RADIO FREQUENCY INTERFERENCE EXCISION IN SOLAR DYNAMIC SPECTRA USING VARIANCE-BASED SPECTRAL STATISTICS*

Xiaoli Wang and Hongya Ge

Department of ECE
New Jersey Institute of Technology
Newark, NJ, 07102, USA

Gelu M. Nita and Dale E. Gary

Center For Solar-Terrestrial Research
New Jersey Institute of Technology
Newark, NJ, 07102, USA

ABSTRACT

In this paper, a radio frequency interference (RFI) excision algorithm using variance-based analysis on power spectral statistics is proposed and tested on the solar data collected from our frequency agile solar radio-telescope (FASR) subsystem test-bed (FST). A threshold approach working on a proposed test statistic $T(\mathbf{x}; f)$, shown to be F-distributed, is developed to effectively identify the presence of non-Gaussian RFI in the Gaussian solar flare background. Detailed discussions on signal duty cycle and threshold setting are provided. Real-data experimental results are presented to demonstrate the robustness and effectiveness of the proposed algorithm.

Index Terms— radio frequency interference (RFI), power spectral estimation, statistical analysis.

1. INTRODUCTION

In radio astronomy, radio-telescopes are used to passively sense the radio spectrum to detect weak emissions from celestial sources. The high sensitivity requirement of radio telescopes makes them vulnerable to various radio frequency interference (RFI). Quite often the RFI is caused by comparatively strong terrestrial and/or satellite communication signals. There has been increasing research interest and needs in RFI excision algorithms based on a *stochastic analysis* of the dynamic power spectrum of the collected signal [1]. Using a stochastic framework, we can approximate the time-domain or frequency-domain data as a random sequence of certain distributions, whose statistical characteristics can be derived hence utilized for RFI identification.

Following such a framework, the spectral kurtosis (SK) estimator, based on spectral-domain statistics, was proposed in previous work [2], with some restrictions, for non-Gaussian RFI identification. The SK estimate was defined as the ratio of the second-order moment over the squared first-order moment of the instantaneous power of data on a given frequency bin. The SK estimate should be centered at unity with a specifiable variance in the absence of RFI, and deviates from the unity in the presence of RFI.

Continued from the work in [2], we present in this paper the statistical relationship of the dynamic spectrum of solar data on adjacent frequency bins. Our study on solar data spectrum reveals that with high probability the power level change of the solar flare for adjacent and sufficiently narrow frequency bins is gradual, while that for the RFI is typically considerable. Based on this observation, we carry out a variance-based spectral statistical analysis, and propose a test statistic $T(\mathbf{x}; f)$ for RFI identification. This statistic, defined as a ratio of the averaged power located at adjacent frequency bins within the data processing window, is later shown to be F-distributed with certain parameters. Using statistics of $T(\mathbf{x}; f)$, we can determine whether the RFI is present at a given frequency bin. The proposed identifier turns out to be reliable, efficient, and suitable for real-time implementation.

It is well known that there is no universal method for RFI mitigation in radio astronomy analysis. The applicability and success of a mitigation procedure depends on a number of factors such as the type of radio telescope, the type of observation, and the type of RFI [1]. The analysis in this paper is based on our FST system operating at the Owens Valley Solar Array (OVSA), which was newly developed to provide a test-bed for studying RFI excision algorithms [3].

2. DATA MODEL

We first introduce in this section the data model of the solar dynamic power spectrum, where only the RFI-free solar emission is considered. Referring to papers [1, 2, 4], we list some of the main characteristics of the solar data as follows:

- The time-domain real-valued solar emission data samples, $\{x_n\}$, can be modeled as zero mean Gaussian random variables, $x_n \sim N(0, \sigma^2_{x_n})$, with a globally varying while locally flat underlying power spectral density (PSD) $P_x(f)$.
- The complex-valued discrete Fourier transform (DFT) coefficients of a given length-N solar data segment $\mathbf{x}$ are given as (for $k = 0, \ldots, N-1$):

*THIS WORK WAS SUPPORTED IN PART BY NSF GRANT AST-0352915 AND NASA GRANT NNG06GJ40G TO THE NEW JERSEY INSTITUTE OF TECHNOLOGY.

$$X(f_k) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x_n e^{-j2\pi f_k n} \triangleq A(f_k) + jB(f_k),$$

where $f_k = k/N$ is the k-th frequency bin. $A(f_k)$ and $B(f_k)$, as linear combinations of Gaussian random variables, are themselves Gaussian distributed with zero-mean and variances $\sigma^2_{A(f_k)}$ and $\sigma^2_{B(f_k)}$ respectively.

- If no special time-domain windowing function is applied, it can be shown that for $k = 1, \ldots, N/2 - 1$ (excluding the DC and Nyquist frequency bins), the real and imaginary parts $A(f_k)$ and $B(f_k)$ are independent and identically distributed (i.i.d.) Gaussian variables with $\sigma^2_{A(f_k)} = \sigma^2_{B(f_k)}$. Hence the estimated PSD, $P(\mathbf{x}; f_k) \triangleq A^2(f_k) + B^2(f_k)$, with $E\{P(\mathbf{x}; f_k)\} = P_x(f_k) = 2\sigma^2_{A(f_k)} = \sigma^2_{x_n}$, follows a scaled chi-square distribution with two degrees of freedom, χ^2_2.

3. VARIANCE BASED TEST STATISTIC

In our previous work [2], the SK estimator was found to be insensitive to intermittent RFI with an on-off duty-cycle near 50%. Now we describe an alternative estimator that avoids this limitation. We use a variance-based analysis, which is a traditional approach in statistical analysis. A representative example is the Fisher's analysis of variance, i.e., Fisher's F-distribution (Fisher-Snedecor distribution) for statistical significance test.

Based on our extensive study of the dynamic power spectrum provided by the FST system at OVSA, we find that typically RFI is caused by various communication signals located at nearly constant frequency bins, occupied continuously or discontinuously in time, with a distinguished instantaneous power level compared with the background solar emission located at adjacent frequency bins. For cases of continuum radio emission, a gradual power change along adjacent frequency bins can be expected even for strong solar flares.

In our study it is assumed that there are available for analysis M adjacent data blocks, $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}$, each containing N time-domain samples. The data block $\mathbf{x}^{(m)}$ ($m = 1, \ldots, M$) is first transformed into N frequency-domain DFT coefficients, $X^{(m)}(f_k)$, and then N corresponding PSD samples, $P(\mathbf{x}^{(m)}; f_k)$. Due to the real-valued realization, only half of the PSD samples contain useful information and are thus used. To facilitate the statistical analysis, we propose to use a test statistic $T(\mathbf{x}; f_k)$, defined as the ratio of the averaged power (across all M data blocks) on two adjacent frequency bins f_k and f_{k-1}:

$$T(\mathbf{x}; f_k) \triangleq \frac{P_{\text{ave}}(\mathbf{x}; f_k)}{P_{\text{ave}}(\mathbf{x}; f_{k-1})}, \tag{1}$$

where $P_{\text{ave}}(\mathbf{x}; f_k) = \frac{1}{M}\sum_{m=1}^{M} P(\mathbf{x}^{(m)}; f_k)$ is χ^2_{2M} distributed (scaled) due to the non-overlapping data segmentation in use. For RFI-free solar emission data, we assume that $P_x(f_k) \approx P_x(f_{k-1})$ for adjacent and sufficiently narrow frequency bins due to the gradual change of its PSD. Accordingly, we have $P_{\text{ave}}(\mathbf{x}; f_k)$ and $P_{\text{ave}}(\mathbf{x}; f_{k-1})$ i.i.d. χ^2_{2M} distributed. Consequently, for $k = 2, \ldots, N/2 - 1$, $T(\mathbf{x}; f_k)$ follows a F-distribution [5],

$$f_{T(\mathbf{x};f_k)}(y) = \frac{\Gamma(2M)}{\Gamma(M)\Gamma(M)} \frac{y^{M-1}}{(y+1)^{2M}}, \tag{2}$$

where $\Gamma(M) = (M-1)!$ denotes the Gamma function. For $M > 2$, the mean and variance of $T(\mathbf{x}; f_k)$ can be found as,

$$E\{T(\mathbf{x}; f_k)\} = \frac{M}{M-1} \approx 1(\text{ for large } M) \tag{3}$$

$$Var\{T(\mathbf{x}; f_k)\} = \frac{M(2M-1)}{(M-1)^2(M-2)} \approx \frac{2}{M}(\text{ for large } M). \tag{4}$$

These statistical characteristics will be later shown as crucial parameters for RFI identification.

4. DETECTION OF DETERMINISTIC SIGNALS

In this section, we model the RFI as a deterministic sinusoidal signal of modulated amplitude α. Moreover, we consider that the frequency of the deterministic signal exactly matches to one of the non-DC nor Nyquist discrete frequency bin f_k.

The power spectrum sample $P^{(m)}(\mathbf{x}; f_k)$ then follows a scaled noncentral chi-square distribution, $\chi^2_2(\lambda_k^{(m)})$ [2]. Hence, the averaged power spectrum $P_{\text{ave}}(\mathbf{x}; f_k)$ follows a scaled noncentral $\chi^2_{2M}(\lambda_k)$ distribution. Therefore, $T(\mathbf{x}; f_k)$ is noncentral F-distributed. The non-centrality parameter λ_k can be given as $\lambda_k = M\alpha^2/2P_x(f_k) = M\eta_k$, with $\eta_k \triangleq \alpha^2/2P_x(f_k)$ denoting the signal-to-noise ratio (SNR) at the frequency bin f_k (viewing RFI as signal). The mean and variance of $T(\mathbf{x}; f_k)$ can then be given as,

$$\begin{aligned} E\{T(\mathbf{x}; f_k)\} &= \frac{2M + \lambda_k}{2M - 2} \\ &\approx 1 + \frac{\eta_k}{2} \quad (\text{for large } M), \end{aligned} \tag{5}$$

$$\begin{aligned} Var\{T(\mathbf{x}; f_k)\} &= \frac{(2M+\lambda_k)^2 + 4(M+\lambda_k)(M-1)}{4(M-1)^2(M-2)} \\ &\approx \frac{1}{M}\left[2 + 2\eta_k + \frac{1}{4}\eta_k^2\right] \quad (\text{for large } M). \end{aligned} \tag{6}$$

In many applications, the interfering signal may not be continuous (or always ON) within our data processing window. In order to model such a situation, we further introduce the concept of signal duty cycle. Particularly, we assume that the randomly occurring transition exactly coinciding with R of the M data blocks results in a duty cycle $\beta \triangleq R/M$. For this given duty cycle, the revised mean and variance of the test statistic $T(\mathbf{x}; f_k)$ can be represented as (for large M):

$$E\{T(\mathbf{x}; f_k)\} \approx 1 + \frac{1}{2}\beta\eta_k, \tag{7}$$

$$Var\{T(\mathbf{x}; f_k)\} \approx \frac{2}{M} + \frac{\beta}{M}\left[2\eta_k + \frac{1}{4}\eta_k^2\right]. \qquad (8)$$

In general, a non-Gaussian RFI located at frequency bin f_k should cause $T(\mathbf{x}; f_k)$ to deviate statistically from its expected value given in eq.(3). Thus the testing of $T(\mathbf{x}; f_k)$ provides a feasible way to discriminate RFI from background signal. From analytical results given in eqs.(3) and (7), we can identify the presence of RFI by testing the value of $T(\mathbf{x}; f_k)$ against thresholds derived from eqs.(3-4). Fig.1 below shows the simulated performance of the proposed testing of $T(\mathbf{x}; f_k)$ when we have deterministic sinusoidal RFI present at f_k. The parameter M is chosen as 1000. We allow SNR η_k varying from 1dB to 10dB to cover the worst case of the theoretical expectation, and duty cycle β varying from 0 to 1. Notice that when M is reasonably large, we can further approximate this F-distributed test statistic $T(\mathbf{x}; f_k)$ for RFI identification to a simple Gaussian distribution $N(1, 2/M)$. Therefore, a 99.7% of confidence interval (CI) for identifying RFI can be set as three times of the standard deviation from the nominal mean, i.e.,$1 \pm 3\sigma_T = 1 \pm 3\sqrt{2/M}$. As seen from the figure, the algorithm works well for this simplified simulation involving both duty cycle and SNR.

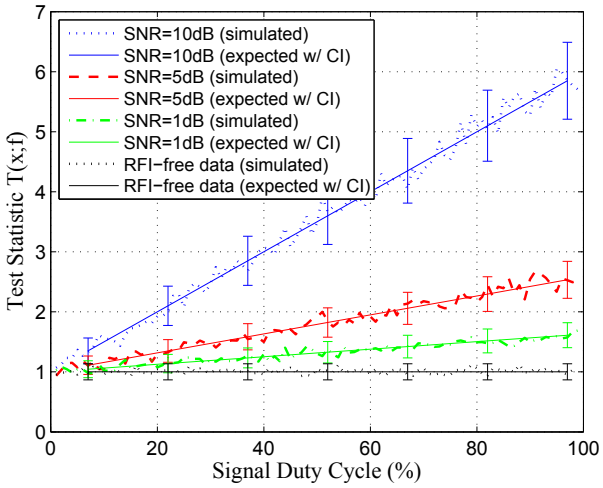


Fig. 1. Testing of $T(\mathbf{x}; f_k)$ vs. duty cycle of the RFI

5. IMPLEMENTATION AND EXPERIMENT RESULTS

As mentioned before, the experiments are carried out using data from our FST system, which was developed recently to provide a test-bed for studying RFI excision algorithms for solar data. The proposed method is tested on selected previously recorded data files, where each includes data of 4 s elapsed time over an instantaneous 500 MHz bandwidth that can be tuned anywhere in the frequency range of 1 to 9 GHz.

Our instrument is able to record time-domain data at a sampling rate of 1 sample/ns, while due to the limitation of the on-board memory of the digitizing system, no more than about 2 ms of contiguous time-domain data segment can be recorded at a time. The data we used were taken in mode 3, as defined in the reference paper [3], where every 0.1 ms of data samples are separated by a 20 ms gap. For the solar spectrum displayed (tested) here, a quadratic time-frequency representation is used. The number of frequency bins N is 8192, and the number of time blocks M is 2400. These parameters provide a frequency resolution of around 0.122 MHz and a time resolution of 1.6 to 1.7 ms. Following our previous analysis, the thresholds are defined as $1 \pm 3\sqrt{2/M}$.

As shown in figures 2−3 next page, we have tested typical scenarios where solar flares are contaminated with RFI. It can be easily seen that when we have RFI present at certain frequency bins, which can be seen as continuous or intermittent horizontal lines across the spectrum (shown in (a) for both figures), sharp spikes can be inspected in the averaged power (shown in (b)). And at these same frequency bins, $T(\mathbf{x}; f_k)$ deviates (exceeds the threshold) from the expected value 1 (shown in (c)). By simply eliminating those frequency bins, we get the cleaned spectrum presented in (e), where RFI is almost eliminated and data of interest is comparatively enhanced. The corresponding averaged power after cleaning is also shown in (d) for reference.

6. CONCLUSIONS

The spectral statistics of the solar radiation data are analyzed in this paper. A variance-based test statistic $T(\mathbf{x}; f)$, shown to be F-distributed, is introduced for RFI identification. The RFI identification procedure is developed based on the confidence interval of our derived statistics of $T(\mathbf{x}; f)$. The proposed identifier is of low-complexity and performs remarkably well on solar data collected by our FST system.

7. REFERENCES

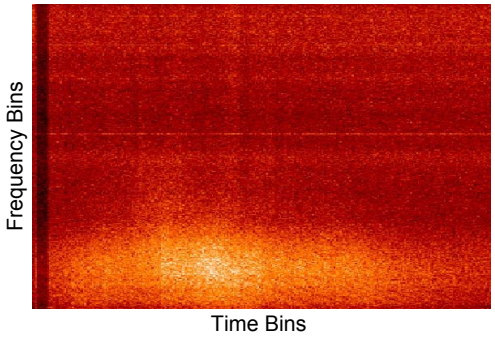
[1] P. A. Fridman and W. A. Baan, "RFI mitigation methods in radio astronomy," *Astronomy & Astrophysics*, 378, 327-344, 2001.

[2] G. M. Nita, D. E. Gary, Z. Liu etc., "Radio Frequency Interference Excision Using Spectral-Domain Statistics," *Publications of the Astronomical Society of the Pacific*, 119(857), 2007.

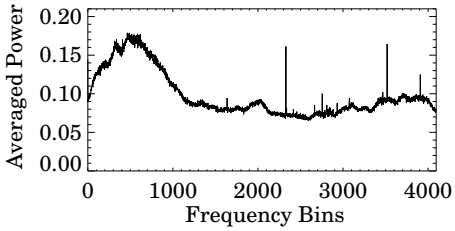
[3] Z. Liu, D. E. Gary, G. M. Nita etc., "A Subsystem Testbed for the Frequency Agile Solar Radiotelescope," *Publications of the Astronomical Society of the Pacific*, 119(303), 2007.

[4] A. Leshem, A-J. van der Veen and A. J. Boonstra, "Multichannel Interference Migitation Techniques in Radio Astronomy," *Astrophysical Journal Supplement*, 131(1), 355, 2000.

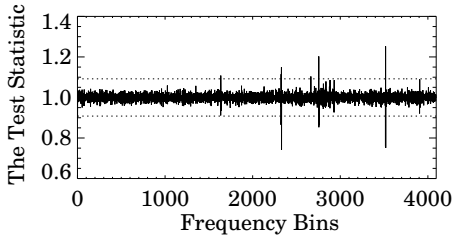
[5] S. M. Kay, *Fundamentals of Statistical Signal Processing*, Prentice Hall, 1993.



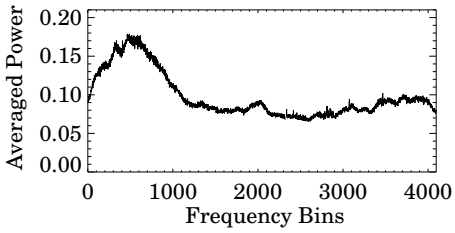
(a) RFI contaminated spectrum



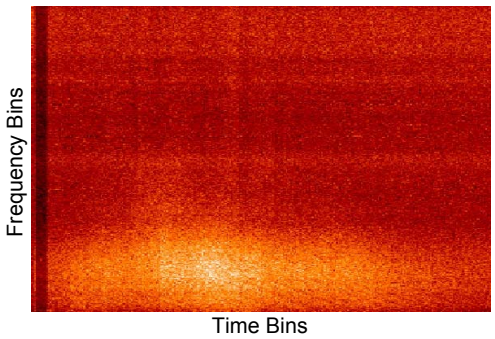
(b) Averaged power (before cleaning)



(c) The test statistic $T(\mathbf{x}; f_k)$ (for RFI identification)

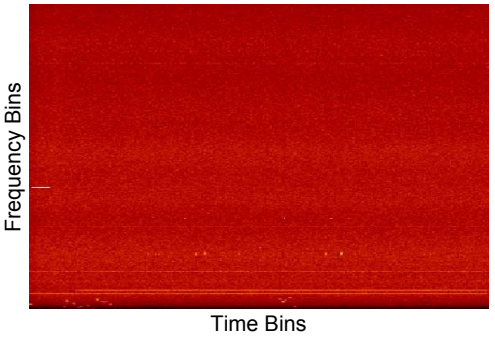


(d) Averaged power (after cleaning)

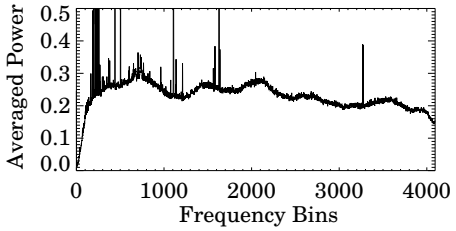


(e) Cleaned spectrum

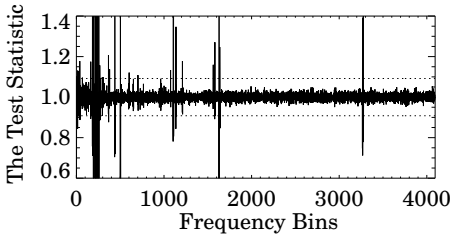
Fig. 2. Experiment results for RFI mitigation (Set 1)



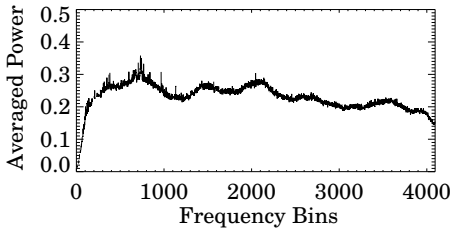
(a) RFI contaminated spectrum



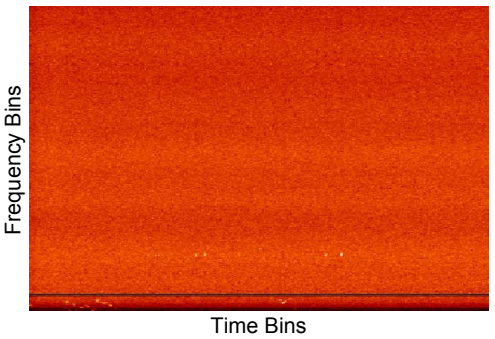
(b) Averaged power (before cleaning)



(c) The test statistic $T(\mathbf{x}; f_k)$ (for RFI identification)



(d) Averaged power (after cleaning)



(e) Cleaned spectrum

Fig. 3. Experiment results for RFI mitigation (Set 2)