = Y(n) \)

2. If \( Y(n) < \) threshold set \( Z(n) = 0 \)

3. estimate \( \theta \) in eq.4

4. if \( \theta_0 \) holds set threshold \( m_1 = \theta \) and \( Z(n) = 0 \)

5. there is an event at sample \( n \) if the following conditions hold:

\[
\bullet Z(n) \text{ is a maximum}^2,
\]

\[
\bullet \text{there was no event in } [\theta - \nu..\theta + \nu].
\]

The constant \( \nu > 0 \) should be chosen such that spurious peaks in the neighborhood of an onset are ignored.

A problem with the above algorithm arises from the fact that \( \theta \) would be close to zero not only for white noise but also for all spectra being symmetric with respect to the midmost frequency band. This problem could be solved by calculating further line fits for shifted band indices and requiring the fulfillment of eq.3 for each of them.

1.2 The Partial Tracker Module

For each onset a PTM is created in which the adaptive gammatone tracking filter is instantiated as a one-to-one copy of the corresponding wavelet filter. The term "one-to-one" is to be taken literally. To avoid unnecessary tracking delay, not only the coefficients are copied but also the current state variables. Initial values for energy normalization and constant-Q property are maintained with each update of the filter coefficients.

Prediction and Frequency Estimation

A first order complex-valued least squares AR model is used to predict the next sample and the frequency localization of the tracked partial. With the estimates of the zeroth and first autocorrelation coefficients \( \phi_0 \) and \( \phi_1 \) the pole of the AR model is at \( z_0 = e^{\phi_1} \). The predicted next sample \( \hat{x}(k+1) \) is

\[
\hat{x}(k+1) = a_0 \cdot x(k) + a_1 \cdot \phi_1, \tag{8}
\]

where \( a_1 \) is the inverse of the transfer function modulus of the gammatone filter at \( f_0 \) given by

\[
a_0 = \frac{\lambda}{k(n-1)}, \tag{9}
\]

The frequency estimate is

\[
\hat{f}(k) = \frac{f_0}{2\pi} \arg(\phi_1). \tag{10}
\]

\( ^2 \)these resulting from setting \( Z(n+1) = 0 \) are ignored
where $f_s$ is the sampling rate. The window size employed for the estimation of the autocorrelation coefficients is continuously adapted to the time window size at on a samplewise basis.

The amplitude of a complex valued signal contaminated by white Gaussian noise is Rice-distributed [Kammeyer, 1992]. By the influence of noise the estimated amplitude is biased towards higher values. In the extreme case of zero amplitude the bias is given by eq.3, while for growing amplitudes it is continuously decreasing to zero. The Rice distribution is difficult to handle, due to the involvement of the modified Bessel function. In the current implementation there is no account for the amplitude bias introduced by the noise.

The Adaption Rule

As in (Wang, 1994) the center frequency $f_c(k)$ of the tracking filter is adapted by the gradient descent rule

$$f_c(k+1) = f_c(k) + \theta \left[ f_c(k) - f_c(k-\delta) \right],$$

(11)

where $f_c(k)$ is the estimated signal frequency, $\theta$ is the tracking gain and $\delta$ is the group delay of the tracking filter plus the one of the frequency estimator in samples.

It can be shown that the adaption given by eq.11 is stable if (Wang, 1994)

$$0 < \theta < 2 \sin \left( \frac{\pi}{1 + \delta} \right).$$

(12)

With $\delta = 0$ in eq.12 we arrive at $0 < \theta < 2$. If the group delay is increased, the upper bound for $\theta$ is diminished. For $\theta \rightarrow \infty$ stable tracking is impossible. Evidently minimal group delay of the tracking filter is desirable for maximum adaption gain. This is why we choose $\delta = 1$ for the order of the gammatone filter, which is (Solbach et al., 1996), was shown to yield minimal group delay.

As the parameters are adapted samplewise it is preferable not to evaluate a transcendental function for the calculation of $\delta$. So we have

$$\delta = \begin{cases} \frac{x}{x < \text{om}(z)} & \text{for } 0 < x < \frac{2}{\pi} \\ \infty & \text{otherwise} \end{cases},$$

it is safe to set

$$\delta = 2 \left( x \right) \infty.$$

(13)

2 Results

The parameters employed in our examples are $Q = 0.1$, $\eta = 1$ and $\theta = 20\text{ms}$. 12 wavelengths per octave are used.

In Gaussian white noise the onset and offset of a time limited sine of $1\text{kHz}$ could be localized with an error of less than $0.4$ ms, down to a signal to noise ratio (SNR) as low as $1\text{dB}$. The sampling rate was $f_s = 44.1\text{kHz}$.

In the more complex example shown in fig.3 and fig.4, a sound composed of 4 partials in $4\text{dB}$ Gaussian white noise was analyzed. The following table shows the partial parameters:

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>freq</th>
<th>SNR</th>
<th>onset</th>
<th>offset</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.16s</td>
<td>0.36s</td>
<td>600 Hz</td>
<td>3.7 dB</td>
<td>-13</td>
<td>+3</td>
</tr>
<tr>
<td>0.13s</td>
<td>0.25s</td>
<td>600 Hz</td>
<td>15.7 dB</td>
<td>-9</td>
<td>-6</td>
</tr>
<tr>
<td>0.15s</td>
<td>0.20s</td>
<td>400 Hz</td>
<td>3.3 dB</td>
<td>-6</td>
<td>-2</td>
</tr>
</tbody>
</table>

The last two columns show the deviation of the estimated time from the actual event instant in samples ($f_s = 22.059\text{kHz}$). In the upper part of fig.3, the residual histogram is displayed, in the lower part we see the corresponding $Z(t)$ and in fig.4 the estimated partial frequencies.

3 Conclusions and Future Work

An architecture for sound onset localisation and partial tracking in Gaussian white noise was presented. The results given are a snapshot of our research in acoustic multicomponent signal separation. They indicate that the onset localization and partial tracking properties of our method are promising in terms of robustness against the superposition of noise and concurrent partials. Obviously there is still much which left to be done. This is part of our agenda for forthcoming work.

- The pre-knowledge requirement for the frequency starting point should be removed.
- The system should be able to automatically discern onsets from offsets.
- The flexibility of the noise model should be improved.
- If the partials originating from one onset are harmonic, the precise partials should be tied together by coupled adaption. This will improve robustness against the disturbance of single partials.
- A partial tracker should be able to terminate itself if the partial fades gradually without offset.
- There should be an account for the noise-induced bias of the amplitude estimation.

References


Solbach & Wöhrmann

ICMC Proceedings 1996