LIGO-P070051-00-Z

Preliminary results from hierarchical glitch pipeline

Soma Mukherjee
on behalf of the LIGO Scientific Collaboration.
Center for Gravitational Wave Astronomy
University of Texas at Brownsville
Brownsville, Texas.

Abstract

This study reports on the preliminary results obtained from the hierarchical glitch classification pipeline. The pipeline that has been under construction for the past year is now complete and end-to-end tested. It is ready to generate analysis results on a daily basis. The details of the pipeline, classification algorithms employed and the results obtained on one day's analysis on the gravitational wave and several auxiliary and environmental channels from all three LIGO detectors is discussed.

Introduction

The three LIGO detectors at Hanford and Livingston [1] have started taking data in continuous science mode since November 2005. This is the fifth science run (S5) of LIGO [2] and it is expected to continue until the end of 2007. As in the previous science runs, data is analyzed in several sub-groups focusing on different aspects of the data. Four categories of astrophysical gravitational wave (gw) signals are searched. These are burst sources, continuous wave sources, stochastic background and inspiral sources. The analysis puts upper limits on various physical parameters associated with the specific sources [3-6]. Raw data from the interferometers are often non-stationary and contains both narrow band and broadband transients or glitches [7]. This poses two kinds of problems. First, the efficiency of the searches are depend on data quality and hence need to define data quality flags [8] to identify segments of data that are best suited to the search analyses. Second, the glitches themselves (often arriving at a high rate) may mimic gravitational waves and hence we need to define efficient vetoes to rule out such possibilities. We thus need to understand the source these glitches are coming from not only for the veto but also to provide feedback to the experimentalists. These activities are performed under the umbrella of the detector characterization and glitch analysis groups [9,10].

Brief review of glitch analysis in LIGO and motivation

The current effort in glitch analysis consists of several algorithms that record and study the glitches from many channels both statistically as well as tracking individual glitches. Kleine-Welle [12] algorithm is a wavelet based analysis that runs on a very large number of channels and records transients. Q-scan [11] is a facility that traces individual glitches seen in pipelines like kleine-Welle in the gw and several auxiliary channels belonging to many different sub-systems. Q-scan produces a set of time domain and time-frequency plots of conditioned data. Block Normal (BN) [13] event display also records loud glitches and looks at the time and time-frequency displays to gain insight into the glitches seen. Binary inspiral glitches [14] are also recorded and viewed in Q-scan. The glitch group members analyze the glitches seen on a daily

basis using all of these tools before drawing conclusions. Analysis is done on single interferometer glitches, as well as, on double and triple coincidences data. Results are stored in password protected web pages accessible to the collaboration.
In recent times, many new kinds of glitch types and events have been seen in the kleine-Welle and BN pipelines in the gw and auxiliary channels. While some of these could be tracked down, a large percentage (>70%) [15] are yet to be understood. This is one of the priorities of the glitch-working group.

Kleine-Welle pipeline picks up glitches at a fairly high rate – several thousands of glitches are seen per day in the gw channel. Some of the auxiliary channels e.g. magnetometer, accelerometer channels, are also found to glitch at a comparable rate. The glitch working group members study these glitches “by eye” and try to conclude about their nature and possible origin by checking if they show up in the BN event display or the Q-scan plots. However, given that it is very difficult if not impossible to study all the glitches manually and that this still leaves a major percentage of the glitches unexplained, more ways of looking into this problem is highly desired.

Following this need, a data mining approach to this problem has been pursued [16] viz. multidimensional hierarchical classification analysis. The aim of this study has been to divide the population of glitches seen in the gw and auxiliary channels into statistically significant distinct similar groups with relatively more uniform members. This effectively reduces the dimensionality of the problem in the sense that we could deal with a smaller number of entities (in this case, the groups) sharing similar characteristics. The problem is dealt in multi-parameter space because it has been seen in astrophysical scenario [17] that higher dimensional analysis reveals features that are not seen in one-dimensional (histograms) or two-dimensional (scatter plots) views. The reason is obvious, since higher dimensions allow simultaneous consideration of many more physical properties. Kleine-Welle trigger database being a multivariate data set, is a natural candidate for a multi-dimensional analysis. In addition to the readily available parameters like duration, central frequency and signal-to-noise ratio (snr) , one can also make use of the information content in the time series itself around the trigger central time i.e. the *shape* of the trigger [16].

Algorithm and pipeline

Several approaches exist in the statistical literature [17-20] for classification of different types of data sets. In the present analysis, we adopt the hierarchical classification scheme [21]. The algorithm is based on computation of metrics between the data points in the multi-dimensional space and using the variance minimization criterion to group them into statistically distinct classes [22]. The metric or the so-called ‘distance’ can be computed in many ways and the specific nature of the problem dictates which criterion to adopt. In our case, we adopt the Euclidean distance. The group-formation stage is guided by the criterion of ‘complete linkage’ i.e. largest distance between objects in the two groups. The choice is made based on pilot study and past experience of working with classification of discreet data sets [17]. The vindication as to how well the grouping fits the data structure under investigation is achieved by well known statistical tests [23-25].

The analysis starts with the kleine-Welle trigger databases with an interface to protected web page access. These databases come with the physical properties of the triggers that the algorithm has detected, viz. duration, central frequency and energy values that contain snr information. Typically, kleine-Welle algorithm picks up thousands of triggers, most of which have low snr values. A selection cut is applied to retain snr values above a certain threshold (snr=6 is the value chosen for this analysis). In order to extract the trigger shape information, 4 seconds of data around the trigger central time is chosen from the raw frame files of the respective channels. The data is pre-processed by first cleaning all the lines and then band-passing around the central frequency with a 16 Hz bandwidth. The data is base-banded [26] and re-sampled down to the given bandwidth. The resulting time series is retained as the best approximation to the trigger shape. This information, along with the three physical parameters mentioned above, go as input to the main classification code. The core classification code is in Matlab [27]. The pipeline uses the compiled version of this code and runs on a 16-node cluster. Figure 1 shows a schematic diagram of the pipeline. The output of the pipeline contains information about the different classes obtained. The individual class members are subjected to time-frequency analysis to look for the specific features typical of that particular class. The present version of the pipeline contains complete end-to-end tested modules up to this point. The next version of the pipeline will incorporate the last module (shown in blue in the figure) that processes the 'pattern recognition' part of the analysis, i.e. indexing well-defined class properties forming the 'basis' and auto-classifying the triggers into one of these classes.

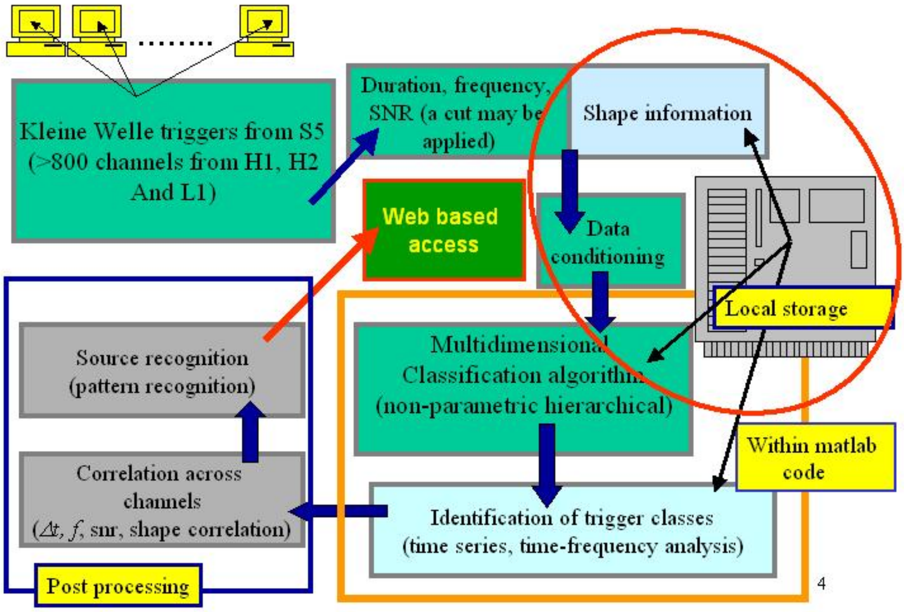

....
Kleine Welle triggers from S5 (>800 channels from H1, H2 And L1)
Duration, frequency, SNR (a cut may be applied)
Shape information
Web based access
Data conditioning
Local storage
Multidimensional Classification algorithm (non-parametric hierarchical)
Within matlab code
Source recognition (pattern recognition)
Correlation across channels (Δt, f, snr, shape correlation)
Identification of trigger classes (time series, time-frequency analysis)
Post processing
4

Figure 1. The figure shows a schematic representation of the pipeline. The analysis starts with the kleine-Welle trigger databases. The physical properties of the triggers viz. duration, central frequency and snr are considered. A selection cut is applied to retain snr values above a certain threshold. In order to extract the trigger shape information, 4 seconds of data around the trigger central time is chosen from the raw frame files of the respective channels that reside in a local storage device accessible to the pipeline. Data-conditioning is applied and a 16 Hz band around the central frequency is retained as the best approximation to the trigger shape. The output of the pipeline contains information about the different classes obtained. The present version of the pipeline contains complete end-to-end tested modules up to the orange box shown in the figure. The next version of the pipeline will incorporate the last module (shown in blue) that processes the 'pattern recognition' part of the analysis.

Analysis results

The results of the analysis are illustrated using one day's worth of LIGO S5 data. The date is February 17 2007. The corresponding GPS time range is 855763200-855849600 s. The channels looked at are DARM_ERR and several auxiliary and environmental channels. The latter is selected for demonstration from the list of auxiliary and environmental channels that are seen in the glitch analysis group to be affected most often.

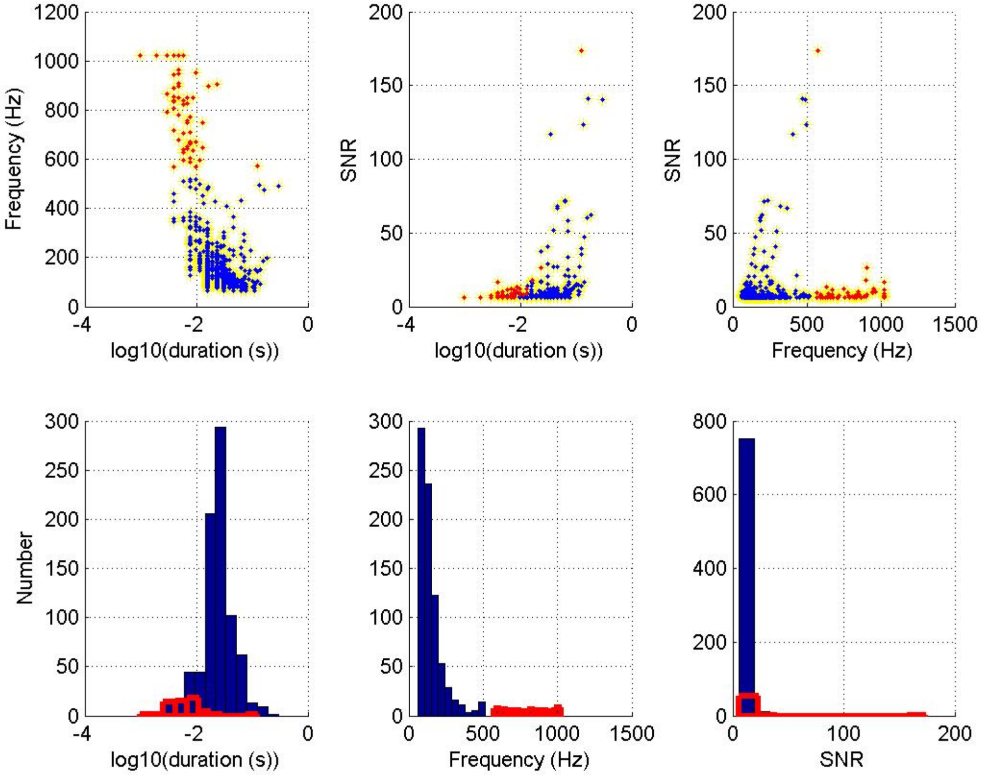


Figure 2. The figure shows the results from H1:LSC-DARM_ERR. On analysis with three parameters, duration, central frequency and snr, two statistically distinct clusters were found. The top panel shows the two-dimensional scatter plots between the three parameters considered and the bottom panel shows the histograms of the two groups thus found. Statistical tests showed that the group structure fitted the actual data with a correlation coefficient r=0.93 and significance p<0.003.

Figure 2 shows the results from H1:LSC_DARM_ERR. Using three physical parameters, two statistically distinct classes were found. The top row of figures shows scatter plots among the variables used and the bottom row shows the histograms of the groups thus found. The two groups are marked with different colours (blue and red) two show their respective positions in the one and two-dimensional snap-shots. Figures 3 and 4 show the same for H2:LSC-DARM_ERR and L1:LSC-DARM_ERR. It is interesting to note that, in all three cases, the two groups clearly separate on the frequency plane. However, the separation is not so sharply visible in the duration-snr plane in this 2-dimensional snap-shot.

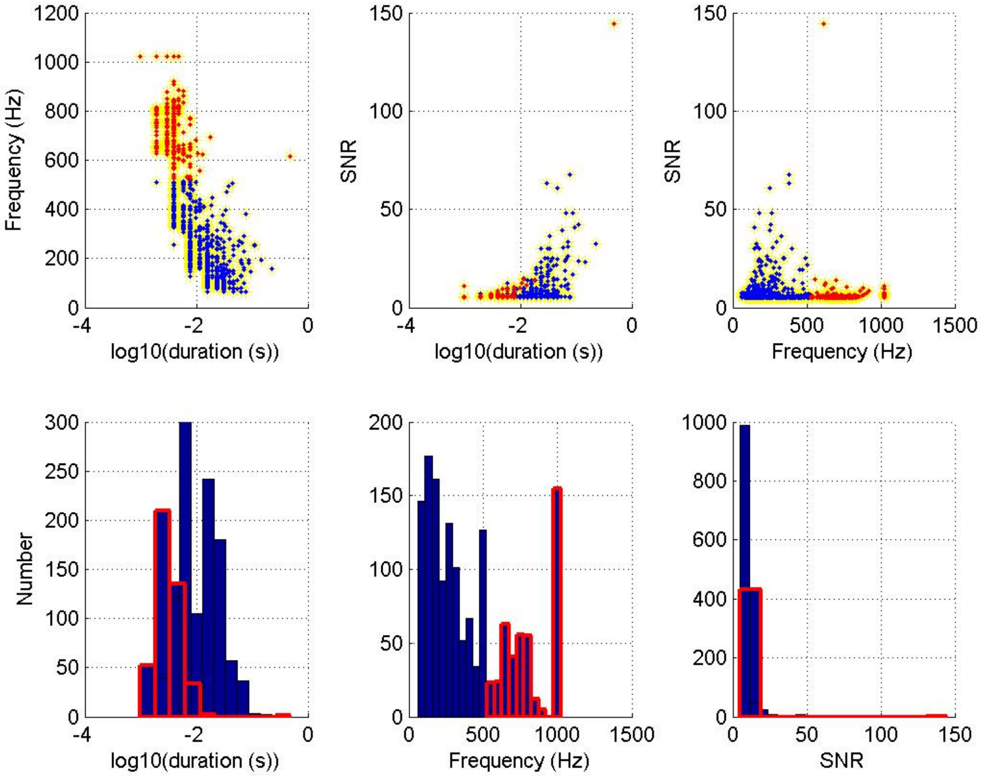


Figure 3. The figure shows the results from H2:LSC-DARM_ERR. On analysis with three parameters, duration, central frequency and snr, two statistically distinct clusters were found. The top panel shows the two-dimensional scatter plots between the three parameters considered and the bottom panel shows the histograms of the two groups thus found. Statistical tests showed that the group structure fitted the actual data with a correlation coefficient r=0.81 and significance p<0.004.

On analyzing the auxiliary channels, it was found that while some of the auxiliary channel triggers show uniformity (i.e. no significant classes are found), some others showed a mixture of different types of triggers i.e. existence of statistically significant classes. Figures 5-8 show some of these different cases. While the magnetometers in Hanford Observatory (HO) clearly show diversity of character in the trigger population (figures 5 a & b). Magnetometers in Livingston Observatory (LO) do not seem to show any such division (figures 6 a & b).

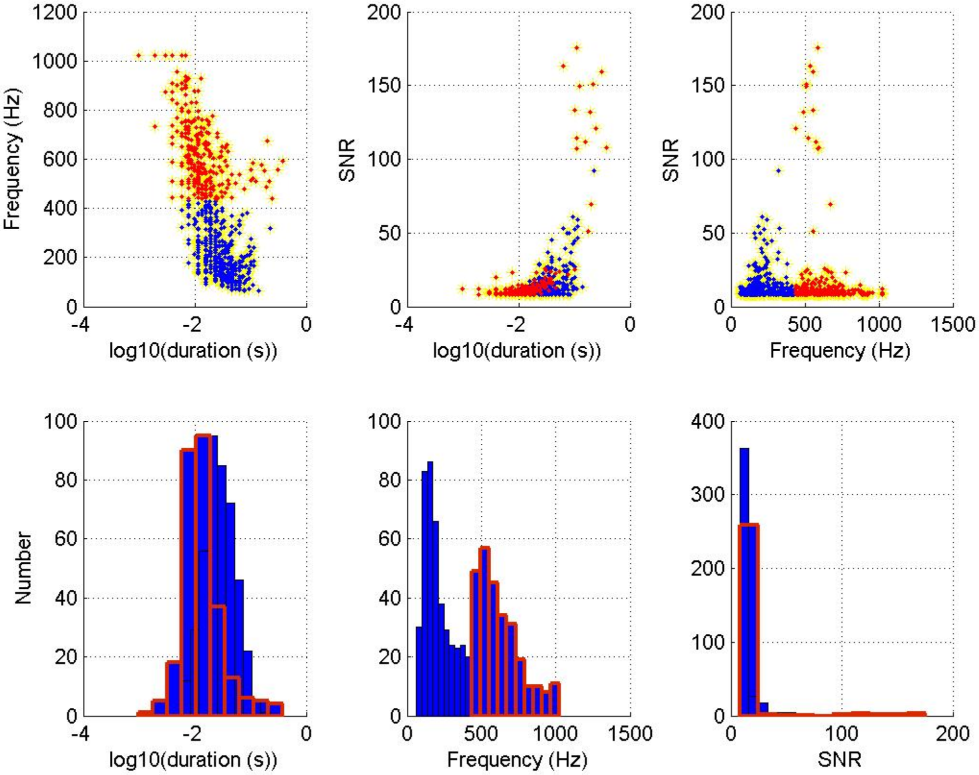
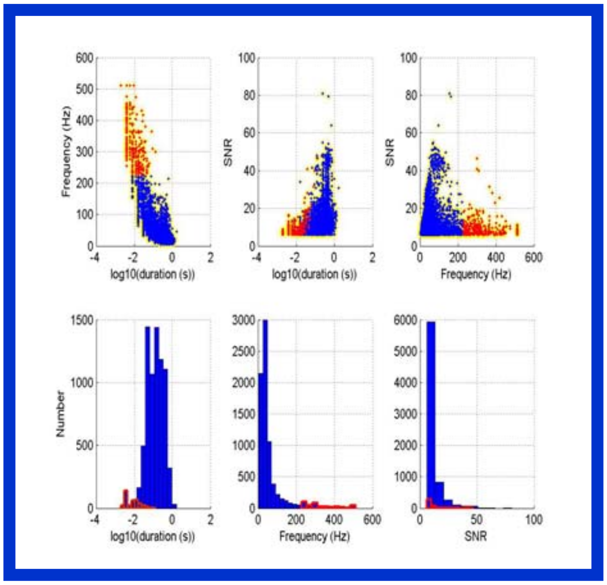


Figure 4. The figure shows the results from L1:LSC-DARM_ERR. On analysis with three parameters, duration, central frequency and snr, two statistically distinct clusters were found. The top panel shows the two-dimensional scatter plots between the three parameters considered and the bottom panel shows the histograms of the two groups thus found. Statistical tests showed that the group structure fitted the actual data with a correlation coefficient r=0.80 and significance p<0.04.



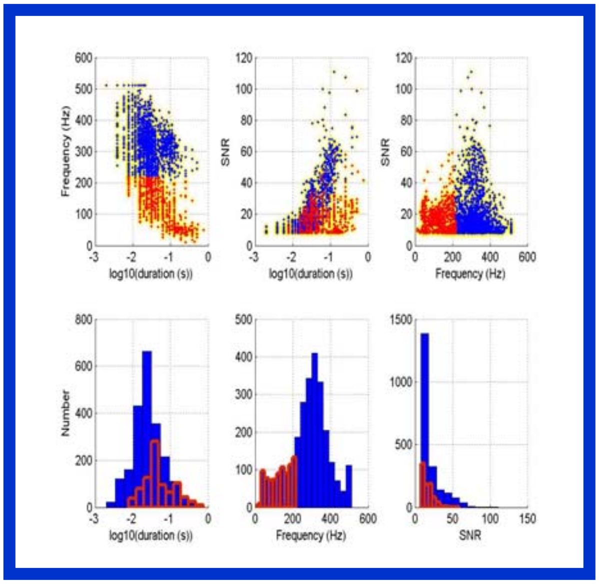


Figure 5. The left panel (5a) shows the classes as projected on to a two dimensional plane and also in histograms for the channel H0:ASC-BSC6_MAGZ. The right panel (5b) shows the same for channel H0:ASC-COIL_MAGZ. How well the present classification fits the given data is given by the usual statistics : r=0.92 (p<1e-6) and 0.70 (p<1e-8) respectively. The clear division in frequency is seen for these trigger populations too, even though the separation is not so sharp in the duration-snr plane .

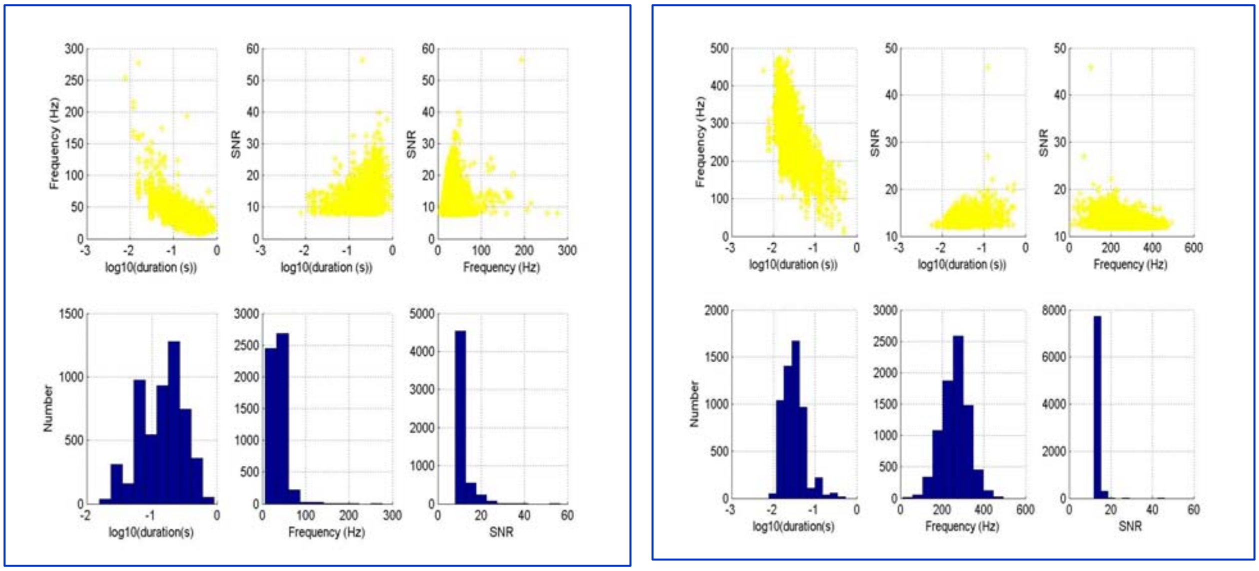
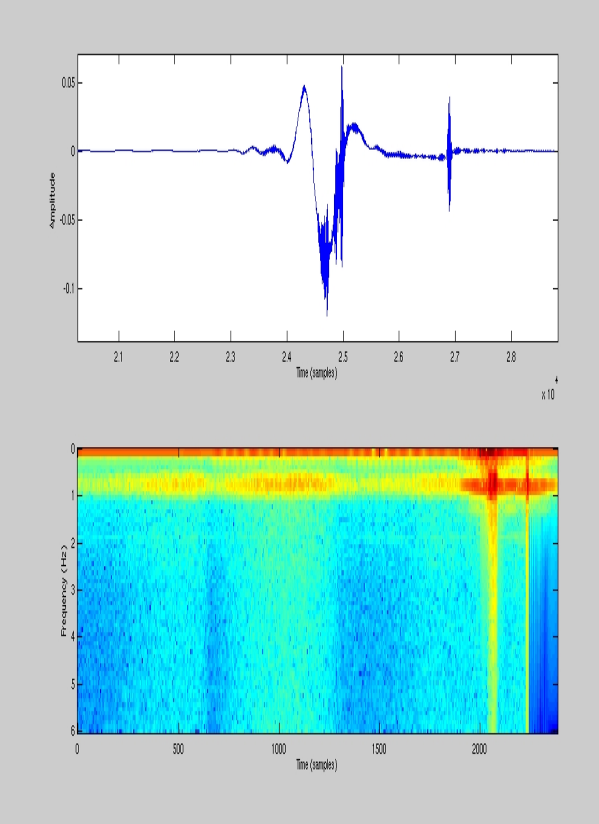
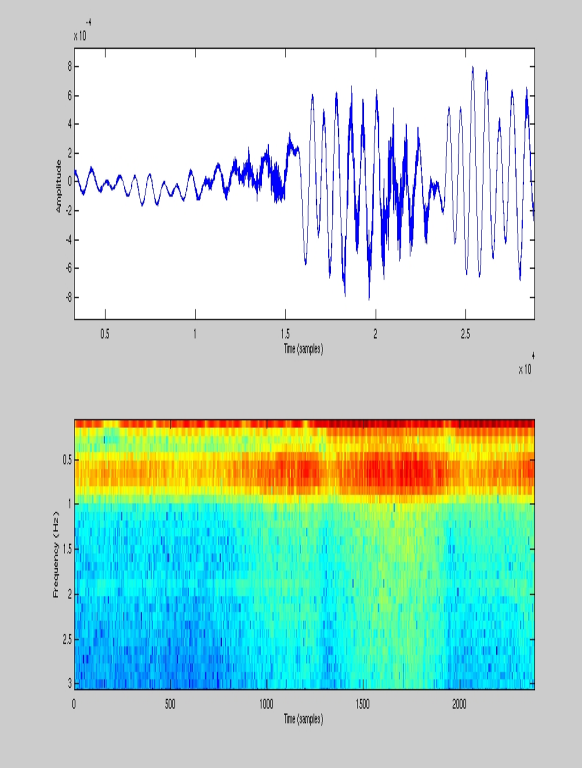
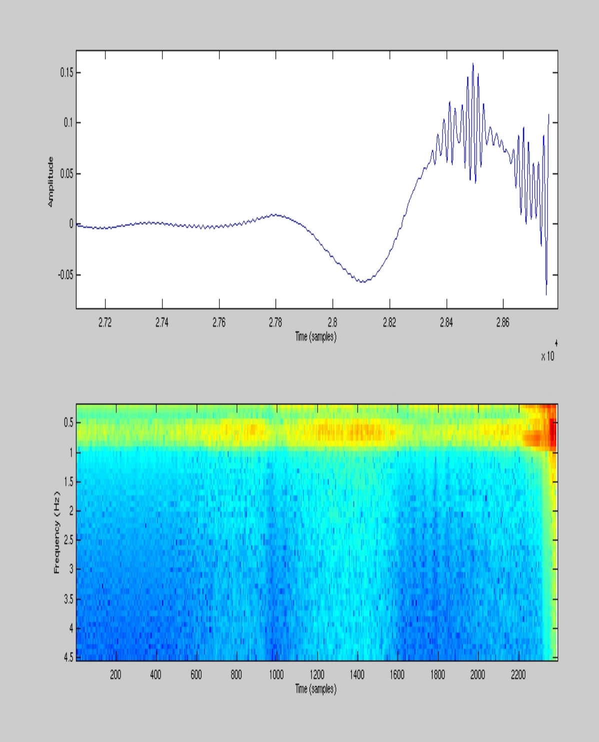


Figure 6. The left panel (6a) shows the classes as projected on to a two dimensional plane and also in histograms for the channel L0:ASC-LVEA_MAGZ. The right panel (6b) shows the same for channel L0:ASC-COIL_MAGZ. The apparent lack of clusters is also corroborated by the statistics : r= 0.79 (p>0.90) and r=0.64 (p> 0.24) respectively.

With the basic class information in hand, it is thus interesting to see what the members belonging to these different classes look like in the time-frequency plane and also in the time domain. The results are shown in figures 7-8. These figures are for H1:LSC-DARM_ERR, but similar picture is also seen in H2 and L1. Figures 7 a and b show two representative triggers from one of the groups that has been described by high frequency, mostly low duration and at the lower end of the snr distribution. The data around the trigger central frequency has been band-passed and all lines removed. The triggers in this group are also broadband, spanning the entire 12 Hz frequency band around the central frequency. Figures 8 a and b show two representative triggers from the groups that has been described by low frequency, mostly long duration and at the higher end of the snr distribution. The data around the trigger central frequency has been band-passed and all lines removed. The triggers in this group are narrowband, being restricted to a smaller bandwidth around the central frequency.







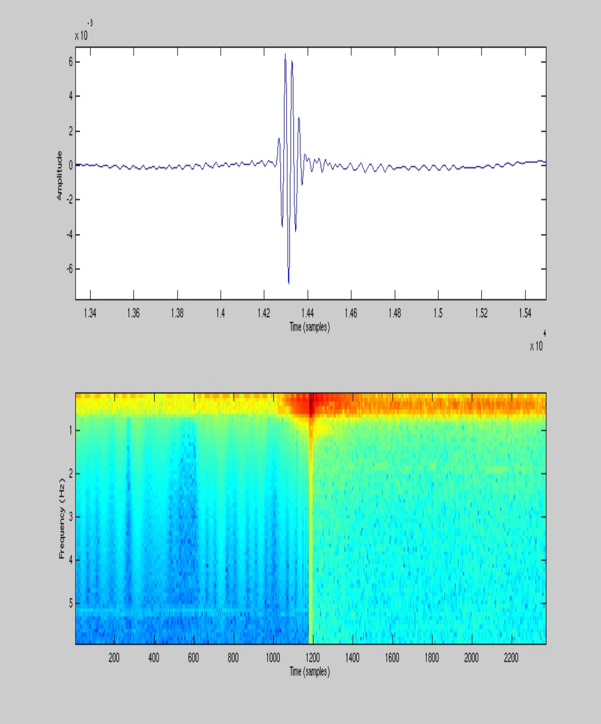
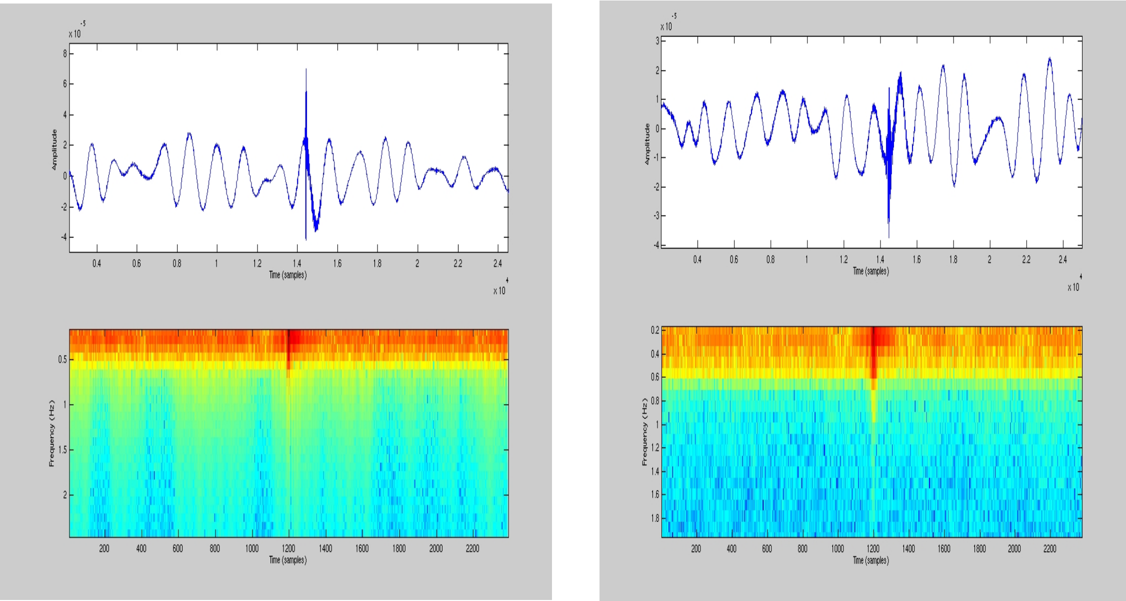


Figure7. Figures 7 a and b show two representative triggers from one of the groups in H1 that has been described by high frequency, mostly low duration and at the lower end of the snr distribution. The data around the trigger central frequency has been band-passed and all lines removed. The triggers in this group are also broadband, spanning the entire 12 Hz frequency band around the central frequency.



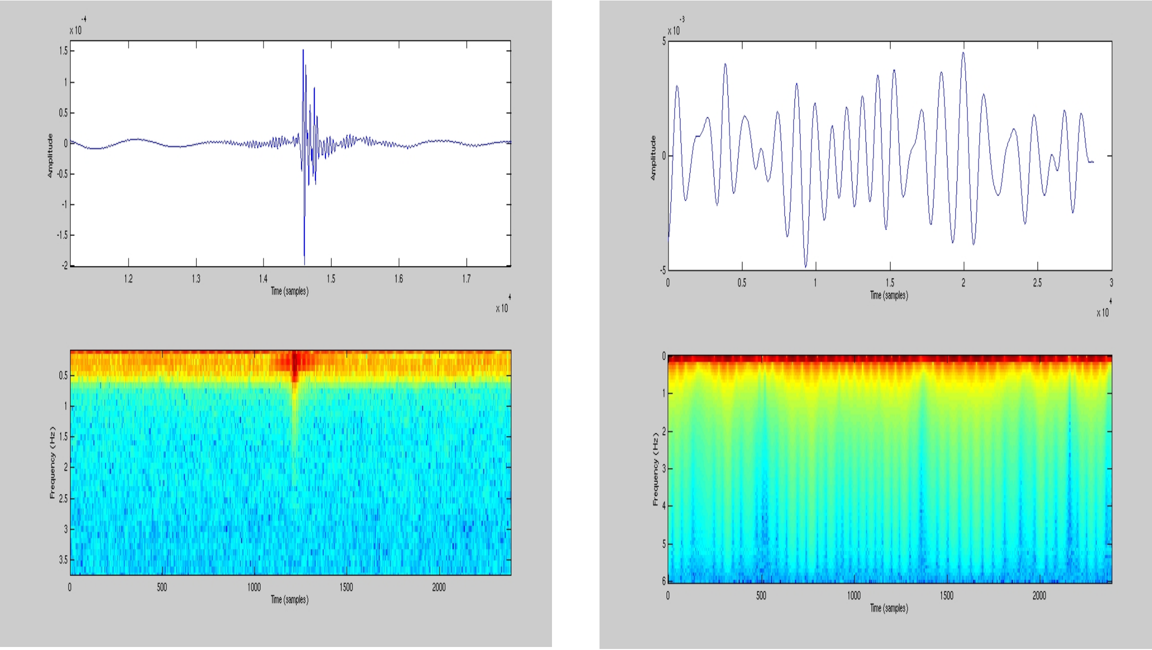


Figure8. Figures 8 a and b show two representative triggers from the groups in H1 that has been described by low frequency, mostly long duration and at the higher end of the snr distribution. The data around the trigger central frequency has been band-passed and all lines removed. The triggers in this group are narrowband, being restricted to a smaller bandwidth around the central frequency.

We now look at the time and time-frequency results from an auxiliary channel, H0:PEM-BSC6_MAGZ, one of the magnetometers from the Hanford detector. This is one of the channels that glitches at a fairly high rate often. As seen in figure 5, triggers from this channel showed presence of more than one class. These results are shown in figures 9-10. The figures represent time domain and time-frequency plots for triggers selected from the two different groups found in the analysis. In figures 9 a and b, we see low frequency (<200 Hz) and bandwidth limited (~10 Hz) triggers while in figure 10 a and b, we find high frequency magnetometer glitches that spread over more than 32 Hz bandwidth. The other auxiliary channels with more than one class, can also be seen on the time-frequency plane in a similar fashion.

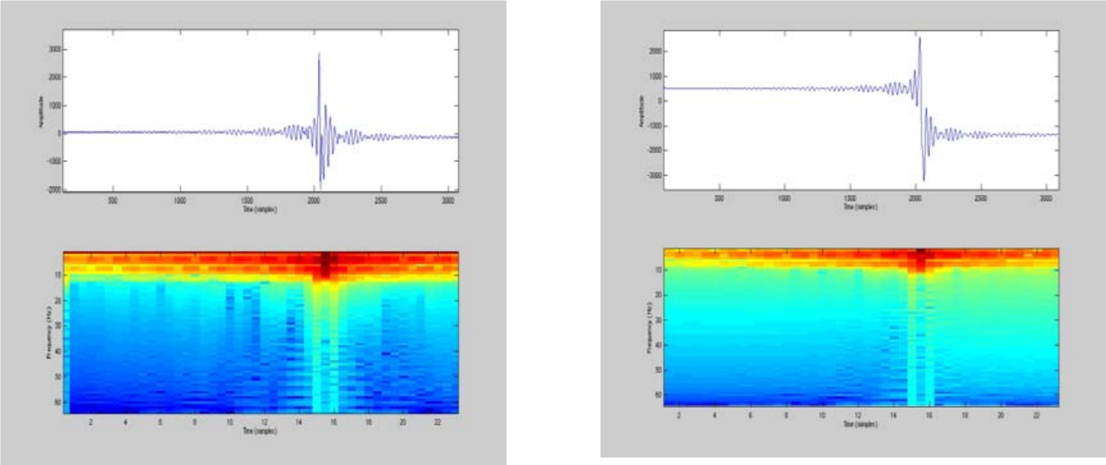


Figure 9. Figure 9 a and b show two triggers selected from group 1 of the channel H0:PEM-BSC6_MAGZ. These triggers are characterized by low frequencies and a relatively low bandwidth (~10 Hz).

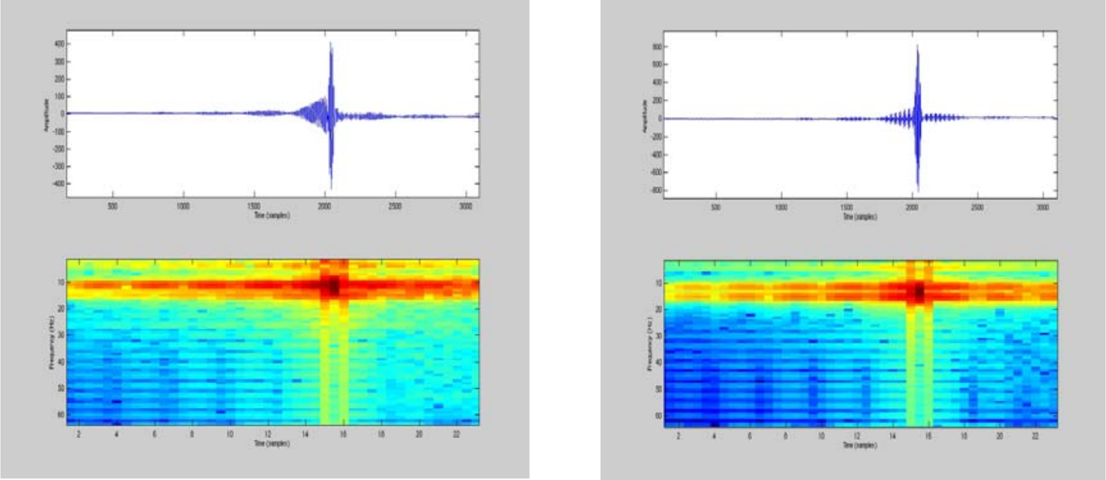


Figure 10. Figure 10 a and b show two triggers selected from group 2 of the channel H0:PEM-BSC6_MAGZ. These triggers are characterized by high frequencies and a wide bandwidth (>32 Hz).

Conclusions

The results in this study point to existence of statistically significant distinct groups in the kleine-Welle trigger population. These groups, when analyzed in the time domain and the time frequency plane reveal triggers with different shapes and physical properties. The variability in the trigger population is shown by the presence of two significant classes in most cases, because three parameters, viz. duration, frequency and snr have been used here. However, the time domain and time-frequency images strongly indicate that trigger shape carries information that could be very useful in further classifying them into sub-categories, i.e. they can contribute additional information in the multi-parameter space for looking in to more classes. This naturally leads the analysis towards the next step in the pipeline, viz. development and implementation of time series classification algorithms. This is being done now within a suggested time frame of July 2007. The aim is to be able to mine maximum information from the trigger database and apply it towards glitch analysis and veto studies. It is expected that this analysis would be able to provide more information towards many of the unexplained glitches that still remain a mystery [13]. One of the most important outcomes of the time series classification is construction of a complete database of different ‘types’ of triggers. There are many questions that could be addressed in the process viz. explanation of double and triple coincidence triggers, correlated triggers etc. The classification analysis is also not restricted to only triggers generated by the LIGO detectors, but can easily be extended to GEO and Virgo as well. These are future goals of this analysis.

Acknowledgment

The author gratefully acknowledges the LIGO laboratory for use of S5 data. This work is supported by the NSF grant PHY-0555842 (2006).

Reference

1. Abramovici, A. et al., LIGO: The Laser Interferometer Gravitational Wave Observatory, Science, 256, 325-33, 1992

2. Zucker, M., S5, S6 and S7, LIGO tech doc. G070070-00-Z, 2007

3. Abbott, B. et al., (LIGO Scientific Collaboration), Search for Gravitational Wave bursts from LIGO’s third Science run, Classical Quantum Gravity, 23, S29-39, 2006

4. Abbott, B. et al., (LIGO Scientific Collaboration), Search for Gravitational Waves from binary black hole inspirals in LIGO data, Phys. Rev. D, 73, 062001, 2006

5. Abbott, B. et al., (LIGO Scientific Collaboration), First all sky upper limits from LIGO on the strength of periodic gravitational waves using the Hough transform, Phys. Rev. D, 72, 102004, 2005

6. Abbott, B. et al., (LIGO Scientific Collaboration), Upper limits on a stochastic background of gravitational waves, Phys. Rev. Lett., 95, 221101, 2005

7. Mukherjee, S., Median based noise floor tracker: robust estimation of noise floor drifts in interferometric data, Classical Quantum Gravity, 20, 925-36, 2003

8. Gonzalez, G. and Cadonati, L., S5 data quality and S5 epochs, LIGO tech doc. G070142-00-Z, 2007

9. gallatin.physics.lsa.umich.edu/~keithr/lscdc/home.html

10. Blackburn, L., Cadonati, L., Chatterji, S., et al., Glitch group S5 activities, LIGO tech doc. G-060407-00-Z, 2006

11. Chatterji, S., S5 Glitch Overview, LIGO tech doc. G070138-00-Z, 2007

12. Blackburn, L., et al., Glitch investigations, with kleineWelle, LIGO tech doc. G050158-00-Z, 2005

13. Desai, S., New glitches seen/studied in Block Normal, LIGO tech doc., G070105-00-Z, 2007

14. Brown, D., S5 online inspiral analysis, LIGO tech doc., G060158-00-Z, 2006

15. Desai, S., Block-Normal, and Event Display and Q-Scan Based Glitch and Veto Studies, LIGO tech doc., G060470-00-Z, 2006

16. Mukherjee, S., Multidimensional classification from kleine Welle triggers from LIGO science run, Classical Quantum Gravity, 23, S661-71, 2006

17. Mukherjee, S., Three types of gamma ray bursts, Astrophysical journal, 508, 314-27, 1998

18. Omohundro, S.M., Efficient algorithms with neural network behaviour, Journal of Complex Systems, 1(2), 273-347, 1987

19. Joachims, T., Making large-scale SVM learning practical, in Schlkopf, B., Burges, C. and Smola, A. eds., Advances in Kernel Methods – Support Vector Learning, MIT Press, Boston, 1999

20. Palau, A., Melgani, F. and Serpico, S.B., Cell algorithms with data inflation for non-parametric classification, Pattern Recognition Letters, 27, 781-90, 2006

21. Prado, R., Molina, F. and Huerta, G., Multivariate time series modeling and classification via hierarchical VAR mixtures, Computational Statistics and Data Analysis, 51, 1445-62, 2006

22. Wilks, S.S., Biometrika, 24, 471, 1932

23. Ashman, K., Bird, C.M. and Zepf, S.E., Astronomical Journal, 108, 2348, 1994

24. Bartlett, M.S., Proc. Cambridge Philosophical Society, 34,33, 1938

25. Johnson, R.A. and Wichern, D.W., Applied Multivariate Statistical Analysis, 3rd ed., Prentice Hall, Englewood Cliffs, 1992

26. gravity.psu.edu/~s3/LSCBootCamp/signalanalysis/src/blimit_example.m

27. www.mathworks.com