Pattern Matching in *COBE* Spectral Analysis

Richard Isaacman
General Sciences Corporation, Code 685.3, NASA/Goddard Space Flight Center, Greenbelt, MD 20771

Shirley M. Read
Hughes STX Corporation, Code 685.9, NASA/Goddard Space Flight Center, Greenbelt, MD 20771

William Barnes
Massachusetts Institute of Technology, Rm. 20F-001, 77 Massachusetts Ave., Cambridge, MA 02139

ABSTRACT

We demonstrate how two simple pattern matching techniques have allowed the Far Infrared Absolute Spectrophotometer (*FIRAS*) aboard the Cosmic Background Explorer (*COBE*) satellite to help us achieve unprecedented photometric accuracy of our spectra while contending with enormous volumes of often heterogeneous data. The first is a robust averaging technique that reduces the volume of data by identifying outlying spectra that in a statistical sense do not "belong" to the ensemble being coadded. The second technique is a variant of the venerable CLEAN algorithm, and is used to remove the effect of cosmic ray hits on our detector. The method is an improvement over the conventional CLEAN in that the algorithm continuously updates the choice of CLEAN profile during the CLEANing process.

Subject headings: pattern matching, spectral analysis, *COBE*, CLEAN, time series, statistics

1. Introduction

The *FIRAS* instrument aboard *COBE* collected over 2 million spectra during its 10 month lifetime and has improved by orders of magnitude the accuracy with which the spectrum of the Cosmic Background Radiation (CBR) is known. Nonetheless, in order to constrain many models of the Big Bang and of galaxy formation in the early Universe, we must work at the sensitivity limits of the instrument. This of course requires extensive averaging of the data, which, though improving the signal-to-noise ratio (S/N), carries some risk. There are at least two dangers: first, there is the danger that anomalous data, taken outside the range of desired operating conditions, may escape inspection and make it into the final average. Some "intelligent" software is required in order to ensure that all of the data to be coadded are consistent, *i.e.*, belong with some statistical certainty to the same parent distribution.

A second danger of coaddition is that cosmic ray hits whose effects may be removable from individual spectra may be washed out in the averaging process and so appear as an elevated random noise level. The power-law distribution of cosmic ray energies requires that some kind of optimal filtering be applied to remove a substantial fraction of all hits ("glitches") before they are lost to view.

2. Consistency Checking

Systematic errors may dominate detector noise if either variations in instrument operating conditions or spatial gradients in the sky emission (*e.g.*, near the galactic plane) cause heterogenous data to be averaged together.

Our selection process operates on data which are collected in the time domain in the form of interferograms ("IFGs"). These are the modulated infrared signals obtained by scanning a moveable mirror platform in the *FIRAS*, which is a Michelson interferometer. Each IFG is the Fourier transform of the average sky spectrum seen by the instrument during the ~ 45 sec collection period.

The IFGs are first sorted by pixel and nominal instrument state (*e.g.*, mirror scan speed). Each sorted group contains an ensemble of anywhere from 5 to 100 IFGs which we hope form a statistically homogeneous set that can be safely coadded. One such ensemble is shown in Figure 1. Each IFG is 512 points long.

We check consistency by first forming a "template", *i.e.*, a robust average of the ensemble in the form of a point-by-point midaverage. That is, if there are k IFGs ${I_j}^i, i = 1...k, j = 1...512$, then the jth element of the template is the average of the middle two quartiles (containing $k/2$ points) of the set $\{I_j\}_k$. The difference between each IFG and the template is then a set of k "residual IFGs" ${R_j}^i, i = 1...k, j = 1...512$ containing noise, residual signal due to spatial or temporal gradients, and any anomalous signal due to other causes.

Because of our coarse pixel size ($\approx 2.6^\circ$) there is a strong gradient in the Galaxy's contribution to the signal within IFG ensembles that span pixels at the edge of the plane. This contribution is expected to vary in amplitude but not in shape (to first order) and so can be recognized and removed. We extract the eigenfunction of the galactic signal by the following procedure.

1. The sign of the peak signal in each residual IFG R^i is checked; if negative, the ith residual R^i is inverted and a flag set.

2. A modified midaverage is performed, using the third to seventh "octiles" (rather than the

middle two quartiles); the result is taken to be the eigenfunction $D_j, j = 1...512$ of the galactic dust spectrum.

3. The best-fit scale factor α^i of D_j (the "secondary template") to the ith residual is determined from the dot product $R^i \bullet D$. This dot product is taken over the range of index j where the modulated signal is greatest.

4. The set of final residuals is defined to be $F^i = R^i - \alpha^i D$ for all indices j.

It is the set of noise-like residuals F^i upon which we perform the statistical consistency checking. For the ith residual we calculate an RMS value σ^i, then in turn form a characteristic overall σ equal to the median of the σ^is. Any IFG whose σ^i differs by more than some threshold (currently 50%) from σ is rejected. Figure I shows the outcome of this test to one small set of IFGs.

A second criterion by which IFGs can be rejected is based on the non-Gaussian (but still well understood) amplitude distribution of the j points within each F^i. We take to be outliers in F^i all the points j where $|F_j{}^i| \geq 5\sigma^i$. More than 5 such points implies either that the noise in this IFG comes from the "wrong" parent distribution or that there is some residual infrared signal.

3. Deglitching

The aggregate effect of untreated charged particle hits on the detectors can reduce the signal-to-noise of the data by up to an order of magnitude at frequencies near the peak of the CBR spectrum. In Figure II, the lower ("Prelaunch") and upper ("Raw Orbit") curves respectively show the noise levels of glitch-free lab data taken prior to launch and untreated data taken in orbit.

Glitches are due to hits on the detector by charged particles (mostly high energy protons) and so their pulse heights obey the ($energy^{-3}$) distribution of cosmic rays. Consequently, a large fraction of the individual glitches are at or below the detector noise limit. Complicating the problem is the fact that the observed shape of the glitches is determined by the transfer function of the onboard electronics and so each glitch rings through an entire IFG. Fortunately, the shape of the profile is known. The upper curve in Figure III shows a large glitch as well as some smaller ones that are not easily resolvable by eye.

We have based our removal technique on the CLEAN algorithm beloved of radio astronomers (Högbom, 1974 A&A *Suppl.*, **15**, 417). Though developed to deconvolve two-dimensional telescope response functions from sky maps, we can apply it to our IFG time series by treating the glitch profile as a one dimensional "beam profile". A major difference with the conventional CLEAN, of course, is that we are interested only in the removal of the components; we have no desire to reconvolve the glitches back into the data afterwards! Thus, we must be conservative in CLEANing

Fig. 1.— Six supposedly homogeneous IFGs selected for coaddition. In fact the middle left IFG is twice as noisy as the others and was rejected by the consistency checker.

Fig. 2.— Noise power density spectra. Prelaunch data are essentially glitch-free due to shielding in the lab. Untreated data from orbit (top curve) are dominated by glitches. Curve DG1 is orbit data deglitched with the traditional CLEAN; DG2 shows improvement after using the "smart" CLEAN.

so as to remove as much of the glitch power as possible while ensuring that misidentifications do not leave spurious residuals.

To reconcile these conflicting demands we first modified CLEAN to use a variable "loop gain" Γ, *i.e.*, the fraction of the profile, centered on and scaled to the current maximum, that is subtracted from the time series. CLEAN uses a constant Γ for every iteration, but we have modified it so that Γ decreases with pulse height. For example, Γ for a large peak may start at the conservative value of 0.10, then increase to *e.g.*, 0.8 as the feature gets "eaten" down into the noise. Curve DG1 in Figure II illustrates that this technique gives noticeable but hardly impressive improvement over un-deglitched data.

The problem here is that CLEAN uses a fixed profile that will in general never match exactly the true glitch. Glitches strike at random times and are of course not synchronized with our onboard sampling clock. Even worse, we compress the data by decimation before telemetering it: that is, the IFG as received on the ground is actually an average of consecutive groups of M points, where M can range from 2 to 12 depending on the operating mode of the instrument. A fixed CLEAN profile will always embody a (generally incorrect) assumption about the arrival time of the glitch with respect to the beginning of the averaging window. Superposing this profile on the true glitch will thus result in a misregistration whose residual after CLEANing will resemble the derivative of the profile. (The problem of CLEANing "sources" which do not fall on sampling gridpoints is discussed by D. Briggs and T. Cornwell elsewhere in these Proceedings.)

We have attacked this problem by making CLEAN smarter. Instead of looking only at the amplitude of a given peak, it looks at the feature's shape as well, fitting a parabola to the peak and its surrounding points and estimating the "true" (*i.e.* noninteger) position of the glitch. The algorithm then uses this "true" position to choose from a library of M profiles, the library having been generated using various arrival phases for the nominal profile with respect to the decimation window. The choice of profile used in the subtraction is updated for every CLEAN iteration.

The results show dramatic improvement. Curve DG2 in Figure II shows that roughly two-thirds of the noise due to glitches has been eliminated. It is doubtful that much additional improvement can be obtained, since the power law energy distribution ensures that most glitches have $S/N \leq 1$. The lower curve in Figure III shows that the CLEANed components sum to an excellent approximation of the true glitches, uncovering some low-level events and one double glitch that are otherwise nearly unobservable.

This research is supported by the Astrophysics Division of the Office of Space Science and Applications at NASA Headquarters.

Fig. 3.— Upper curve: an unusualy large glitch, as well as some smaller ones, in an IFG residual after the template has been subtracted. Lower curve: The summed CLEAN components show a near-perfect match after dynamically updating the glitch profile (curve offset for clarity).