Application of a data cleaning technique to LIGO data to be used in continuous signal searches

Paola Leaci[1], Pia Astone[2], Maria Alessandra Papa[1,3] and Sergio Frasca[2]
[1] Max-Planck-Institut für Gravitationsphysik, Albert-Einstein-Institut, 14476 Golm, Germany
[2]University of Rome "La Sapienza", 00185 Rome, Italy
[3]University of Wisconsin, Milwaukee, WI 53201

High frequency short events, due for instance to delta-like spurious disturbances, may affect the broad band noise level and thus produce a loss in the efficiency of detection of continuous (CW) signals [1]. We apply this technique to LIGO data and characterize its performance.

Technique of identification and subtraction of disturbances in the Time Domain

Disturbances in time domain data, such as for instance delta-like spurious signals, might affect the data in the frequency domain, enhancing more or less the noise level in the spectrum. This may result in a loss in the efficiency of detection of continuous signals. The procedure we use here, first proposed and implemented by the Rome Virgo group, identifies such events and removes them from the original data time series. In this method, the events are identified in a high-pass filtered series. They are then registered and subtracted in order to produce the cleaned data sets. We apply this cleaning procedure to part of the LIGO's fifth science run data and study what improvements it yields, resulting in a decrease of the spectral power.

Cleaning procedure:

- *Butterworth filter* (at 38 Hz). To minimize the contribution of the prominent low frequency component.
- *Tukey window*. The 0.1% of the data are effectively lost unless the data are interlaced.
- *High-pass filter* (at 100 Hz). The events are identified on this data.
- *Time Domain cleaning procedure*. After finding events, their amplitude is subtracted to the amplitude of the original time series.

How are the events identified?

In order to identify the events, we set a threshold on the critical ratio (CR) on the data produced in the third step above. The CR is defined as

$$\mathrm{CR} = \left| \frac{y(i) - \mu_A(i)}{\sigma_A(i)} \right|, \qquad (1)$$

where $y(i)$, $\mu_A(i)$ and $\sigma_A(i)$ are the data samples, the mean and the standard deviation, respectively. These last two quantities are estimated in an auto-regressive way, hence we have to use an 'adaptive' threshold, i.e. a threshold changing with time because the sensitivity of the detectors changes with time. The **threshold on the CR** is equal to 5, that is considered a reasonable veto. There are other parameters that play an important role in this step of identification of events. One of them is **the memory time for the autoregressive estimation**, that depends on the characteristics of the apparatus. We used here 20 s. The procedure also makes use of the concept of **dead-time** Θ, that is the minimum time between two events. This is set depending on the apparatus, the noise and the expected signal. Here we used 0.1 s.

Event Subtraction

We call EVENT an ensemble of samples (in the high-pass data) which has at least one sample with CR above threshold. An event may have more than a single sample when more samples are above threshold and their distance is smaller than the dead-time Θ. Once an event has been identified, we clean the data by subtracting it from the original data set. Note that the event data is not the same as the original data because the event data does not have the low frequency component of the original data. A few samples before and after the event are used to smoothly connect the original and cleaned data.

Here we used ten samples, that correspond roughly to $6\text{x}10^{-4}$ s, being the sampling time of $\sim 6\text{x}10^{-5}$ s. Samples in the dead-time are not included in the duration of the event.

Looking for events

Figure 1 (a) shows the original time data series $\mathrm{h(t)}$ (blue curve) and the time series (red curve) on which the events are identified.

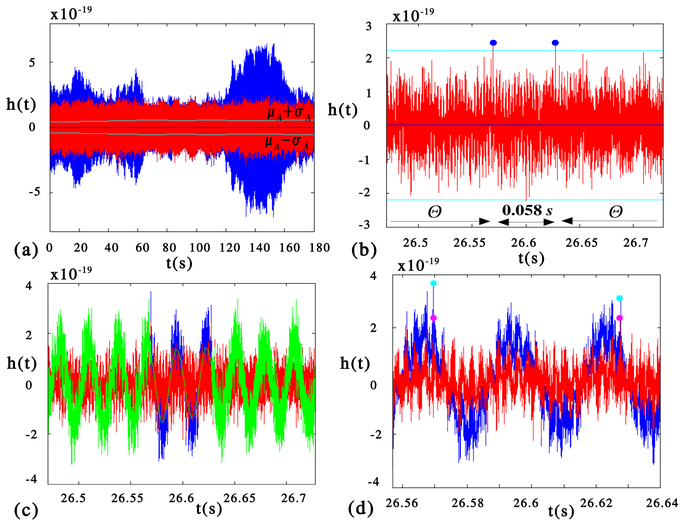

FIG. 1: Strain amplitude $\mathrm{h(t)}$ curves versus the time. The blue curve represents $\mathrm{h(t)}$ after a Butterworth high-pass filter, with a knee frequency of 38 Hz. The red curve consists of the further application of a high-pass filter, with a cut frequency of 100 Hz to identify "high frequency spikes". The 3 horizontal lines represent the auto-regressive estimation, $\mu_A \pm \sigma_A$ (cyan lines) and μ_A (blue line). The green curve corresponds to the cleaned data set and all the circles in the subplots label the beginning and the end of the considered event, lasting for about 0.058 s.

We show here an event that lasts 954 samples, corresponding to roughly 0.058 s in the S5 Hanford detector (H1) data. As we can see in Figure 1 (b), the event goes above threshold ($\mathrm{CR} > 5$) twice (blue circles). After the subtraction of the amplitudes of all the samples of the event, the cleaned time series is obtained and plotted in Figure 1 (c) (green curve). A zoom of the region where the subtraction is applied is shown in Figure 1 (d).

Illustrative results

We illustrate in Figure 2 ($\mathrm{b_1}$) the improvements that may be obtained when the data is particularly disturbed. We compute the power spectral density (PSD) of 1800 s of H1 data based on a running median algorithm, with and without cleaning. We can appreciate a decrease of the noise floor of almost an order of magnitude in the band around 200 Hz. As shown in Figure 3, the increase in the level of the noise floor is due to a population of loud disturbances in time. The gain in sensitivity (of order 3) more than counter-balances the effective loss in observation time for having effectively decreased

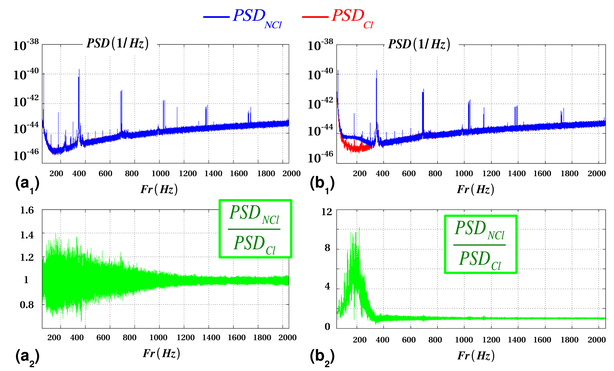

FIG. 2: Example of noise floor estimation. Panels ($\mathrm{a_1}$), ($\mathrm{b_1}$): the PSD of not cleaned data (blue curve) is compared to the PSD of the corresponding cleaned data (red curve). It is evident that the data analized in the panel ($\mathrm{b_1}$) is more disturbed. Therefore, the cleaning procedure is more effective in this case. In the bottom panels ($\mathrm{a_2}$), ($\mathrm{b_2}$), the ratio between the PSDs is also shown.

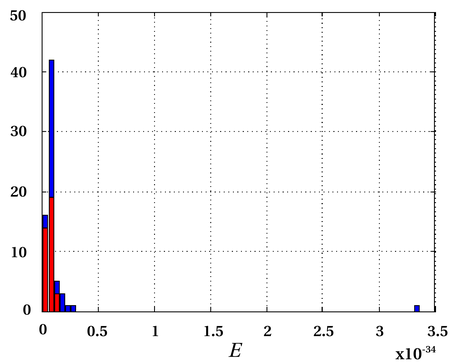

FIG. 3: Distribution of the event energies E of 36 (red bars) and 69 (blue bars) events found for the data whose PSDs are plotted in red in Figure 2 ($\mathrm{a_1}$) and ($\mathrm{b_1}$), respectively.

the observation time, ~ 1.1 s out of 1800 s.
In general, the cleaning procedure will not yield so dramatic improvements. We expect a few to several percent gain in sensitivity depending on the data quality flags previosuly used to select the data and on the instrinsic quality of the time domain data. On Virgo VSR1 data improvements of the order of 1.07/1.3 at high/low frequencies were observed analyzing roughly 10 days of data. We expect the improvements of this procedure on S5 and S6 LIGO data after the application of basic data quality flags, not to significantly differ from these.

References

[1] F. Acernese et al. *Class. Quantum Grav.*, page S491, v. 24, 2007.