Classification of glitch waveforms in gravitational wave detector characterization

S. Mukherjee1, R. Obaid, B. Matkarimov2
Center for Gravitational Wave Astronomy
University of Texas at Brownsville, Brownsville, USA

1 Corresponding author: soma@phys.utb.edu
2 Permanent address: Kazakh-British Technical University, Kazakhstan

PACS numbers: 95.55.Ym, 95.75.Wx, 95.75.Pq

Abstract

The paper describes a new fast and efficient algorithm to classify waveforms from signals that arise in gravitational wave detectors like LIGO. Such waveform classification is useful and important for detector characterization as well as for understanding glitches seen in the analysis pipelines that detect signals from astrophysical burst and compact object inspiral sources. Classification of glitches based on discrete parameters has been reported earlier by the author. In the current study, a new feature-mining approach has been developed that uses the additional information based on shape of the glitch waveforms. This has been a challenging problem because of the unique structures present in the real gravitational wave data. The paper presents results from simulations as well as real data. Studies show that the proposed method is fast and efficient in classifying glitches in noise.

PACS numbers: 95.55.Ym, 95.75.Wx, 95.75.Pq

1. Introduction

The three LIGO detectors at Hanford and Livingston [1] have been taking data in ‘science mode’ since July 2009. This is the sixth science run (S6) of LIGO [2]. Raw data from the gravitational wave (GW) interferometers can often be non-stationary, containing narrow band noise, as well as broadband glitches (also known in literature as transients or triggers) [3]. Presence of glitches can severely restrict the efficiency of astrophysical searches and hence we need to understand the origin of these glitches – instrumental, environmental or otherwise. The glitches themselves (often arriving at a high rate) may mimic gravitational waves. It is thus of utmost importance to develop data analysis methods that can mine the glitch database to extract information about the origin and properties of the glitches. There is a twofold advantage in this method. First, building up a knowledge base for the different types of observed glitches can help model the GW noise accurately and realistically for refinement of GW detection statistic. Secondly, tracing their instrumental origin can help experimentalists with debugging noise.

There are several pipelines operating online on LIGO data viz. kleine welle (KW) [4] and waveburst (WB) [5]. The KW pipeline, which works on the GW channel and several hundreds of auxiliary and environmental channels, picks up glitches at a fairly high rate. Unsupervised data mining methods like the multidimensional hierarchical classification analysis [6, 7] have been developed and LIGO science data have been analyzed in recent past. The aim of these studies has been to classify the population of glitches seen in the GW, environmental and auxiliary channels into statistically significant distinct groups with uniform characteristics. These methods use discrete properties of the glitches and produce groups in the higher dimensional parameter space, effectively reducing the dimensionality of the problem. However, in these analyses methods, we do not incorporate an important aspect of the glitches.
Glitches in GW data contain very rich information in terms of shape or glitch waveform. One can also make use of this information content in the time series data itself in classifying the glitches. This brings us to the aspect of development and application of methods in temporal classification [8]. Temporal classification methods e.g. S-means and Constraint Validation have been developed and studied on simulated GW data [9,10]. While these methods showed very promising results on simulated waveforms, they have not been tested on data from real GW detectors. Moreover, S-means has been built around the Dynamic Time Warping [11] which may not be optimal for heterogeneous datasets as seen in GW data from detectors like LIGO and Virgo [12]. Dynamic programming can also be slow quite often since it involves an element-by-element comparison of the time series. While his problem can be countered by developing index-based methods [10, 13-15], it is not totally free from problems because these methods may require using dynamic programming on the superset.

The current paper describes a new fast and efficient method for unsupervised classification of glitch waveforms. This method is based on distances calculated using Longest Common Subsequence (LCSS) [16]. The technique is studied in GW data for the first time. Results are demonstrated using, (i) noise-free simulated waveforms and (ii) waveforms embedded in Gaussian white noise. After benchmarking, the method has also been applied to KW database from LIGO’s fourth science run (S4).

2. Kleine Welle algorithm

Kleine Welle (KW) is an algorithm uses a dyadic Haar wavelet [17] transform to look for excess energy in time scale decompositions. The statistical outliers are identified by setting a threshold on the individual pixel energy. The significance of the nearby pixels that cluster on the time–frequency plane is calculated from known distributions.

3. Glitch waveform classification based on LCSS based algorithm

There are two main issues that one must address while developing an efficient feature-based classification method – (i) the data may have temporal gaps, i.e. there may be same pattern occurring at different time epochs and (ii) the classification preferably should be automatic (to keep up with the online feedback systems) and thus unsupervised classification methods robust to noise need to be explored.

Let us first take a look at why LCSS is efficient and how it fits into GW data analysis. In any unsupervised classification method, the first step is to calculate the distance between points in the parameter space. Even though Euclidean distance is the most commonly used method, it is not suitable for addressing the first difficulty mentioned above. Thus, two glitches that have the same waveform may show a high Euclidean distance if they are not occurring simultaneously. This will be considered a redundant cluster from the physical point of view. LCSS is able to compute a match between two time series by calculating distances that do not necessarily occur at the same time without having to rearrange the sample sequence.

Let \( X \) and \( Y \) be two waveforms with size \( p \) and \( q \) respectively.

\[
X = [(x_{1,1}, x_{j,1}), \ldots, (x_{1,p}, x_{j,q})] \quad (1)
\]

\[
Y = [(y_{1,1}, y_{j,1}), \ldots, (y_{1,p}, y_{j,q})] \quad (2)
\]

Let us define \( \Delta X = [(x_{1,1}, x_{j,1}), \ldots, (x_{1,p-1}, x_{j,q-1})] \quad (3) \)

Given an integer \( d \) and a number \( e (0<e<1) \), we define the LCSS\(_{d,e}(X,Y)\) as

\[
\text{LCSS}_{d,e}(X,Y) = 0 \text{ if } X \text{ or } Y \text{ is empty sequence} \quad (4i)
\]
The distances arrived at from the LCSS can be defined as follows –

\[
\Gamma_{d,e}(X, Y) = 1 - \frac{\text{LCSS}_{d,e}(X, Y)}{\min(p,q)} \quad (5)
\]

Once the distances are calculated according to equation (5), a generalized k-means [29] algorithm is employed to form the clusters. The reason as to the choice of k-means is dictated by the fact that this allows formation of homogeneous clusters that are insensitive to outliers. Generalized k-means [30] uses two parameters to start with – the number of clusters \(m\) and an initial seed \(s\). The algorithm works as follows.

Let \(T\) be a set of \(M\) waveforms. \(T = \{T_1, ..., T_M\}\). Let the number of clusters be \(m\). Let there be a set of \(m\) seeds denoted by \(S = \{s_1, ..., s_m\}\). Each seed \(s_j\) is composed of \(s\) time series representation of the waveforms, chosen randomly in \(T\). The choice of seed can be random in absence of a priori knowledge about the dataset, as is often the case with unsupervised clustering. The distance between the waveform \(T_i\) and the seed \(s_j\) is given by

\[
I(T_i, s_j) = 1/s \sum_{z \in G_j} \text{LCSS}(T_i, z). \quad (7)
\]

\(I(T_i, s_j)\) is usually called the inertia function. In the final step we calculate the intra-class variance

\[
\sigma^2(G) = \frac{1}{M} \sum_{j=1}^{k} \sum_{T_i \in G_j} I(T_i, s_j). \quad (8)
\]

The iterative process is continued until the value of \(\sigma^2(G)\) does not fall below a given threshold.

4. **Analysis Results**

We carry out the analysis in three steps. First we implement the algorithm on different types of simulated signals. In the second phase we analyze simulated signals of various forms embedded in noise and in the third and final phase we run the algorithm on LIGO’s fourth science run data (S4).

a. **Simulated waveforms without noise**

We first generate a data set of 350 simulated waveforms with variable parameters, each 1024 samples long. The waveforms are in the shapes of sine Gaussians, Gaussian pulses, Chirps and Mixture sinusoids with varying parameters, i.e. varying amplitude, frequency, width, location on the time axis etc. The
waveform database is introduced to the pipeline as the prime input. Figure 1 summarizes the output of the classification. The algorithm correctly classified the waveforms into the right categories. No misclassification was noted. This is an expected result in the absence of noise.

![Figure 1](image1.png)

Figure 1. This figure shows the result of classifying 350 simulated waveforms. Each subplot shows classes of waveforms as selected by the LCSS classification scheme. The input waveforms came in four different shapes - sine Gaussians (bottom right), Gaussian pulses (bottom left), Chirps (top right) and Mixture Sinusoids (top left) with varying parameters. The x-axis gives the time samples and the y-axis gives amplitude measured in arbitrary units.

b. Simulated waveforms embedded in noise

In the next step of this study, we generate a data set of 350 simulated waveforms of four different types of shape with variable parameters, each 1024 samples long, as described in the previous section. Each waveform data stream is mixed with Gaussian white noise. The output is thus a noisy waveform. This is fed to the classification pipeline. The results are summarized in Figure 2. In this case too, the LCSS algorithm could classify the waveforms with 100% accuracy. This robust-against-noise result is rather optimistic and points towards its aptness to be used in classification of glitches noted in GW data. It was interesting to observe that, when the same data was analyzed using standard k-means algorithm (without invoking the LCSS), approximately 23% of the glitches were misclassified.
c. Glitches in LIGO science data

Having gained insight with the simulations, we now apply the LCSS algorithm to classify glitches seen in LIGO science data. We have used glitches found in the KW glitch catalog during LIGO’s fourth science run (S4) [18]. The complete pipeline is explained in figure 3. This involves two main stages – (i) Extraction of the glitch from noisy output of the gravitational wave channel and (ii) Analysis of the extracted S4 glitch waveforms by the LCSS algorithm. Stage (i) involves the following sequence.

(i) Selection of glitches with signal-to-noise ratio (snr) greater than 10 from KW database;
(ii) Extraction of raw GW channel data, centered around the glitch (extracted time series noted by, say, \{q_i\})
(iii) Whitening \{q_i\} [19] and dynamically removing [20] the narrowband noise present in \{q_i\}. The resulting time series is denoted by \{cq_i\}.
(iv) Filtering \{cq_i\} with a bandwidth of 256 Hz around \(f_c\), where \(f_c\) is the central frequency of the glitch as noted in the KW catalog. The resulting time series is denoted by \{fcq_i\}.
(v) Heterodyning \{fcq_i\} and re-sampling the resulting data.

All the extracted glitches are normalized and stored as a time series of 1024 samples. This is introduced as the input to the LCSS pipeline for classification of the glitches.
In the data set used (an arbitrary day in S4), we extracted 150 glitch waveforms using the method described above. Figure 4 shows some examples of the typical glitch shapes noticed in the population under study. The examples amply demonstrate that there is a wide repertoire of glitch shapes in the data and thus merits classification. Figure 5 shows the end results of classification pipeline. Eight distinct and statistically significant classes of glitches were detected. It was found that the representative groups (classes) were consistent in picking up glitches of similar shapes or waveforms.

Figure 4. This figure shows examples of some of the typical types of glitch shapes seen in the population. The x-axis gives the time samples and the y-axis gives amplitude measured in arbitrary units.
Figure 5. The figure shows eight distinct and statistically significant classes of glitches as detected by the LCSS pipeline on a population of glitches seen in an arbitrary day during LIGO’s fourth science run. The x-axis gives the time samples and the y-axis gives amplitude measured in arbitrary units.

5. Discussion and future direction

Once the classes of glitches are determined, it becomes imperative to subject them to follow-up studies to determine their possible sources, whether instrumental or environmental. There exists a rich collection of various methods within the LIGO Scientific Collaboration (www.ligo.org) that helps in this direction. Methods such as Omega Scan [21], H-veto [22], detector-wide channel scan (DWCS) [23] or online KW analysis [24] can be very helpful in following up on the classified glitch groups. There also exists a large repository of data quality (DQ) flags (25-28). These can also serve as complementary information in conjunction with the follow-up methods mentioned above. The examples in this paper are shown on S4 data. Since then, a huge development has taken place in these areas of detector characterization and the current search tools and glitch characterization machineries can indeed bring out valuable information regarding explanation of observed glitches, thus providing a better handle to the detector noise. For example, one could take the possible set of correlated auxiliary and environmental channel glitches in the same time window (following the schemes given by say, H-veto or DWCS) and check the waveforms of the coincident (within the accuracy of the same time window) glitches for consistency. It would be an interesting study to see if a set of waveforms of similar shape could be correlated to a particular subset of auxiliary channels. One would thus develop a complete repository of glitch waveforms that could be used as a discriminator in detector characterization studies as well as in astrophysical search engines where either the waveform from a source could be known (coalescence of compact objects) or unknown (as in the case of burst sources). This is not done yet and can, in principle, open up a new window of understanding glitch sources in GW data. This is under construction by the authors. The plan is to run the complete analysis on the current S6 data.

To summarize, we have developed and tested a new fast and efficient method based on distances calculated using Longest Common Subsequence (LCSS) for unsupervised classification of glitch waveforms. The technique is studied in GW data for the first time. The method proves promising for classification of glitches in presence of noise because of its robust character. The method is now being applied on S6 data.
Acknowledgment

The authors gratefully acknowledge the LIGO Laboratory for use of LIGO S4 data. This work is supported by NSF grant #PHY0855371, 2009. The document is assigned LIGO Technical document number P1000057.

References

1. Abramovici A et al 1992 Science 256 325–33
2. Kawabe, K. 2008 J. Phys.: Conf. Ser. 120 032003
3. Mukherjee S 2003Class. Quantum Grav. 20 925–36
4. Chatterjee, S. et al. Classical and Quantum Gravity, 21, S1809-18, 2004
5. Klimenko, S et al 2004 Class. Quantum Grav. 21 S1685
6. Mukherjee S 2006 Class. Quantum Grav. 23 S661–71
7. Mukherjee S (for the LSC) 2007 Class. Quantum Grav. 24 S701
8. Schoenfeld, W. N. and Cumming, W. W. 1960 Proceedings of the National Academy of Sciences of the United States of America 46 753-758
9. Tang L R et al. 2008 Class. Quantum Grav. 25 184023
10. Lei, H et al 2007 International Workshop on Knowledge Discovery from Ubiquitous Data Streams, 18th European conference on Machine Learning (ECML) and 11th European conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), Warsaw, Poland. ECML, PKDD proceedings, Springer-Verlag (LNAI series)
11. Sakoe, H and Chiba, S 1978 IEEE Trans. Acoustics, Speech and Signal Processing, ASSP, 26, 43-49
12. Cavalier, F 1999 34th Rencontres de Moriond : Gravitational Waves and Experimental Gravity, Les Arcs, France
13. Agarwal, P. K. et al 2000 19th ACM symposium on Principles of database systems (PODS)
14. Pfoser, D 2000 VLDB, Cairo
15. Saltenis, S et al 2000 ACM SIGMOD
16. Vlachos, M 2002 ICDE 673-84
17. Mallat, S 1999 A Wavelet Tour of Signal Processing (Wavelet Analysis and its Applications) 2nd edn (New York: Academic)
18. Credico, A 2005 *LIGO technical document* G050292-00-Z

19. Mohanty, S 2002 *Class. Quantum Grav* **19** 1513

20. Weerathunga, T. S. 2010 *M.S. Thesis, University of Texas at Brownsville*

21. Sutton, P. et al 2009 *gr-qc* 0908.3665

22. Smith, J (for the LSC) 2008 *Class Quantum Grav* **26** 114013

23. Mukherjee, S 2007 *GWDAW12*, Boston, *LIGO technical document* G070814-00-Z

24. Christensen, N 2010 *Class. Quantum Grav* This issue

25. Gonzalez G et al 2007 *LIGO technical document* G070142-00-Z

26. Blackburn L et al 2006 *LIGO technical document* G-060407-00-Z

27. Chatterjee S 2007 *LIGO technical document* G-070138-00-Z

28. Blackburn L 2005 *LIGO technical document* G-050158-00-Z

29. Diday, E 1973 *International Journal of Computer Sciences* **2**

30. McQueen, J 1967 5th *Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA