SURF 2003: Second Progress Report
Gravitational Wave Bursts: Characterization of Transients in LIGO Interferometer Data

Ed Brambley
Mentor: Dr. John Zweizig

May 2, 2004

1 Introduction

Gravitational waves were predicted by Einstein's 1915 Theory of General Relativity. However, they have yet to be detected, owing to the tiny effect they are predicted to have on the world around us, and even the types of waves passing through the Earth are unknown. The Laser Interferometer Gravitational-Wave Observatory (LIGO) project aims to detect these waves, via interferometry.

The signals from the laser interferometers, as well as other measurements of interest, are digitized and recorded electronically, 24 hours a day. Each observatory generate between 3 and 6 MB/s of data, which is far too much for a human to process. One method used to alleviate this is to use software to automatically note sections of particular interest; these notes are called triggers. One class of artifacts identified are transients, which are unexpected dramatic, short lived blips in the data. At present, triggers for transients are being generated many times a second, which precludes human examination of each event. This SURF project is concerned with the further analysis of these transients, and their detailed classification.

The first progress report detailed the achievements made during the first three weeks of this project. After a period of familiarization with a sample of the data, a method was developed called "Classical BlockNormal" (CBN) analysis, which looked for a single change in behavior of the data, whilst assuming everything was distributed normally. This idea was further developed using the Kolmogorov-Smirnov (KS) test to look for a single change in behavior, without the normal distribution assumption. In addition, a program called "manufacturegwa" was created that would provide a framework to try out different ideas and algorithms, without the overhead of re-writing input/output code. It provided a scripted language with which to randomly generate various types of signal or read in actual LIGO data, and then perform various analyses on these data.

2 Data Preprocessing

While statistical procedures are a good method for identifying a signal, they are greatly helped by first processing the data to eliminate, or at least reduce, some of the noise present. The first method developed was, as suggested at the end of the first progress report, to look at the different frequencies present in the data. This yielded a power spectrum, which plots the power present in each frequency against that frequency. To eliminate bleeding of frequencies into adjacent bands, a Hanning window was used. The power spectrum allowed the dominant 30Hz signal component to be ignored, and other, more subtle signals to be discovered. Moreover, since the power spectrum was produced by a Fast Fourier Transform (FFT), this preprocessing incurs a one-off overhead of complexity only $\Theta(N \log N)$, where N is the number of data points.

The problem with the power spectrum was that it gave no time reference for a signal. To combat this, a method was developed which cut the data into small chunks in time (typically 30ms or so in length), and then produced a power spectrum of each chunk. Hence, the time evolution of distinct frequency bands (typically 16Hz in width) could be analyzed, giving approximate time and frequency information. This is known as a spectrographic method, and is what is planned on being used for performing all statistical analysis on. It also only incurs a one-off overhead of complexity $\Theta(N)$.

An additional problem faced was that the data produced by the spectrographic method above was not distributed normally. Aside from attempting to use other distributions, a Box-Cox transformation was implemented, which is a family of transformations designed to map positive data into a form more normally distributed, while not contorting the data too much. The Box-Cox transformation is given by

$$t_\lambda(x) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x) & \text{if } \lambda = 0 \end{cases}$$

where λ is a real parameter, and $x > 0$. It is continuous in λ, and monotonically increasing and smooth in x. The optimal value of λ to get the best normal fit was estimated using a maximum likelihood estimator. The entire transformation yields a one-off overhead of complexity $\Theta(N)$, which is comparable to the spectrograph generation.

A Box-Cox transformation was also attempted to better fit the spectrographic data to a Gamma distribution. Unfortunately, this did not yield good results, since the Gamma distribution also requires all data points to be positive, and adapting the Box-Cox transformation to have a strictly positive range as well as domain proved problematic. It was therefore decided to abandon this line of investigation due to time constraints.

3 Statistical Procedures

3.1 Methods Assuming a Normal Distribution

The first new statistical procedure developed was a modification of the CBN method which looked for a small interval of interest, either side of which was background noise. The procedure still assumed all data to be normally distributed, and hence was termed the Classical BlockNormal Interval (CBNI) method. The CBNI method showed far more sensitivity to small intervals of interest than the CBN procedure. In addition, it did not suffer from the edge effects that the CBN was prone to, since the data was in effect wrapped round on itself. The CBNI procedure, when used with a limit on the maximum size of intervals of interest, is still of complexity $\Theta(N)$, which is comparable to the complexity of the CBN method.

Unfortunately, as mentioned above, the spectrographically process LIGO data ended up not being normally distributed. In addition to the Box-Cox transformation described above, other methods were investigated that might provide a better fit. One of these methods was to take the real and imaginary parts of individual DFT coefficients, rather than the square of their modulus (as taken for the frequency power). If the original data had been normally distributed, so should these coefficients be. I developed two different, but related, methods to analyze these frequency coefficients. One involved just looking at both the real and imaginary parts at the same time, assuming they were distributed normally — this method was termed the Classical Block Di-Normal Interval (CBDNI) method. A second method was to take into account the possible correlation of the real and imaginary parts, producing a Classical Block Bi-Variate Normal Interval (CBBVNI) method. Both of these two methods have complexity $\Theta(N)$, so are comparable to the CBNI test. In practice, the CBDNI method produced better results than the CBBVNI procedure when applied to LIGO data, as considering the possible correlation of real and imaginary parts produced more false alarms than it helped pick out strong signals.

3.2 Other Parametric Methods

The first non-normal distribution considered was the Gamma distribution. This was motivated by the fact that if the original data had been normally distributed, the real and imaginary coefficients of the DFT would have also been normally distributed, and hence their moduli squared (i.e. the frequency power) would have been distributed according to a Scaled Chi-Squared distribution, which is a special case of a Gamma distribution. In practice, the Gamma distribution shows a much better fit than a Normal distribution to spectrographically transformed LIGO data, and is a good fit for most of the data, apart from a problematic band from about 1–3kHz. The Gamma distribution was used to search for intervals of interest in a similar way to the CBNI method, and hence was termed the Classical BlockGamma Interval (CBGI) method. Best fitting a Gamma distribution using maximum likelihood estimation proved too computationally expensive, so the parameters of the maximum likelihood estimators where themselves estimated using the mean and variance of the data. This proved adequately accurate to give good results, whilst being computationally cheap; the overall complexity of the CBGI method is only $\Theta(N)$, which is comparable to that of the normal methods. The results of this method tend to be similar to the CBNI method's, except that false alarm levels are kept better under control.

An alteration to the CBGI method considered was to consider a Shifted Gamma distribution; i.e. a Gamma distribution plus a constant. Interval searches based on this distribution were termed Classical Block Shifted Gamma Interval (CBSGI) methods. Again, maximum likelihood estimation of parameters proved too computationally expensive, as in this case it would have yielded a $\Theta(N^2)$ algorithm. The parameters were first estimated using mean, variance, and skewness. Unfortunately, this estimation was too inaccurate to give good results, and ended up performing worse than the CBGI method! A second method was to estimate the shift by setting it to the minimum value of all data points. This produced more predictable results, but which were identical to those produced by the CBGI method! It was concluded that only considering a non-shifted Gamma distribution provided enough flexibility, so the CBSGI method was abandoned.

Another parametric distribution used was the Weibull distribution, since it is similar in form to the Gamma distribution. This was implemented in the Classical BlockWeibull Interval (CBWI) method. The Weibull distribution gave just as good a fit to the spectrographically transformed LIGO data as the Gamma distribution, while fitting much better in the 1–3kHz range. Unfortunately, no shortcuts were found to estimate the maximum likelihood estimators, and their direct computation was of $\Theta(N)$, giving the CBWI method an overall complexity of $\Theta(N^2)$; it was estimated that the CBWI method would take about 5 days to process one 60 second length of data. Due to these limitations, the CBWI method was also abandoned.

3.3 Non-Parametric Methods

As detailed in the first progress report, the KS method is a good non-parametric method similar to the CBN test, without assuming normality. An obvious extension to this was to look for intervals of interest, rather than distribution changes, which was implemented in the Kolmogorov-Smirnov Interval (KSI) test. This test is unfortunately of complexity $\Theta(N^2)$, but in practice takes about 15 minutes to complete on a 60 second section of spectrographically transformed LIGO data, so is still usable. However, tests on spectrographically transformed LIGO data showed that this method performed far worse than the CBNI method. This was probably caused by two factors: parametric methods need far fewer points to give accurate results than the KSI method does, and for spectrographically transformed data, a 60 second sample probably did not give enough data to the KSI method to produce good results; also, the Kolmogorov-Smirnov test is known to be insensitive in the tails of the distribution, and this was exactly where tests for transients occurred.

In an attempt to combat the second of these failings, a modified Kolmogorov-Smirnov test, called the Anderson-Darling test, was used. The Anderson-Darling test gives more weight to the tails of the distribution, whilst maintaining the same level of computational complexity. This test was implemented in the Anderson-Darling Interval (ADI) test. The ADI test gave marginally better results than the KSI method, but was unfortunately still nowhere near as good as the parametric methods.

4 Data Analysis Results

Figures 1–4 show the time evolution of power in a selection of frequency bands in a 60 second sample of LIGO data. They were created by the program “manufacturegwa”, using a spectrographic transformation of the data. The sample was chosen so that a prominent transient noted by a trigger occurred after about 30 seconds. Figure 1 shows this transient, and is the sort of thing that current analyses can pick out. Figure 2 shows the transient in detail, and demonstrates features such as a number of peaks during the transient, and a second transient slightly delayed from the first, that it would be useful if the data analysis would pick out; this is currently work in progress.

Figures 5 and 6 show Figures 1 and 3 respectively, after an optimal Box-Cox transformation to make them appear more normally distributed.

In order to test how well certain different distributions fit these data, another variant of the Kolmogorov-Smirnov test (also known by the same name) was applied. This test gives a statistic which represents how much the data deviates from a given distribution, with higher numbers representing larger deviations, and a value of zero indicating a perfect fit. An approximate 99.99% confidence level for this statistic is 1.4, although the precise number depends on the distribution being tested against.

Figures 7–11 show the goodness of fit of several different parametric distributions against varying frequencies of the LIGO data sample. As can be seen, the best fit was achieved by the Box-Cox transformed Normal test (Figure 8), where the statistic remained well below the threshold for all frequencies. The Gamma (Figure 10), Weibull (Figure 11), and Di-Normal (Figure 9) distributions also achieved this level of fit, but only above about 3kHz; between 1kHz and 3kHz, these distributions showed the characteristic deviation of all but the Box-Cox corrected Normal distribution. Of these, the Di-Normal and Weibull distributions equally showed the best fit, with the Gamma distribution showing a reasonable fit, although considerably worse than the first two. The plain Normal distribution (Figure 7) showed the worst fit, with all points well above the threshold.

Figures 12–17 show the results of running the block interval analyses on this section of LIGO data. The graphs plot likelihood of a signal being present against the different frequency bands. An ideal graphs would have a huge peak at 288Hz, indicating the signal present in this channel, and would be as small as possible everywhere else.

The CBNI method (Figure 12) gave an almost ideal graph. The noise in the 1–3kHz range was probably due to non-stationarity of the data in this region, in addition to the worse fit the normal distribution provided in this range, as shown in Figure 7. By preceding the analysis by a Box-Cox transform, the results in Figure 13 were obtained. This would appear to give worse results, despite the normal distribution being a better fit. Note the spikes at 4320Hz and 7168Hz, in addition to the noise in the 1–3kHz range.

Figure 14 shows the results of the analyses using the CBDNI and CBBVNI algorithms, and demonstrates the extra noise caused by considering the covariance of the real and imaginary parts of the DFT coefficients. The spikes at 4320Hz and 7168Hz were still present, but were of much smaller magnitude than either the 1–3kHz noise, or the transient signal at 288Hz.

The results of the CBGI method are shown in Figure 15. They show very similar features to the CBNI results, except the noise levels beyond 3kHz were more stable. Tiny peaks at 4320Hz and 7168Hz were present, but were little more than the general noise. These results seem the most promising.

Figures 16 and 17 show the KSI and ADI results, respectively. As can be seen, they gave very little information. The main 288Hz transient signal that was so prominent in Figure 1 is shown as less significant than the 1–3kHz noise, and significant spikes are shown at 4320Hz and 7168Hz. The ADI test demonstrated more dramatic spikes than the KSI method, but its conclusions are unfortunately no better.

From all the above analysis, the LIGO data sample used showed a large transient at 288Hz, at roughly the expected time. However, there was a large amount of noise between 1kHz and 3kHz that was probably caused by the non-stationarity of the signal in these bands. In addition, two isolated frequencies caused several false signals of significant magnitude — the 4320Hz and 7168Hz bands. Whether there is some physical significance to these features is a question beyond the scope of this SURF project.

5 Work in Progress

Currently, the block interval analysis used above picks out the best interval of interest by various methods, and attempts to classify how interesting it thinks it is. In Figure 2, it generally picks out the first three closely grouped peaks. However, it would be useful to clarify this further, by noting that there are three grouped peaks in this first interval, and another close group of three a short while later. Work on an algorithm to recursively apply the block interval analysis, both to the interval found and on either side of it, is currently ongoing in an effort to detect these features.

6 What's Next

The most likely next steps in this project are to:

- Finish work on the recursive block detection.
- Develop post-processing of the results of the recursive block detection, to detect things like repetitive signals, or trends with neighboring frequency bands. For example, it may well be the case that one transient excites several different frequencies. At present, each frequency band is analyzed in isolation. Post processing the recursive block detection results would allow these different frequency responses to be grouped together as the result of one transient.
- Integrate the current program code (written in C) into the existing DMT C++ framework. The final product of this SURF project will likely be a derived class which fits into the DMT to automatically analyze triggers, and flag important results. There are various base classes already in the DMT which are designed to be extended in this way.
- Run the algorithms on sections of the LIGO data for which there are not any significant known transients, to detect its ability to discover new signals that may previously have been missed.

Other possible avenues of investigation if time permits are to:

- Look at covariance of the signal in time. This may help eliminate problems in the 1–3kHz frequency range due to non-stationarity, and possibly reduce overall noise caused by low levels of non-stationarity else where.
- Investigate other distributions similar in shape to the Weibull and Gamma distributions, which fit the data better than the Gamma distribution, but are less computationally expensive to fit than the Weibull distribution.

List of Figures

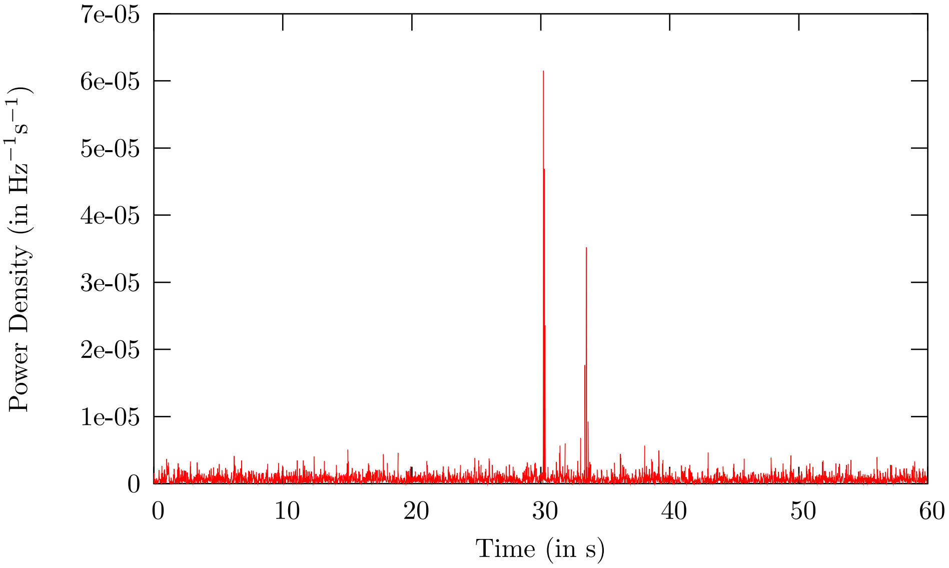

Figure 1: The time evolution of power in the 288Hz frequency band, showing a signal around 30 seconds.

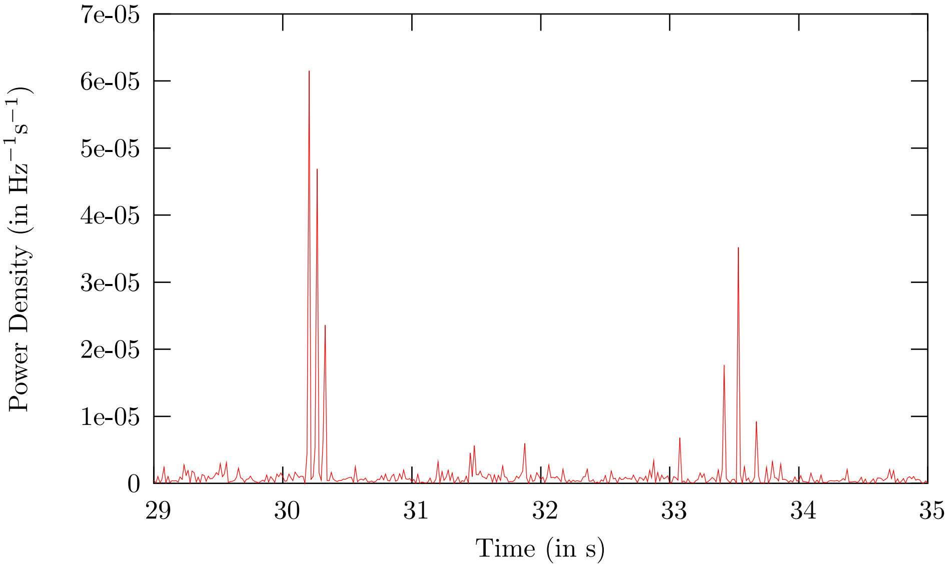

Figure 2: A closeup of the signal in the 288Hz frequency band.

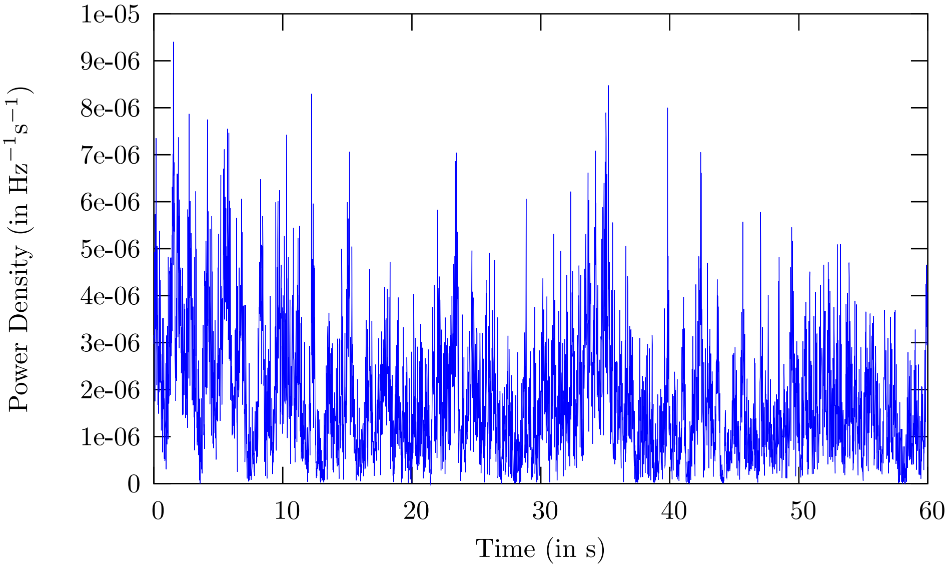

Figure 3: The time evolution of power in the 4320Hz frequency band.

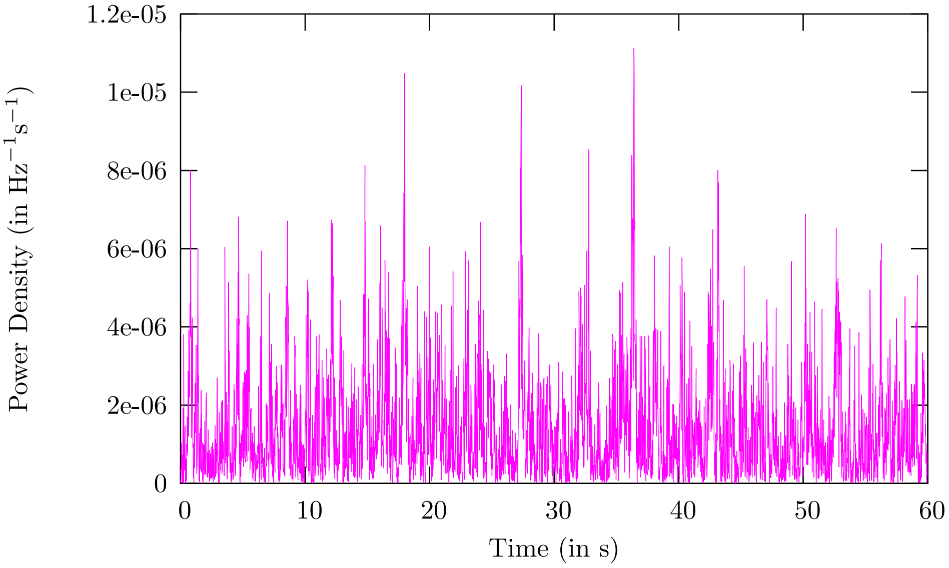

Figure 4: The time evolution of power in the 7168Hz frequency band.

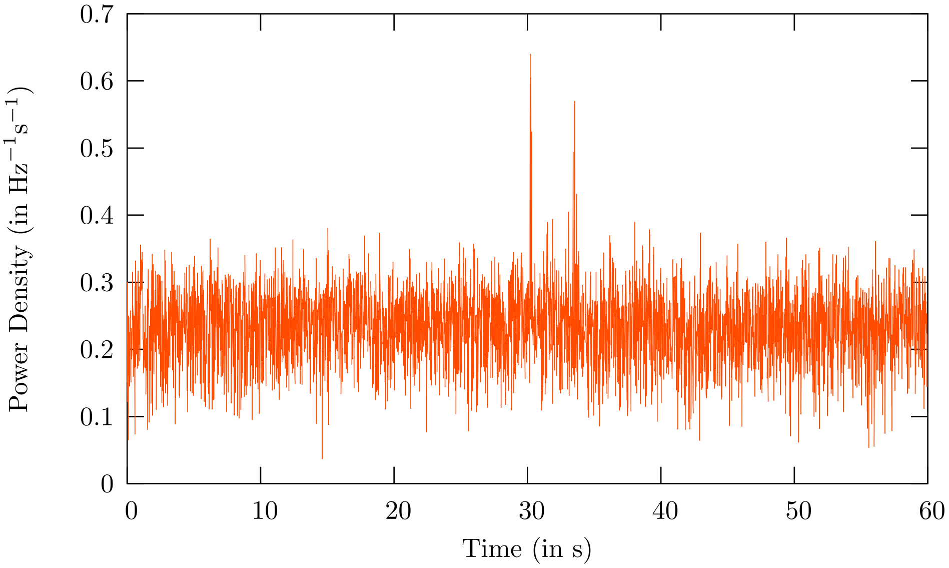

Figure 5: The time evolution of power in the 288Hz frequency band, after a Box-Cox transform.

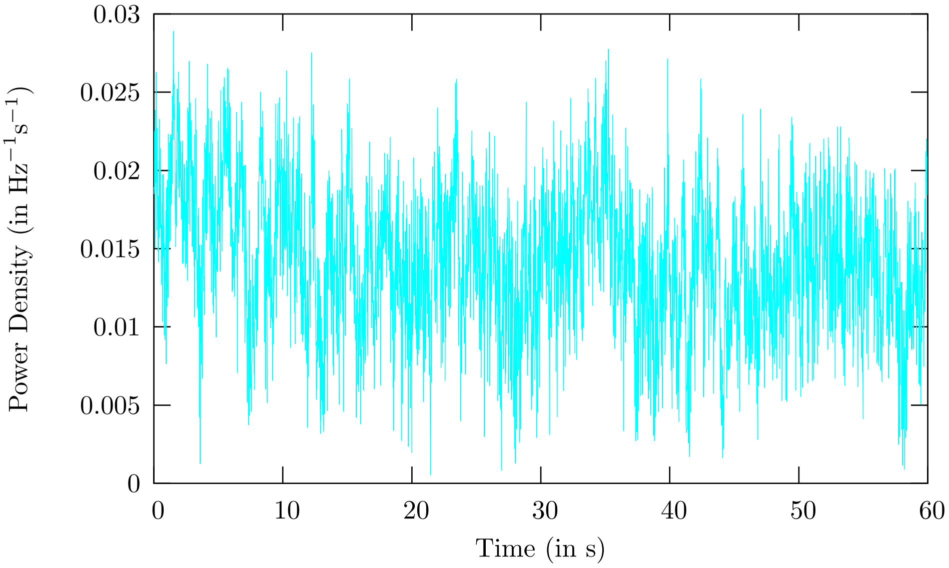

Figure 6: The time evolution of power in the 4320Hz frequency band, after a Box-Cox transform.

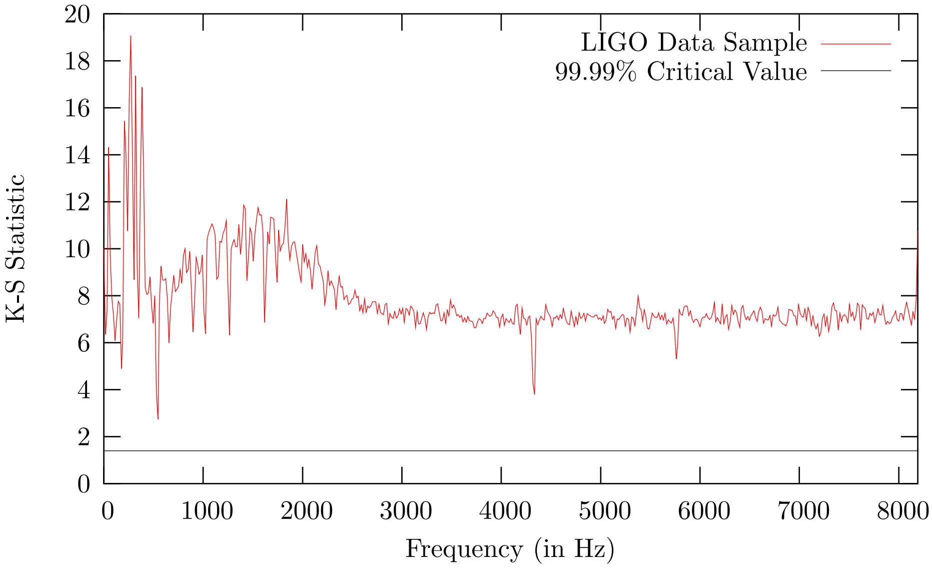

Figure 7: Goodness of fit of a Normal distribution to the time evolution of frequency powers.

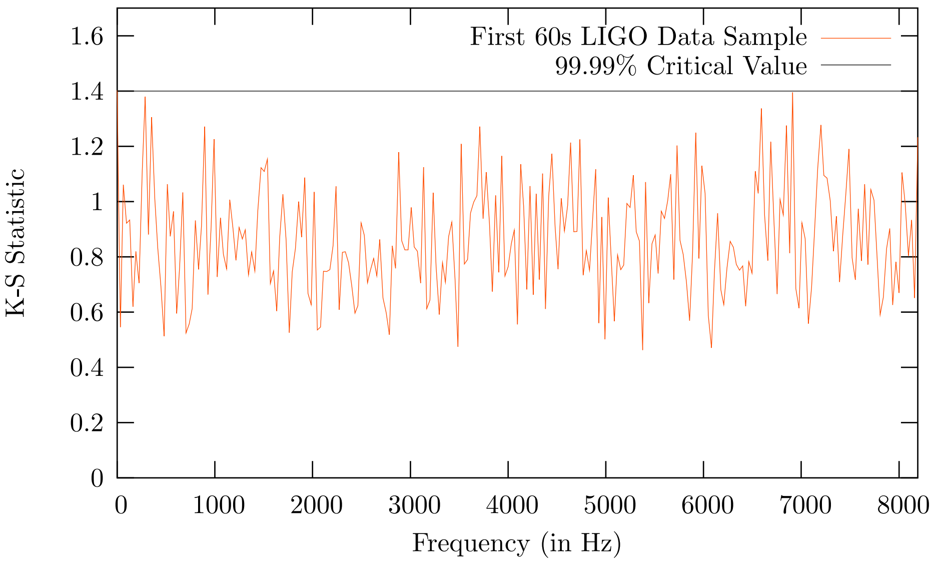

Figure 8: Goodness of fit of a Normal distribution to the time evolution of Box-Cox transformed frequency powers.

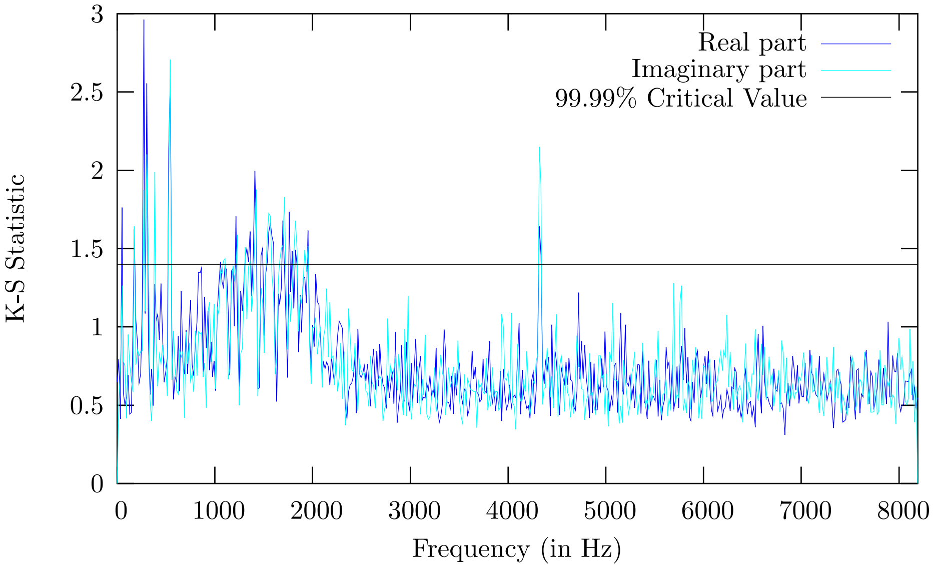

Figure 9: Goodness of fit of two Normal distributions to the time evolution of the real and imaginary parts of the DFT frequency coefficients.

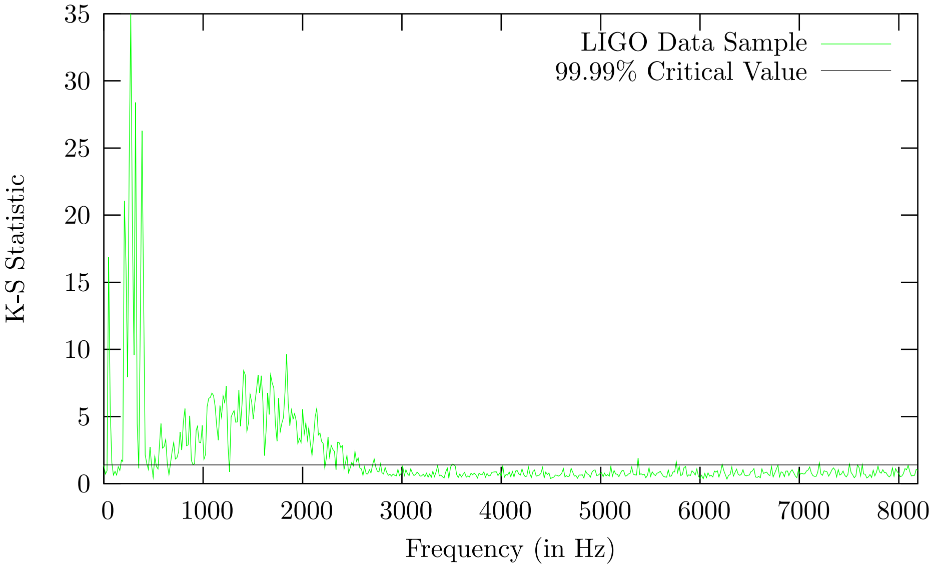

Figure 10: Goodness of fit of a Gamma distribution to the time evolution of frequency powers.

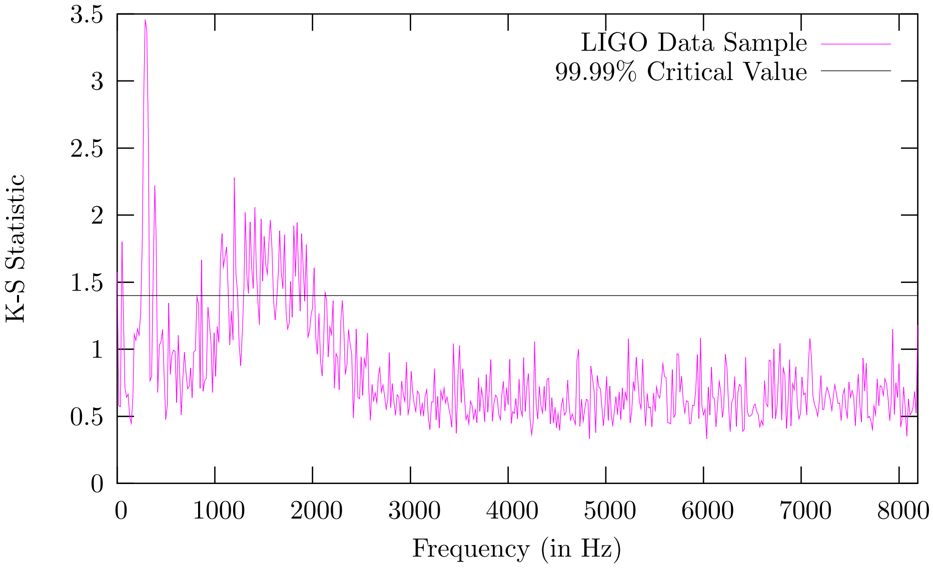

Figure 11: Goodness of fit of a Weibull distribution to the time evolution of frequency powers.

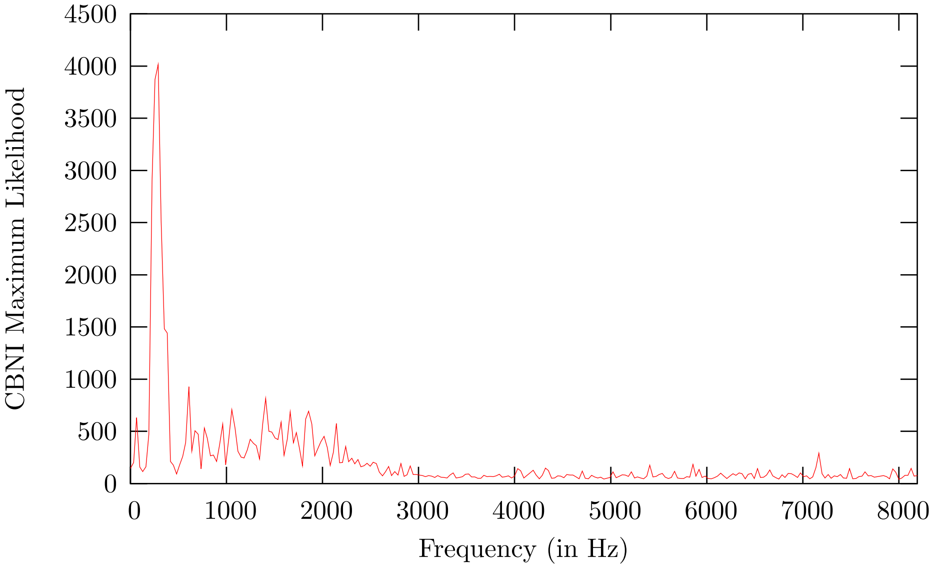

Figure 12: CBNI analysis of the time evolution of frequency powers.

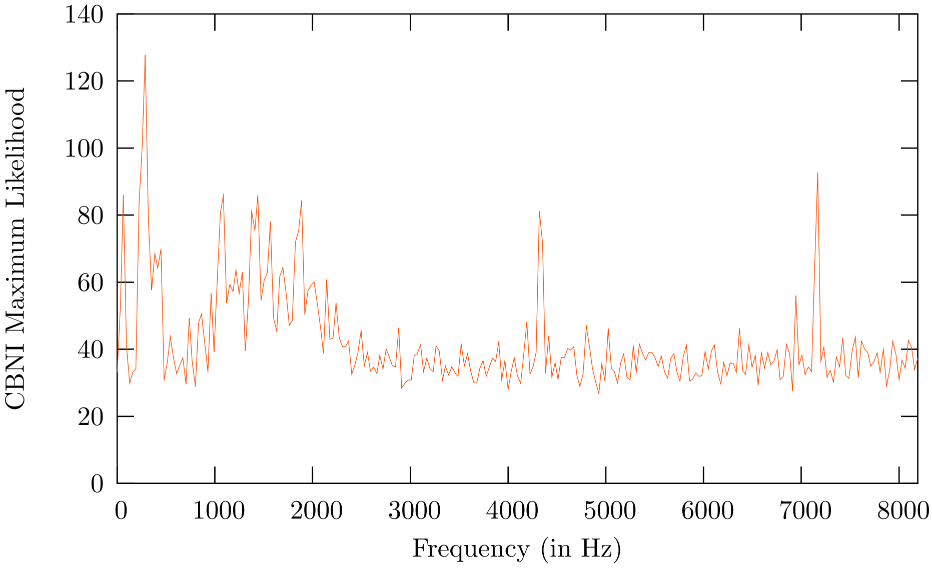

Figure 13: CBNI analysis of the time evolution of Box-Cox transformed frequency powers.

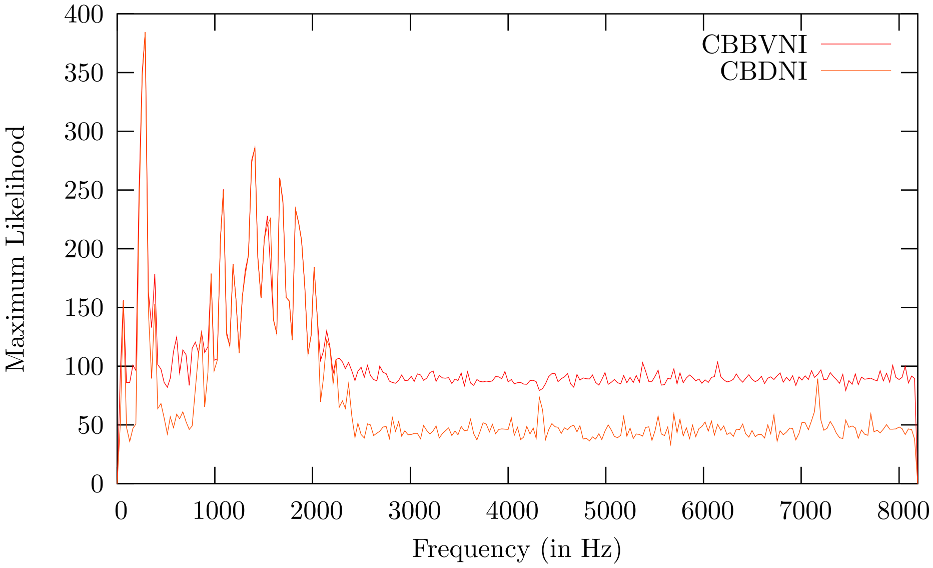

Figure 14: CBBVNI and CBDNI analyses of the time evolution of the real and imaginary parts of the DFT frequency coefficients.

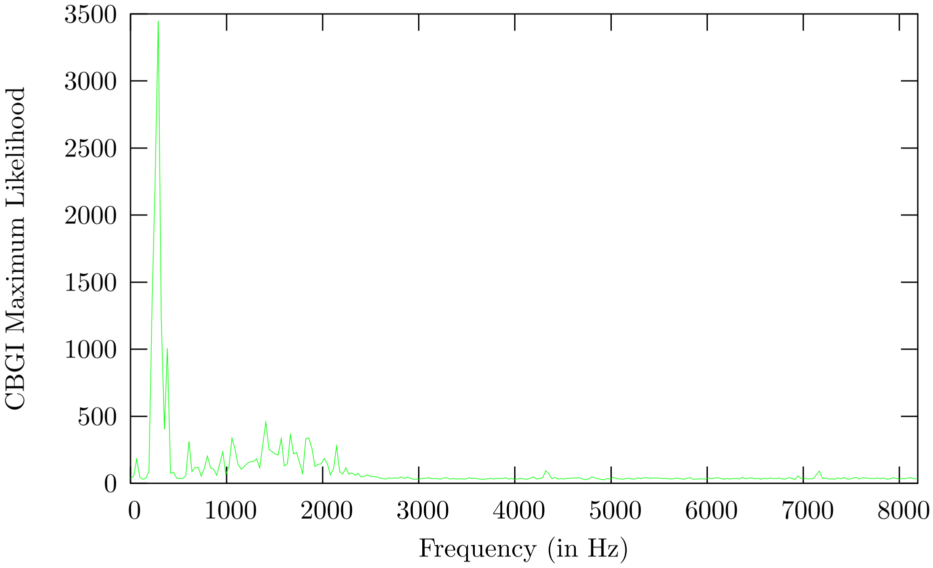

Figure 15: CBGI analysis of the time evolution of frequency powers.

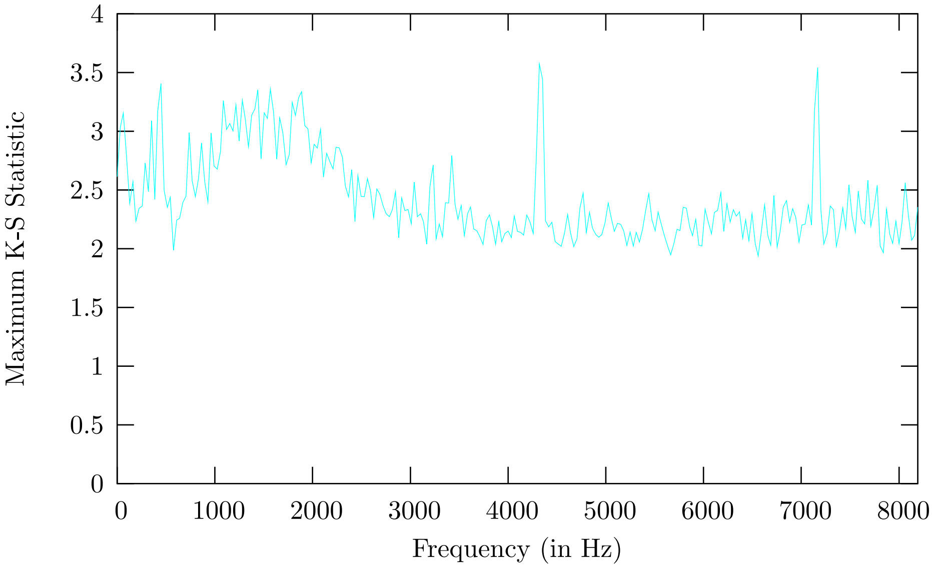

Figure 16: KSI analysis of the time evolution of frequency powers.

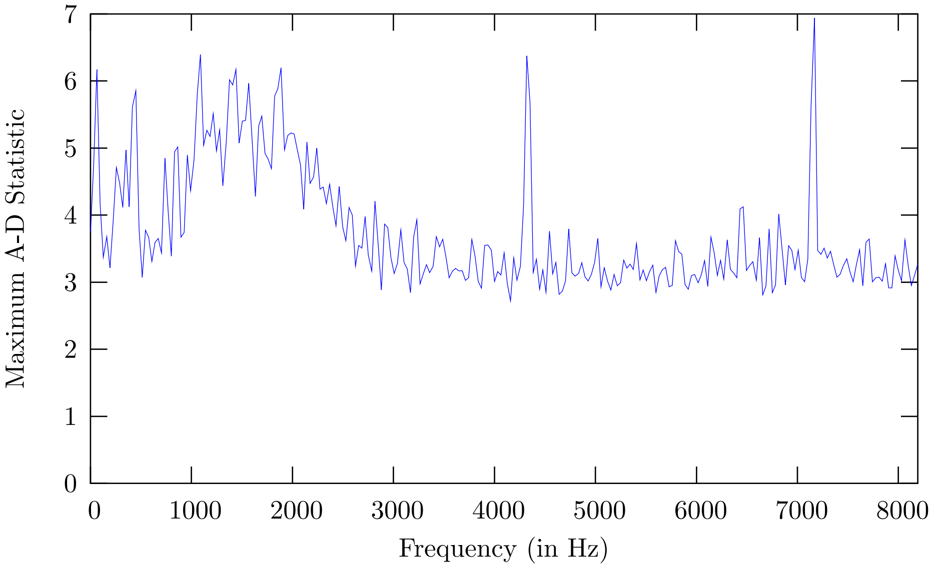

Figure 17: ADI analysis of the time evolution of frequency powers.