



Radio Frequency Interference Excision Using Cyclostationary Signal Processing

Ryan S. Lynch⁽¹⁾, Evan T. Smith⁽²⁾⁽³⁾, and Matthew Harrison⁽¹⁾

(1) Green Bank Observatory, PO Box 2, Green Bank, WV, 24944, USA, <http://greenbankobservatory.org/>

(2) Department of Physics and Astronomy, West Virginia University, White Hall, Box 6315, Morgantown, WV 26506, USA

(3) Center for Gravitational Waves and Cosmology, West Virginia University, Chestnut Ridge Research Building, Morgantown, WV 26505, USA

Abstract

A cyclostationary process is one whose autocorrelation function is periodic or nearly periodic. The modulation schemes used to encode information give rise to cyclostationarity in many human-generated sources of interference. In contrast, nearly all astrophysical signals are expected to be wide-sense stationary on timescales of interest, making cyclostationarity a potentially robust way of discriminating between interference and astronomical sources. We are developing an algorithm that employs a well-known method of detecting cyclostationary signals and testing its efficacy against a suite of simulated interference covering a wide range of modulation schemes. We present receiver operating characteristic curves and binary classification scores for different types of interfering signals. Our algorithm performs well for many modulation schemes, with F1 and ϕ coefficient scores in excess of 0.9 in some cases, though it shows weaknesses in the case of frequency modulation. We also apply our algorithm to archived Robert C. Byrd Green Bank Telescope observations of a bright millisecond pulsar. We use standard pipelines for blindly detecting and timing pulsars and preliminarily find improvement in data quality according to several metrics, though some undesirable effects are still present. We also show that our algorithm has no negative impact when detecting Galactic HI emission. We thus believe that cyclostationary signal processing shows promise as a means of interference mitigation and discuss opportunities and challenges for employing it more widely.

1 Introduction

Human-generated radio frequency interference (RFI) is a long-standing and growing challenge for radio astronomy. Ultra-wide bandwidth instruments are becoming more common in radio astronomy at the same time as the number of interference sources and their spectrum occupancy is increasing. This confluence threatens the abilities of existing and next-generation radio telescopes to deliver on their scientific promise.

A number of methods exist for identifying and mitigating the impact of RFI, both during real-time data collec-

tion and during post-processing. These methods usually assume that astrophysical signals closely follow Gaussian statistics, and flag data as being contaminated by RFI when the data exhibit non-Gaussian statistical moments [9, 10]. Some other approaches attempt to use the data themselves to subtract RFI at an early stage of post-processing (e.g. [4]). While these approaches can be very successful, they can also break down when extremely strong astrophysical sources depart from the assumption of Gaussian statistics (e.g. bright fast radio bursts).

One robust feature of nearly all astrophysical sources, regardless of their intensity, is that they closely approximate a wide-sense stationary process (one import exception is pulsars, which we discuss in more detail below). Some process, $x(t)$, is wide-sense stationary when its autocorrelation function, R_{xx} is time invariant, i.e.

$$R_{xx}(t, \tau) = E \left\{ x \left(t + \frac{\tau}{2} \right) x^* \left(t - \frac{\tau}{2} \right) \right\} \quad (1)$$

$$= R_{xx}(\tau)$$

where E is the expectation operator, τ is a time-lag, and the $*$ denotes complex conjugation.

In contrast, many sources of RFI are not wide-sense stationary, but instead are *cyclostationary* [5], meaning that R_{xx} is periodic on a timescale T_0 :

$$R_{xx}(t, \tau) = R_{xx}(t + T_0, \tau). \quad (2)$$

Cyclostationarity arises in many sources of RFI because information is encoded by periodically modulating some property (phase, frequency, and/or amplitude) of a carrier wave on a timescale determined by the bit rate, $f_{\text{bit}} = 1/T_{\text{bit}}$. Thus, cyclostationary signal processing (CSP) has the potential to robustly discriminate between RFI and most astrophysical signals. One important exception is pulsars, which are cyclostationary on the timescale of the pulsar rotational period, P [3]. CSP may still be used when observing pulsars as long as care is taken to analyze data in segments that are shorter than P .

We are exploring the efficacy of a CSP-based approach to RFI mitigation using both simulated RFI and real-world

data collected with the Robert C. Byrd Green Bank Telescope (GBT). We have developed an algorithm that detects the presence of cyclostationary signals and apply it to simulated signals that employ a variety of modulation schemes. We then use the optimal algorithmic parameters to mitigate archived baseband data collected between 1100 and 1900 MHz on the bright millisecond pulsar J1713+0747. We process both the mitigated and original, unmitigated data using standard pulsar data analysis tools and compare relevant metrics of scientific data quality. We also compare the integrated spectrum of Galactic HI, which was present in our observing band. We find promising results and some areas for improvement. In §2 we provide more details about our algorithm, in §3 we describe our simulations, and in §4 we present results from our tests with GBT data. We discuss possible improvements and future avenues of research in §5.

2 CSP RFI Identification Algorithm

The autocorrelation function of a cyclostationary process that is periodic on a timescale T_0 can be expanded as a Fourier series with coefficients that can be estimated as

$$C_{xx}(\tau; \alpha) = \frac{1}{T_0} \int_0^{T_0} R_{xx}(t, \tau) e^{-2\pi i \alpha t} dt \quad (3)$$

$$= \frac{1}{T_0} \int_0^{T_0} x\left(t + \frac{\tau}{2}\right) x^*\left(t - \frac{\tau}{2}\right) e^{-2\pi i \alpha t} dt \quad (4)$$

where the second line holds true if $x(t)$ is ergodic and R_{xx} can be estimated from the observed data [5]. C_{xx} is known as the *cyclic autocorrelation function* and is non-zero only at discrete *cycle frequencies* given by $\alpha_k = k/T_0$, where k is an integer. The *spectral correlation function* (SCF; also known as the cyclic spectrum [5]) can be obtained by taking a Fourier transform of C_{xx} with respect to τ :

$$S_{xx}(\nu; \alpha) = \int_{-\infty}^{\infty} C_{xx}(\tau; \alpha) e^{-2\pi i \nu \tau} d\tau \quad (5)$$

where we will refer to ν as the *spectral frequency*. Note that S_{xx} reduces to the typical definition of the power spectral density (PSD) when $\alpha = 0$. The SCF can be normalized by a frequency-shifted version of the PSD so that it is defined on $[0, 1]$, in which case it is referred to as the *spectral coherence function*:

$$\rho_{xx}(\nu; \alpha) = \frac{S_{xx}(\nu; \alpha)}{[S_{xx}(\nu + \frac{\alpha}{2}; 0) S_{xx}(\nu - \frac{\alpha}{2}; 0)]^{1/2}}. \quad (6)$$

This normalized version of the SCF is useful for setting detection thresholds for signals of arbitrary mean and variance.

In practice we will not usually know T_0 for all of the RFI that may be present in a data set, so to detect the presence of a cyclostationary signal we need to have some way of estimating S_{xx} with sufficient resolution in both ν and α . We employ a technique known as the strip-spectrum correlation analyzer (SSCA; [12]), which is a method of time-averaging spectral correlations. In words, the steps in the SSCA are:

1. Take a data set, denoted as $x(n)$, of length N points and duration Δt .
2. Use a windowing function and sliding Fourier transform to channelize subsets of $x(n)$, each of length N' , yielding $X(r, \nu_k)$. Note that r is the time-index and ν_k are the channelizer frequencies (*not* the final spectral frequencies that appear in Equation 5).
3. Multiply $X(r, \nu_k)$ by $x^*(r)$.
4. Perform a discrete Fourier transform of the result of step 3.
5. If desired, compute the spectral coherence using an over-sampled estimate of the PSD.

Mathematically this can be expressed as

$$\hat{S}_{xx}\left(n, \frac{\nu_k}{2} - \frac{q\Delta\alpha}{2}; \nu_k + q\Delta\alpha\right)_{\Delta t} = \sum_r X(r, \nu_k) x^*(r) w(n-r) e^{-2\pi i q r / N} \quad (7)$$

where the Δt subscript indicates averaging, $\Delta\alpha = \Delta t^{-1}$ is the cycle frequency resolution, q is an integer index running from $-N/2$ to $N/2$, and w is a windowing function. Note that the SSCA estimates S_{xx} at discrete spectral and cycle frequencies which are given by $\nu = \frac{\nu_k}{2} - q\frac{\Delta\alpha}{2}$ and $\alpha = \nu_k + q\Delta\alpha$.

Our implementation closely follows [2]. We use `cupy`¹ and `cusignal`² to implement the SSCA in Python with GPU acceleration. $X(r, \nu_k)$ is computed via the `cusignal` short-time Fourier-transform (STFT) routine using a Hann window and a time-domain step size of four samples.

Having used the SSCA to obtain an estimate of the spectral coherence, we define a detection statistic using the *maximum*³ value of $|\rho_{xx}|$. The maximum value follows a Gumbel distribution (a special case of the generalized extreme value distribution), with shape parameters that depend slightly on the choice of windowing function used in the STFT and on the choice of N and N' . We determine the shape parameters for various combinations of N and N' by simulating 10^6 independent realizations of complex Gaussian random values, passing them through our SSCA implementation, and fitting a Gumbel distribution to the observed $\max\{|\rho_{xx}|\}$. We can then set a detection threshold, σ_{thresh} , such that we consider the data under analysis as having significant evidence of cyclostationarity, and thus flag it as being contaminated by RFI, when

$$\max\{|\rho_{xx}|\}_{\text{observed}} > Q(\sigma_{\text{thresh}}) \quad (8)$$

where Q is the quantile function for a Gumbel distribution.

¹<https://cupy.dev/>

²<https://github.com/rapidsai/cusignal>

³According to [6], a near-optimal detection statistic is given by $Y = \sum_{\alpha \neq 0} \int_{\nu} \rho_{xx}(\nu; \alpha) \rho_{xx}^*(\nu; \alpha) d\nu$. We are still investigating use of this statistic but so far have had better results with our approach.

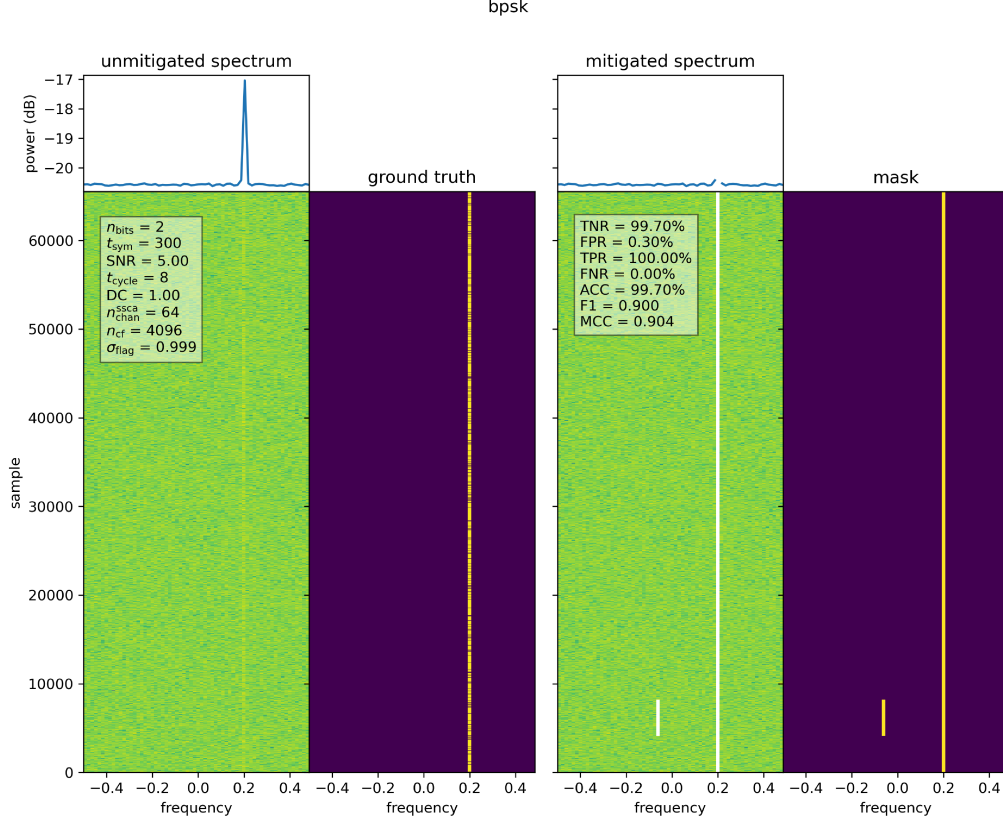


Figure 1. An example of our simulated data for a BPSK signal. The left panel shows the simulated spectrum as a function of time and the ground-truth for when a signal is present. The right panel shows the same data but with the samples flagged by our algorithm having been “masked” (i.e. removed from the data set), and the mask used for this purpose. We also show some parameters of the simulated signal and algorithmic metrics in the text-boxes.

3 Tests Using Simulated Data

Our goal is to explore the efficacy of our algorithm for various choices of N , N' , and σ_{thresh} when applied to real astronomical data. As a first step, we simulated various types of human-generated signals with well-known parameters so that we could compare the results of our algorithm with ground-truth and characterize its performance.

3.1 Simulated RFI Signals

We simulated human-generated signals using a variety of modulation schemes, namely: amplitude-shift keying (ASK), on-off keying (OOK), binary phase-shift keying (BPSK), quadrature phase-shift keying (QPSK), frequency-shift keying (FSK), minimum-shift keying (MSK), Gaussian minimum-shift keying (GMSK), and quadrature amplitude modulation (QAM). We generated a random symbol sequence with some energy-per-bit, E_{bit} , with one bit per symbol (except in the case of QPSK/QAM, which by definition uses two bits per symbol). We then added random Gaussian noise with some noise spectral density, N_0 . To better match the digital backend system used at the GBT, we performed a first-stage channelization via a 64-channel, 24-tap polyphase filterbank (PFB). The result is a complex-value voltage time series for each PFB channel.

We worked entirely in normalized units, i.e. $f_{\text{samp}} = 1$ Hz. We explored various combinations of N , N' , and σ_{thresh} for signal-to-noise ratios (S/N) of five and ten. We also varied the bit-duration, T_{bit} , using values of 100 and 300 samples. In all the simulations presented here we used a duty cycle of 100%, i.e. the signals were always “on” (OOK signals by definition involve an “off” state, but we do not apply any additional cycling).

In some of our test cases the simulated signal had a bandwidth greater than that of a single PFB channel. Furthermore, in the case of the frequency-switched family of signals and OOK, the signal state was modulated using a random symbol sequence, such that we did not know a priori which time/frequency samples would contain RFI. We therefore defined a ground-truth data set by considering an interfering signal to be present when $S/N \geq 1$. We then passed the data set through our algorithm and compared the results to ground-truth, counting the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From these we derived the following metrics:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (\text{true positive rate}) \quad (9)$$

$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}} \quad (\text{false negative rate}) \quad (10)$$

$$\text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}} \text{ (true negative rate)} \quad (11)$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \text{ (false positive rate)} \quad (12)$$

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \text{ (accuracy)} \quad (13)$$

$$\text{F1} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \quad (14)$$

$$\phi = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{[(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})]^{1/2}} \quad (15)$$

We prefer to use ϕ to judge the efficacy of our algorithm since it performs better than the commonly-used F1 score when the size of the TP, TN, FP, and FN classes are very different [1].

3.2 Simulation Results

In Figure 1 we show an example of our simulated data, ground-truth, mitigated data (with samples flagged as containing RFI removed), and mitigation mask. In this example we used a BPSK signal with $T_{\text{bit}} = 300$ and $S/N = 5$. Our algorithmic parameters were $N' = 64$, $N = 4096$, and $\sigma_{\text{thresh}} = 0.999$. The algorithm performs very well, with $\phi = 0.904$, $\text{TPR} = 100\%$, and $\text{FPR} = 0.3\%$.

A more useful metric than this snapshot is a receiver operating characteristic (ROC) curve, which compares TPR and FPR for different values of σ_{thresh} . Fig. 2 shows ROC curves for $S/N = 5$ and 10 and for $T_{\text{bit}} = 100$ and 300 samples for the different modulation schemes in our simulations. Our algorithm performs very well for most signal types, especially at high S/N , but shows weaknesses in the case of FSK and, especially, MSK/GMSK. This is because in the current implementation of the algorithm, each PFB channel is analyzed independently; thus, when a frequency-switched signal moves between different PFB channels our algorithm does not recognize it is as cyclostationary *unless* it returns to a given channel within N samples (note that in these plots $N = 4096$) and is of sufficiently high S/N . We also note that, in general, signals with a shorter T_{bit} are also classified more poorly compared to signals with a longer T_{bit} . This is because a shorter T_{bit} requires a wider-bandwidth signal, which in turn can spread energy across multiple PFB channels and reduce the S/N in any one PFB channel. We discuss potential ways of reducing both of these weaknesses in §5.

Broadly speaking, though, our algorithm performs quite well. When considering a range of factors we find that the best combination of algorithmic parameters is $N' = 64$, $N = 4096$, and $\sigma_{\text{thresh}} = 0.999$, and we use these parameters in tests with real astronomical data.

4 Tests Using Astronomical Data

One of the long-term goals of our project is to implement real-time RFI mitigation for the GBT. Since the data would

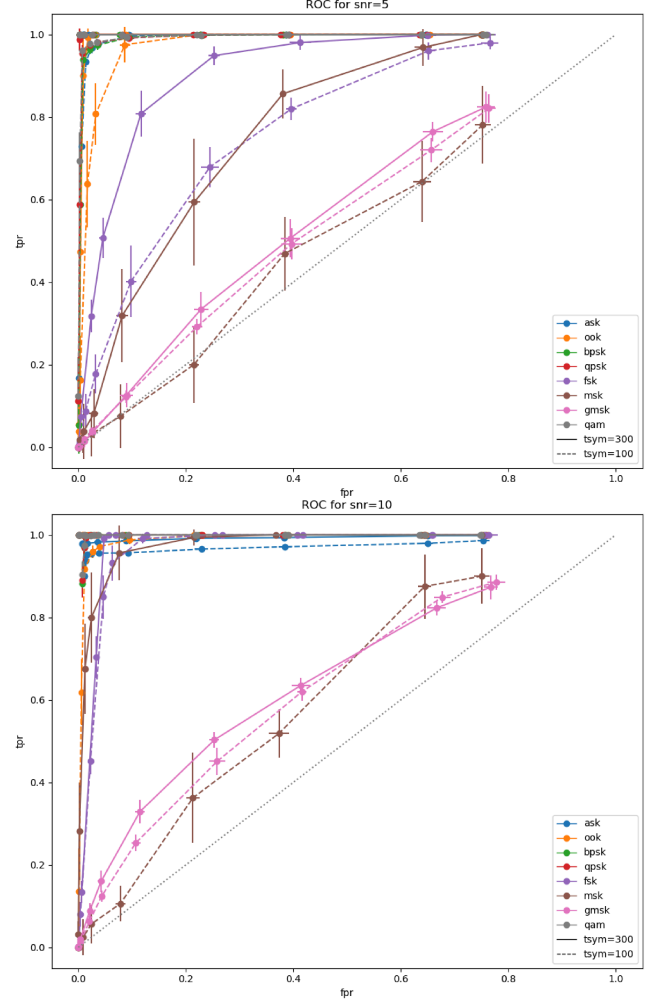


Figure 2. ROC curves for $S/N = 5$ (top) and 10 (bottom) for different signal-types (indicated by color) and T_{bit} (solid lines for 300 samples and dashed lines for 100 samples). Error bars indicate the standard deviation observed in ten independent trials. In all cases we used $N' = 64$ and $N = 4096$. The dotted line is the ROC curve for a non-informative classifier (which has a slope of exactly one). The curve for a perfect classifier would appear at a right-angle. Good classifiers approach close to the upper-left of the plot. Our algorithm performs well for most modulation schemes but shows weaknesses in the case of the frequency-switched family of signals.

be altered in-place, it is vital that we show that our algorithm not only removes RFI, but also does not deleteriously impact the quality of the final science data products. To this end, we have recorded a large amount of baseband data on a variety of astronomical sources, using the Versatile Green Bank Astronomical Spectrometer (VEGAS) and the previous-generation Green Bank Ultimate Pulsar Processing Instrument (GUPPI). This allows us to produce a variety of final science data products from both the original, unmitigated data, and from data that we have modified using our algorithm, and compare the results. In this way we hope to rigorously test our algorithm and increase

Table 1. Blind Search Parameters for PSR J1713+0747

	P (ms)	Best DM (pc cm ⁻³)	Fourier Bin	N_{det}	σ	$\mathcal{P}_{\text{inco}}$	$\mathcal{P}_{\text{coh}}$
Unmitigated	4.570147	16.0	9397.88	553	150.11	11384	169918.6
Mitigated	4.570147	16.0	9397.88	593	173.28	15135	229867.1

Notes — Green highlighting indicates parameters that show improvement after applying our algorithm. Best DM and Fourier Bin refer to those which have the highest significance; N_{det} is the number of trial DMs in which the candidate is detected; σ is the highest Gaussian-significance; $\mathcal{P}_{\text{inco}}$ and $\mathcal{P}_{\text{coh}}$ are the highest incoherent and coherent powers, respectively. This pulsar has an actual $P = 4.57013$ ms and $\text{DM} = 15.917$ pc cm⁻³.

confidence among observers that it can be used safely. We should at least see no change in derived astrophysical properties, and ideally we would see improvement in metrics like S/N.

4.1 GBT Data

For the preliminary tests presented here we used the bright millisecond pulsar (MSP) J1713+0747, which is an important source for projects like the North American Nanohertz Observatory for Gravitational Waves (NANOGrav; [8]). We were careful to analyze data in segments less than a rotational period to avoid flagging the pulsar as RFI due to its own cyclostationary nature.

The data were recorded with the GBT L-Band receiver and VEGAS, using 800 MHz of bandwidth centered at 1500 MHz. The data were channelized via a 512-channel, 24-tap PFB, and were recorded in the GUPPIRAW format. The sampling rate after channelization was 0.64 μ s. There are a number of interfering sources in this frequency range, including the Global Positioning System and Iridium communication satellites.

We used the `dspsr`⁴ [13] package to create standard filterbank “search-mode” data in the PSRFITS format [7]. We did not further channelize the data (i.e. the final number of frequency channels is the same as the number of PFB channels) and we wrote a spectrum to disk every 40.96 μ s. We then simulated a blind search for the pulsar using the PRESTO⁵ package [11] by creating dedispersed time series at dispersion measures⁶ $0 \leq \text{DM} \leq 32$ pc cm⁻³ in step sizes of 0.05 pc cm⁻³. We performed both Fourier-domain and single-pulse searches using standard PRESTO routines.

We also used `dspsr` to coherently dedisperse and phase-fold the baseband data using the known rotational properties of the pulsar, and then measured pulse times-of-arrival (TOAs). TOAs are the most important input for pulsar

timing experiments; deviations between predicted and observed TOAs can be used to study a wide range of phenomena, including fundamental physics, and possibly to directly detect nanohertz-frequency gravitational waves [8]. As such, it is vitally important that our algorithm does not bias TOAs.

We analyzed both the original, unmitigated data and the data passed through our algorithm in the exact same way.

4.2 Results from GBT Data Analysis

As noted above we used a flagging threshold of $\sigma_{\text{thresh}} = 0.999$, $N' = 64$ spectral channels in the SSCA, and analyzed data in segments of length $N = 4096$ samples (a duration of 2.62144 ms, which is less than the 4.57-ms period of the pulsar). We replaced samples flagged as containing RFI with randomly-generated values drawn from a Gaussian distribution with the same mean and standard deviation

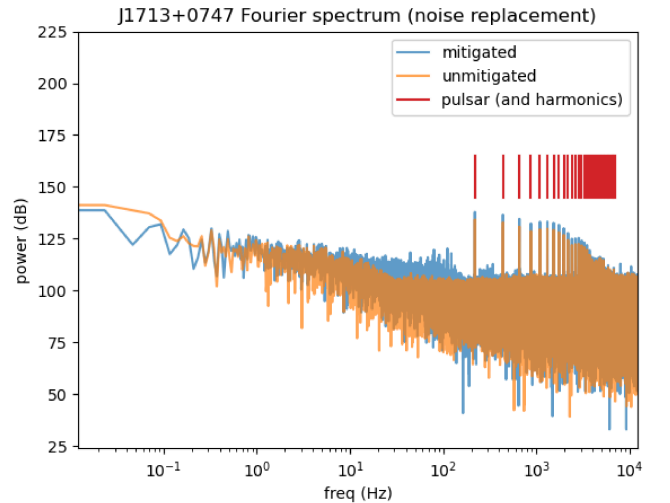


Figure 3. Fourier power spectrum from our observation of PSR J1713+0747. The data were dedispersed at the true DM of the pulsar and integrated in frequency to produce a time series, which was transformed via the PRESTO routine `realfft`. The frequencies corresponding to the pulsar and its harmonics are indicated with red tick marks. The pulsar has higher power in the mitigated data, but there is also excess noise, especially at low frequencies.

⁴<http://dspsr.sourceforge.net/>

⁵<https://github.com/scottransom/presto>

⁶Dispersion measure (DM) is the electron column density towards a pulsar. This ionized medium causes a dispersive delay of a broadband pulse which goes as ν^{-2} . The delay is typically much greater than a rotational period. Thus, without correcting for this effect, a pulsar becomes undetectable when integrating over a wide bandwidth.

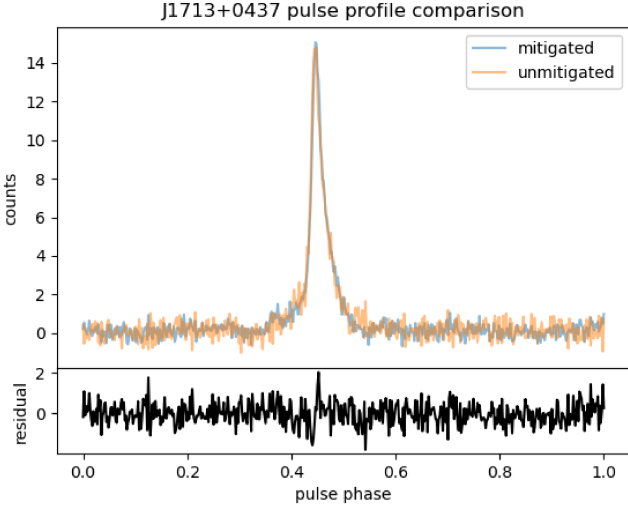


Figure 4. Comparison of integrated, phase-folded pulse profiles for PSR J1713+0747. The bottom panel shows the residuals after subtracting the unmitigated and mitigated data. We see an improvement in S/N with no distortion of the pulse shape.

Table 2. Timing Parameters of J1713+0747

	S/N	Mean σ_{TOA} (μs)
Unmitigated	74	1.51
Unmitigated w/ manual removal of RFI	88	1.50
Mitigated	118	1.57

Notes — We used the PSRCHIVE package (<http://psrchive.sourceforge.net/>; [7]) to manually remove RFI in the second case presented above, and to derive TOAs.

as our actual data. We also explored replacing data with zeros but found that this introduced artifacts that negatively impacted Fourier-domain pulsar searches.

4.2.1 Blind Pulsar Search

We were able to detect the pulsar through a blind search in both the unmitigated and mitigated data. Table 1 compares relevant parameters returned by the `accelsearch` routine that is part of PRESTO. We find good agreement in relevant astrophysical parameters (P and DM) and improvement in the significance of the detection. However, we can also see in Fig. 3 that the Fourier spectrum of the mitigated data exhibits excess noise at several frequencies, especially at $f \lesssim 100$ Hz, when compared to the unmitigated data. This is obviously undesirable, and if the pulsar had a lower rotational frequency then we expect that our algorithm would have made it harder to detect. We are still investigating the source of this excess noise.

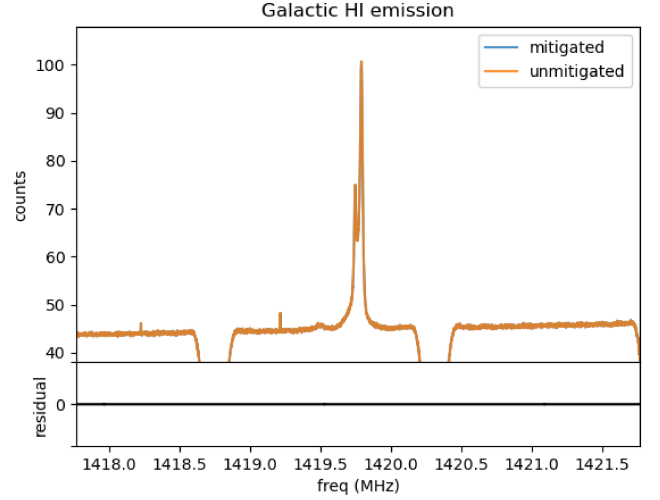


Figure 5. Comparison of Galactic HI emission in our mitigated and unmitigated data. We see no change in the intensity of the HI line.

4.2.2 Pulsar Timing

Fig. 4 shows the integrated, phase-folded pulse profile of PSR J1713+0747 for both our unmitigated and mitigated data. We see an improvement in S/N (see Table 2) and no significant distortion of the pulse shape, which is important because a change in pulse shape could bias the measured TOAs. We also show the mean TOA uncertainty in Table 2. Interestingly, this shows no improvement (and even gets slightly worse, though the change is small). The TOA uncertainty should be proportional to S/N, so this result is unexpected and still requires further investigation.

4.2.3 Galactic HI Emission

In Fig. 5 we show Galactic 21-cm HI emission for both unmitigated and mitigated data. The data are total power, with no subtraction of off-source emission, and no calibration. To resolve the HI line, we further channelized the output of the PFB to a final frequency resolution of ~ 0.7 kHz. The periodic drop in power across the spectrum is due to intentional tapering of the coarse PFB channels used in the VEGAS spectrometer, which minimizes leakage of power into neighboring channels and improves resistance to RFI. The key result of these tests is that, as expected, our algorithm does not flag any astrophysical HI emission.

5 Discussion

Overall, we believe that our algorithm shows a great deal of promise. In both simulations and tests with astronomical data it performs well and leads to some improvement in data quality. However, we identify several ways in which the algorithm could be made better.

The first is the weakness to frequency-switched signals. As noted previously, we analyze PFB channels independently,

and when a signal changes frequency by more than the bandwidth of a PFB channel, the cyclostationary nature of the signal may not be captured by our algorithm. We can alleviate this weakness by analyzing multiple PFB channels jointly to capture signals over a wider bandwidth. This would also help alleviate the weakness to non-frequency-switched signals that occupy a wider bandwidth. This will require more parallelization of our code but is conceptually straightforward.

Another weakness in our approach is that we currently analyze data in segments of fixed size. A more adaptive approach that analyzes data in segments of varying size would allow us to better match a variety of signal types.

Next, when we replace samples that we flag as containing RFI we do so in blocks that are at least as wide as a PFB channel. While this is a natural choice given the GBT's current backend architecture it has the potential to throw out good data along with the bad. For example, if RFI was present in the PFB channel containing the H1 line (see Fig. 5), the entire channel would have been lost, even if the RFI did not overlap with H1. This scenario is unlikely in the case of H1 since it is in a protected band, but this will not generally be the case for other spectral lines. One possible approach is to oversample the PFB and then reduce the frequency resolution after removing bad data.

Finally, we are currently replacing data that has been flagged by our algorithm with statistical noise. This method simplifies post-processing but is conceptually flawed, since it decreases the effective integration time. A better approach would be to "blank" samples flagged as containing RFI, i.e. replacing them with a special value that indicates that they should be ignored entirely, and adjusting the true on-source integration time accordingly. Unfortunately this requires significant modification to existing post-processing software — data are currently stored as 8-bit integers and there is no integer representation of a NULL or NaN value other than zero, which is a possibly valid data value. The modifications that would be required to implement a true null-value replacement scheme are beyond the scope of the current work.

Most of these improvements, however, increase computational complexity. Our long-term goal is to implement a real-time RFI mitigation system, but our current code is far from running in real-time — it takes many multiples of the observation length to process and flag data. So far we have used Python to test our algorithms, and while this is a good choice for rapid prototyping, it is not optimal from a performance standpoint. We are currently developing a pure C++/CUDA implementation, which we hope will speed up processing by at least a factor of a few. However, even this will not be fast enough to process wide observing bandwidths in real-time. Either we will need to find a more efficient SCF estimator than the SSCA, or we will need to significantly increase the amount of computational

resources of our backend systems. Another potential approach is to drop the requirement for real-time processing and instead record baseband data to disk, as we have done for our tests. The data could then be further processed offline and removed after making the data products of interest (e.g. spectra or phase-folded pulsar profiles). This would require significant disk space, since the data rates would be very high, but may ultimately be a more attractive option.

6 Acknowledgments

This work is supported by the National Science Foundation through Advanced Technologies and Instrumentation grant #1910302. This material is based upon work supported by the Green Bank Observatory which is a major facility funded by the National Science Foundation operated by Associated Universities, Inc. We are very grateful to Dr. Chad Spooner for maintaining cyclostationary.blog, which has been an invaluable resource, and for answering our questions about CSP.

References

- [1] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric", *PLoS ONE*, **12**, 6, 06/2017, pp. e0177678, 10.1371/journal.pone.0177678.
- [2] N. J. Carter, "Implementation of cyclic spectral analysis methods", MsT, 1992.
- [3] P. B. Demorest, "Cyclic spectral analysis of radio pulsars", *Monthly Notices of the Royal Astronomical Society*, **416**, 4, October 2011, pp. 2821-2826, doi:10.1111/j.1365-2966.2011.19230.x.
- [4] R. P. Eatough, E. F. Keane, and A. G. Lyne, "An interference removal technique for radio pulsar searches", *Monthly Notices of the Royal Astronomical Society*, **395**, 1, May 2009, pp. 410-415, doi:10.1111/j.1365-2966.2009.14524.x.
- [5] W. A. Gardner, "Exploitation of spectral redundancy in cyclostationary signals", *IEEE Signal Processing Magazine*, **8**, April 1991, pp. 14-36, doi:10.1109/79.81007.
- [6] W. A. Gardner and C. M. Spooner, "Detection and source location of weak cyclostationary signals: simplifications of the maximum-likelihood receiver", *IEEE Transactions on Communications*, **41**, 6, June 1993, pp. 905-916, doi: 10.1109/26.231913.
- [7] A. W. Hotan, W. van Straten, and R. N. Manchester, "PSRCHIVE and PSRFITS: An Open Approach to Radio Pulsar Data Storage and Analysis", *Publications of the Astronomical Society of Australia*, **21**, 3, January 2004, pp. 302-309, doi:10.1071/AS04022.

- [8] M. A. McLaughlin, "The North American Nanohertz Observatory for Gravitational Waves", *Classical and Quantum Gravity*, **30**, 22, November 2013, pp. 224008, doi:10.1088/0264-9381/30/22/224008.
- [9] G. M. Nita and D. E. Gary, "Statistics of the Spectral Kurtosis Estimator", *Publications of the Astronomical Society of the Pacific*, **122**, 891, May 2010, pp. 595, doi:10.1086/652409.
- [10] G. M. Nita and D. E. Gary, "The generalized spectral kurtosis estimator", *Monthly Notices of the Royal Astronomical Society*, **406**, 1, July 2010, pp. L60-L64, doi:10.1111/j.1745-3933.2010.00882.x.
- [11] S. M. Ransom, S. S. Eikenberry, and J. Middleditch, "Fourier Techniques for Very Long Astrophysical Time-Series Analysis", *The Astronomical Journal*, **124**, 3, September 2002, pp. 1788-1809, doi:10.1086/342285.
- [12] R. S. Roberts, W. A. Brown and H. H. Loomis, "Computationally efficient algorithms for cyclic spectral analysis", *IEEE Signal Processing Magazine*, **8**, 2, April 1991, pp. 38-49, doi:10.1109/79.81008.
- [13] W. van Straten and M. Bailes, "DSPSR: Digital Signal Processing Software for Pulsar Astronomy", *Publications of the Astronomical Society of Australia*, **28**, 1, January 2011, pp. 1-14, doi:10.1071/AS10021.