A Wideband Spectrometer with RFI Detection

DALE E. GARY, ZHIWEI LIU, AND GELU M. NITA

Physics Department, New Jersey Institute of Technology,
Newark, NJ

ABSTRACT. We report on the design and construction of a wideband spectrometer of 500 MHz instantaneous bandwidth that includes automatic radio frequency interference (RFI) detection. The implementation is based on hardware developed at the Center for Astronomical Signal Processing and Electronics Research (CASPER). The unique aspect of the spectrometer is that it accumulates both power and power-squared, which are then used to develop a spectral kurtosis (SK) estimator. The SK estimator statistics are used for real-time detection and excision of certain types of RFI embedded in the received signal. We report on the use of this spectrometer in the Korean Solar Radio Burst Locator (KSRBL). This instrument utilizes four of these 500 MHz bandwidth SK spectrometers in parallel, to achieve a 2 GHz instantaneous bandwidth that is time multiplexed over the entire 0.24–18 GHz radio frequency range, to study solar bursts. The performance of the spectrometers for excising RFI over this range is presented. It is found that the algorithm is especially useful for excising highly intermittent RFI but is less successful for RFI due to digital signals. A method we call multiscale SK is presented that addresses the known blindness of Kurtosis-based estimators to 50% duty-cycle RFI. The SK algorithm can also be applied to spectral channels prior to correlation to remove unwanted RFI from interferometer data.

1. INTRODUCTION

The Korean Solar Radio Burst Locator (KSRBL, described in detail in Dou et al. 2009) is the first spectrometer to employ the technique of spectral kurtosis (SK, Nita et al. 2007) to identify and remove radio frequency interference (RFI) signals in the radio spectrum. Man-made terrestrial radio signals, such as TV, FM/AM radio, cellular telephone, digital data links, radar, and many others, occupy an ever increasing part of the radio spectrum. Several techniques have been investigated to address the problem of RFI in data from radio astronomical instruments (for a recent discussion, see Offringa et al. 2010 and references therein). Given the continuous frequency coverage of many new and proposed instruments, including KSRBL, it is unavoidable that RFI signals are recorded by the instrument, which thus represent unwanted interference that must be removed. RFI at fixed frequencies can in principle be flagged, but highly intermittent signals affect the data in subtle ways that are very difficult to find and remove. The SK algorithm described by Nita et al. (2007) is highly sensitive to such intermittent RFI and has been implemented for the first time in the KSRBL digital hardware described here.

The KSRBL hardware, described in Dou et al. (2009), uses four identical boards designed by the Center for Astronomical Signal Processing and Electronics Research (CASPER, see, e.g., Parsons et al. 2008). These are referred to as IBOBs (Internet break-out boards). Each IBOB is equipped with a field programmable gate array (FPGA) programmed as a 2048 channel digital spectrometer, operating on a 500 MHz instantaneous

bandwidth, implemented using the CASPER design tools. KSRBL was temporarily operated at the Owens Valley Radio Observatory (OVRO), California, US, and in August 2009 was moved to Daejeon, Republic of Korea, where it is now operated by the Korean Astronomy and Space Science Institute (KASI).

Based on our early experience with KSRBL, described here, we have found that the thresholds for RFI detection suggested by Nita et al. (2007) and initially used in Dou et al. (2009) are not optimum. Nita & Gary (2010) have extended the theoretical understanding of the SK algorithm, in particular finding expressions for the probability density function (PDF) and the related cumulative function (CF) of the SK estimator, to further refine the thresholds for RFI detection. This article demonstrates the improved performance of the refined thresholds using real-time data.

In § 2, we provide those details of the SK algorithm needed to understand the design and operation of the hardware spectrometer. In § 3 we introduce the KSRBL spectrometer hardware design. The SK distribution calculated from solar data obtained with the spectrometer is compared with the theoretical SK estimator PDF to verify its operation in § 4. In § 5 we describe a method for identifying and classifying various types of RFI based on data from the spectrometer. In § 6 we investigate strategies for improving RFI elimination, including an additional approach we call multiscale SK, to reduce the duty-cycle limitation of the SK algorithm for detecting RFI. We conclude in § 7.
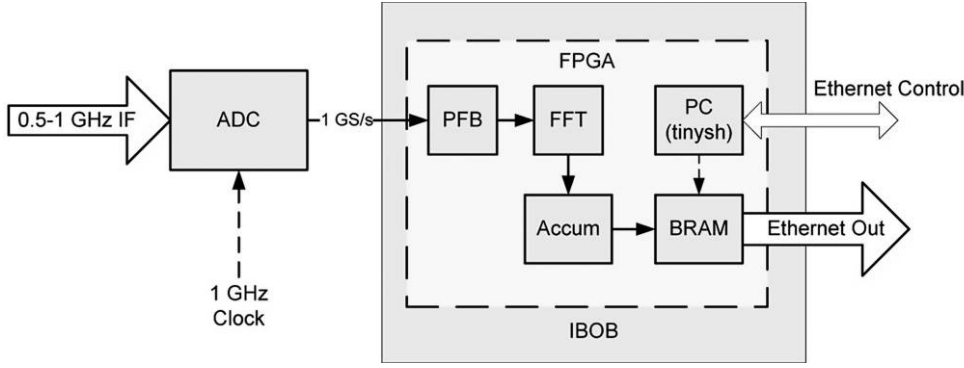
Fig. 1.—The block diagram of the SK spectrometer. The IBOB FPGA does the digital processing by implementing a four-tap PFB and a 4096-point FFT, accumulates the PSD samples, and stores them in block memory (BRAM). The FPGA also implements a PowerPC that runs a rudimentary form of Linux called the tiny shell (tinysh), with which the control system can set registers for controlling the spectrometer and which supervises the broadcasting of the accumulated data via fast Ethernet.

2. SPECTRAL KURTOSIS ALGORITHM

Before describing the hardware implementation, we briefly describe the practical aspects of the SK algorithm, which was developed and described in more detail by Nita et al. (2007) and extended by Nita & Gary (2010). This simple algorithm offers a powerful means of identifying certain kinds of RFI, as we will show. A related statistical quantity, time-domain kurtosis (TDK), is well known as a test for departures of a series of measurements from a Gaussian distribution, and it has been applied successfully in radio sensing (e.g., Ruf et al. 2006; De Roo et al. 2007). In the spectral domain, given a time series of power spectral density (PSD) estimates in each spectral channel k, an estimator for SK is found by forming sums of power and power-squared:

$$S_1 = \sum_{m=1}^{M} P_k, \qquad S_2 = \sum_{m=1}^{M} P_k^2, \qquad (1)$$

where k is the frequency channel index. Using the same nomenclature as Nita & Gary (2010), the SK estimator, $\widehat{SK}$, for an accumulation of M samples is then

$$\widehat{SK} = \frac{M+1}{M-1}\left(\frac{MS_2}{S_1^2} - 1\right) \qquad (2)$$

and has a value close to unity for accumulated samples distributed according to Gaussian statistics. The variance of the SK estimator from unity was shown by Nita et al. (2007) to be simply

$$\mathrm{Var}(\widehat{SK}) \simeq \frac{4}{M}. \qquad (3)$$

In Nita et al. (2007) and Dou et al. (2009), we assumed that the SK estimator is normally distributed and defined the thresholds (criterion for interpreting the sample as Gaussian noise) as $\widehat{SK} = 1 \pm 3\sigma = 1 \pm 3\sqrt{\mathrm{Var}(\widehat{SK})}$, and conversely, an accumulated spectral channel was considered to have RFI present if

$$|\widehat{SK} - 1| > \frac{6}{\sqrt{M}}. \qquad (4)$$

However, Nita & Gary (2010) have since derived the theoretical PDF of equation (2) and found it to be considerably skewed, with a shape that is highly dependent on the number of accumulated
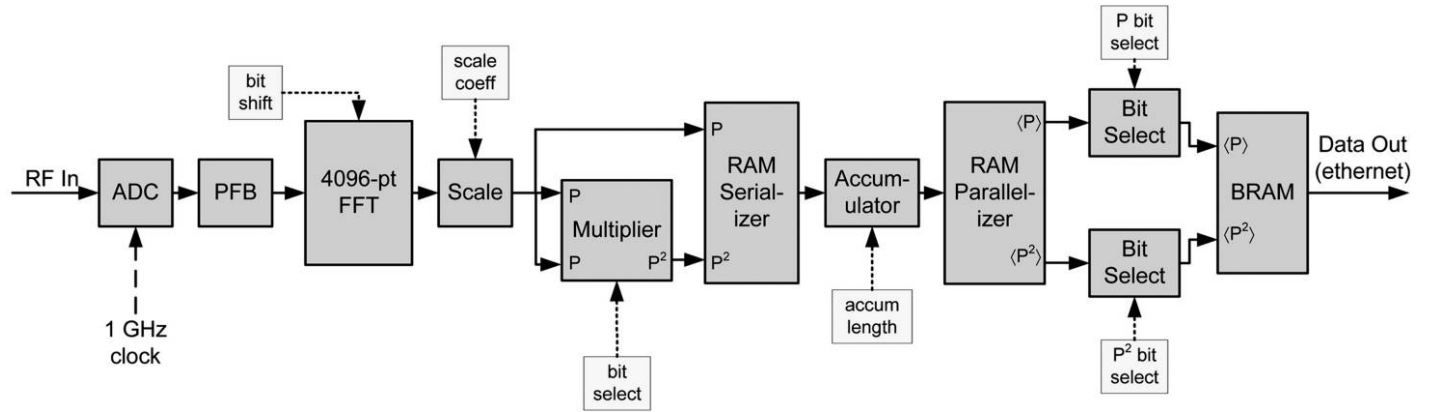

Fig. 2.—The block diagram of the firmware implemented in the FPGA. The light gray blocks represent register settings as described in the text.
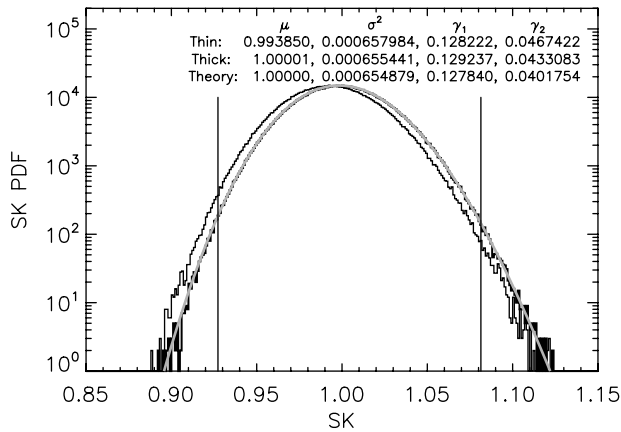
FIG. 3.—Histogram of SK values for KSRBL data when pointed at the Sun in the frequency band 7500–8000 MHz (no RFI). *Thin line*: histogram of data with bit-select set to 1, resulting in truncation. *Thick black line*: histogram of data with bit-select set to 0. *Thick gray line*: overlay of theoretical curve. The vertical lines show the thresholds given by equation (6). The moments μ, σ^2, γ_1, and γ_2 for the three curves are given in the inset.

samples M. A main result of that article is that, in the limit of large M, the first four moments of the PDF (mean μ, variance σ^2, skew γ_1, and kurtosis excess γ_2) are

$$\mu = 1; \qquad \sigma^2 \simeq \frac{4}{M}; \qquad \gamma_1 \simeq \frac{10}{\sqrt{M}}; \qquad \gamma_2 \simeq \frac{246}{M}. \quad (5)$$

Although three of these moments approach the normal distribution behavior as $1/M$, the skewness decreases only as $1/\sqrt{M}$.

Nita & Gary (2010) describe the setting of thresholds for the general case. For the purposes of this article we limit discussion to our particular case of $M = 6104$. Recall that M is the number of independent spectra (PSD samples) accumulated in the KSRBL dump time ($T = 25.001984$ ms). Given our 1 GHz sample rate and 2048 spectral channels (4096 time samples, or $t_{\mathrm{PSD}} = 4096$ ns), we have $M = T/t_{\mathrm{PSD}} = 6104$. As discussed in Nita & Gary (2010), if we wish to keep the false-alarm rate for both lower and upper thresholds at the equivalent $\pm 3\sigma$ values for a normal distribution ($\approx 0.135\%$), the correct lower and upper thresholds equivalent to equation (4) must be chosen as $1 - 5.6799/\sqrt{6104}$ and $1 + 6.3596/\sqrt{6104}$, i.e., samples should be flagged as RFI if

$$\widehat{SK} < 0.9273 \quad \mathrm{or} \quad \widehat{SK} > 1.0814 \quad (\text{for } M = 6104). \quad (6)$$

The reader is referred to Nita & Gary (2010) for guidance in setting thresholds for other values of M.

We note that a key requirement in using the SK algorithm is that the power-squared in equation (1) must be formed on individual PSD estimates prior to summation (accumulation). Accumulating PSD estimates before squaring changes the statistics and invalidates their use for this algorithm. For that reason, the
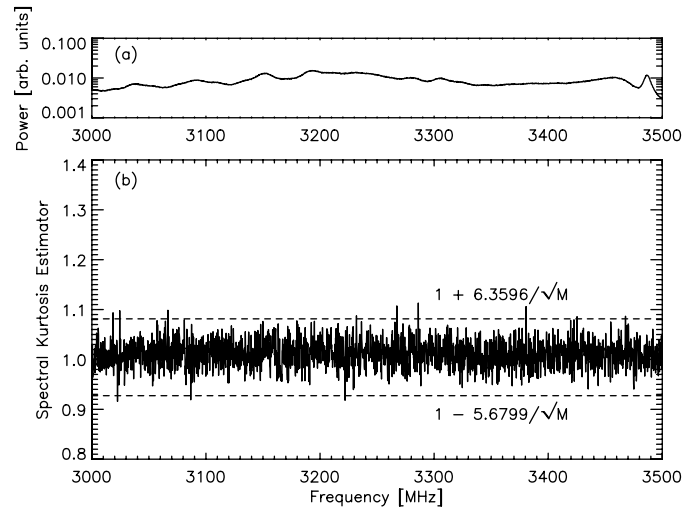

FIG. 4.—SK estimator for a snapshot spectrum with no RFI in the frequency band 3000–3500 MHz. (*a*) The snapshot spectrum itself, showing S_1 for an accumulation of $M = 6104$ PSD estimates. (*b*) The corresponding SK estimate, with dashed lines showing the equation (6) thresholds for Gaussian noise.

SK algorithm requires that individual samples from the fast Fourier transform (FFT) be squared in the high-speed digital electronics, where, for a total power spectrometer of the type described here, the sum S_1 is the primary data (the accumulated power spectrum $M\langle P \rangle$), while the sum S_2 (the accumulated power-squared spectrum $M\langle P^2 \rangle$) is an additional data product to be output for later application of the SK algorithm. In principle, the calculation of $\widehat{SK}$ and generation of flags, equation (6), can be done in hardware, and the flags either applied in hardware or read out for later application. This is not done in our implementation, however, because our application is observations of solar bursts, some types of which can mimic RFI (see Nita et al. 2007). Also, saving both S_1 and S_2 for each accumulation allows the application of multiscale SK, which will be defined in § 6.1.

3. BOARD IMPLEMENTATION

A block diagram of the relevant part of the system is shown in Figure 1. The CASPER analog/digital converter (ADC) board that we use employs a dual 1 GS s^{-1} (gigasample per second) 8 bit digitizer based on the Atmel/ev2 AT84AB001B ADCs. The spectrometer uses only one of the two available ADC inputs, sampled at a clock speed of 1 GHz. The intermediate frequency (IF) input to be sampled has a bandwidth of 500 MHz, ranging from 0.5 to 1 GHz. Hence, the IF is undersampled relative to the Nyquist criterion, which performs an implicit down-conversion to the DC–500 MHz band. The IF power level is maintained at -14 dBm by an upstream analog system that employs automatic leveling control (ALC). This power level results in at least a 7 bit range for the sampled data. The ADC supplies its digital data to a CASPER IBOB, equipped with a Xilinx Virtex-II Pro FPGA. The FPGA firmware is developed using
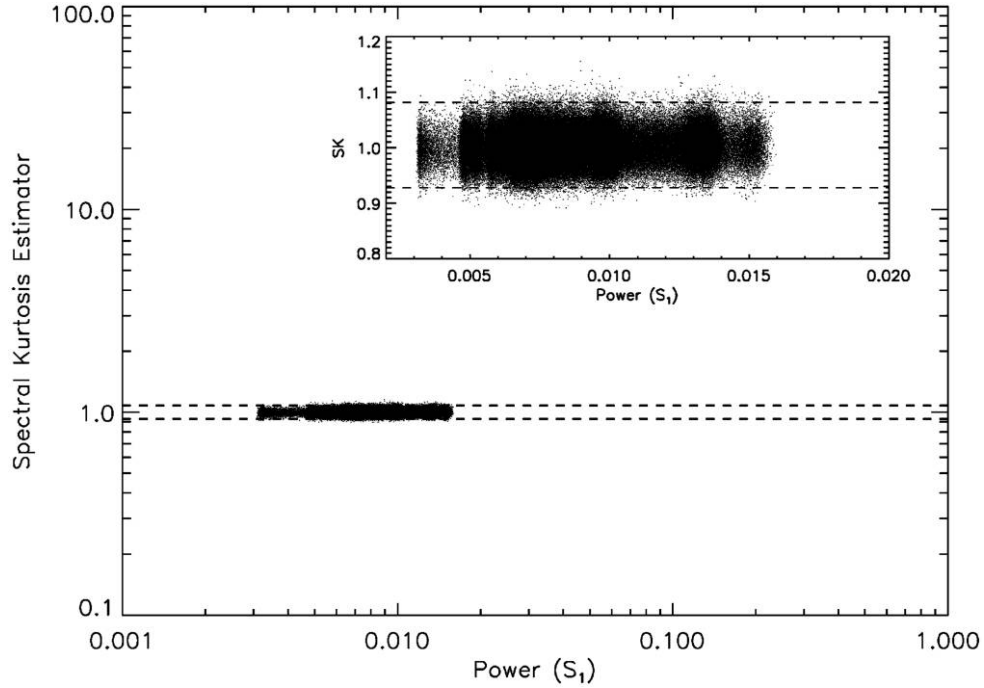
FIG. 5.—Plot of SK estimator vs. S_1 for 150 accumulated spectra with no RFI. The horizontal dashed lines indicate the thresholds for Gaussian noise given by equation (6). This plot is scaled for consistency with subsequent plots, so the zoomed view inset is included to provide better visibility.
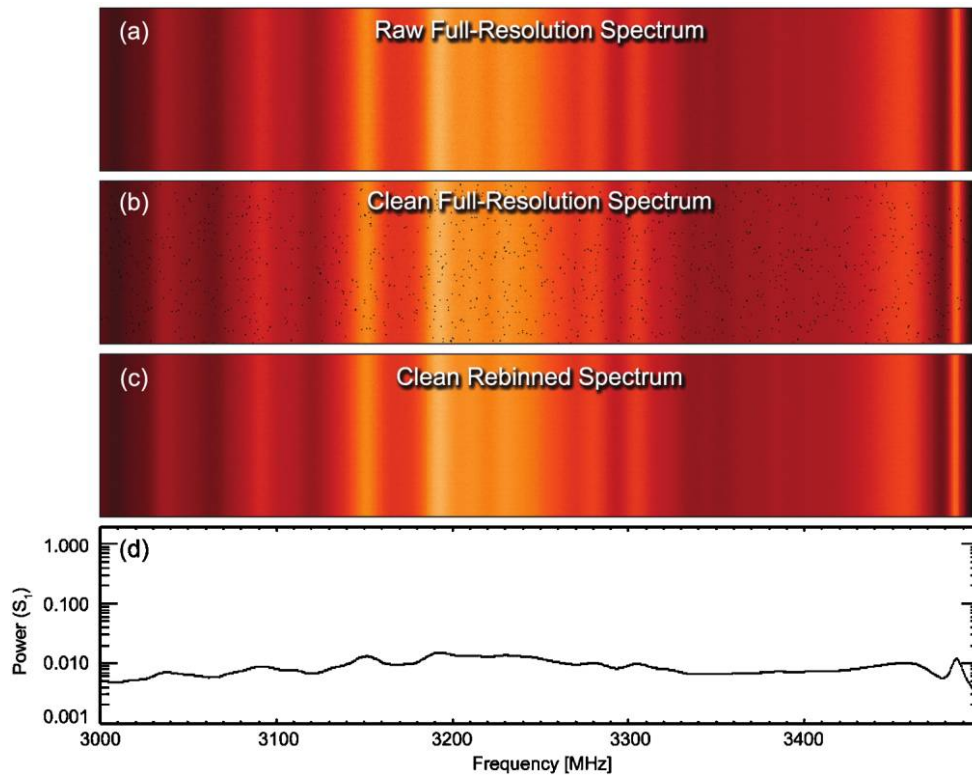

FIG. 6.—(a) The 150 full-resolution snapshot spectra (2048 frequency subchannels), shown as a dynamic spectrum. (b) The clean dynamic spectrum, again with 2048 frequency subchannels, after applying the SK flags. Bins where the SK estimator exceeds the threshold are black. (c) The same as in (b), but now rebinned to 512 subchannels by averaging 4 adjacent frequency bins. (d) Average spectrum formed by summing over the 150 snapshot spectra.
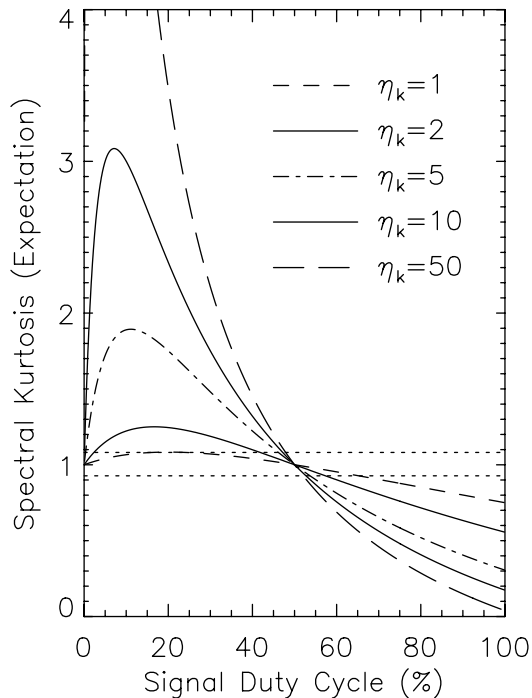
FIG. 7.—SK estimator vs. duty cycle of transient RFI for various $S/N = \eta_k$. The different line styles show the theoretical expectation for η_k ranging from 1 to 50. The horizontal dotted lines represent the detection thresholds for $M = 6104$ PSD estimates.

the software development toolflow created by the CASPER group and their collaborators, which in turn is built on the Xilinx-supplied blocks and Matlab Simulink.

Several firmware designs have been tested during the development of the spectrometer. The baseline design uses a four-tap polyphase filter bank (PFB), which is a modification to the familiar windowing function commonly applied to time-domain signals to improve spectral resolution. This filtered signal then feeds a 4096-point FFT, which results in 2048 frequency channels (244 kHz channel width). Variations on this design include either no FFT prefilter or a two-tap PFB and varying the number of channels from 1024 to 4096 channels. Dramatic differences in spectral resolution are found in going from no prefilter to the two-tap PFB but minor improvement in going to a four-tap PFB. The final version employed by KSRBL, however, does use the four-tap PFB. Likewise, RFI excision is improved in going from 1024 to 2048 channels, but only minor improvements are seen in going to 4096 channels, as will be described later. A C program running in the FPGA's embedded PowerPC handles the broadcast of the data onto the 100 Mbit s^{-1} Ethernet, and we found that it was not able to successfully transmit 4096 channels of S_1 and S_2 within the desired 25 ms accumulation time. We note, however, that the IBOB is equipped with 10 Gbit Ethernet (10 GigE, or XAUI) ports, which our design does not use but which could be used for 4096-channel output. We elected not

to implement that because 2048 channels is enough for reliable RFI excision and we wanted to avoid the additional data volume resulting from 4096 channels. We also note that the 4096channel design leaves about 20% of the FPGA resources unused, but we did not investigate whether a 10 GigE block would fit within the remaining resources. For specificity, then, we limit further description in this section to the four-tap, 2048-channel design.

Figure 2 shows a simplified diagram of the Simulink blocks used in the design. Because this FPGA design implements a fixed-point algorithm, the bit width needs to be carefully selected to avoid the overflow or loss of precision of the data. Scaling can be applied at various key points in the design, indicated as light gray blocks representing register settings. The 8 bit digitized data out of the PFB, after passing through the FFT block, retains 36 bits of resolution in order to maintain the validity and precision of the data. For the convenience of calculating and transferring data, we set the dynamic range of data to be 32 bits. The scale-coeff register is used to select which 32 bits to feed into the vector accumulator. In our design, the multiplier block that calculates P^2 for accumulation of S_2 takes 16 bit input and outputs 32 bit-wide values; hence, this block requires a bit-select register to set which 16 bits are selected from the 32 bit-wide power data, with bit-select = 0, 1, or 2 specifying the least significant, middle, or most significant 16 bits, respectively. This design choice was made in order to optimize high-speed operation and maintain low-power consumption by using the Virtex-II Pro embedded 18 bit by 18 bit multiplier. A multiplier with input bit width of more than 18 bits would have to be implemented with multiple embedded multipliers, which could not satisfy the very strict clock time constraint (250 MHz in our case). The bit-select is implemented as a multiplexer, which selects a 16 bit data range at various offsets from the least significant bit (LSB). The P^2 output is then 32 bits wide, matching the P data width. After vector accumulation, the data stream becomes 48 bits wide, which limits the number of accumulations to 2^{16}. Finally, the P bit-select and P^2 bit-select register settings are used to select which 32 bits are to be extracted from the 48 bit data to maximize precision without losing the validity of the data through overflows. All of these scaling registers must be set correctly to avoid overflow or excessive loss of precision. Inappropriate scaling would cause the SK estimator to deviate from unity. This will be discussed in the next section.

4. VERIFYING SK OPERATION

It is of interest to compare the histogram of SK values for real solar data with the $\widehat{SK}$ PDF theoretically determined by Nita & Gary (2010) to verify that the KSRBL hardware implementation matches theory. In fact, our first test after achieving the results of Nita & Gary (2010) matched well in shape but was found to be shifted by a small amount, as shown by the histogram plotted as a thin line in Figure 3. The histogram shown by the thin line was taken on 2010 January 5, when the bit-select register in Figure 2 was set to 1, and represents about 10^6 individual SK
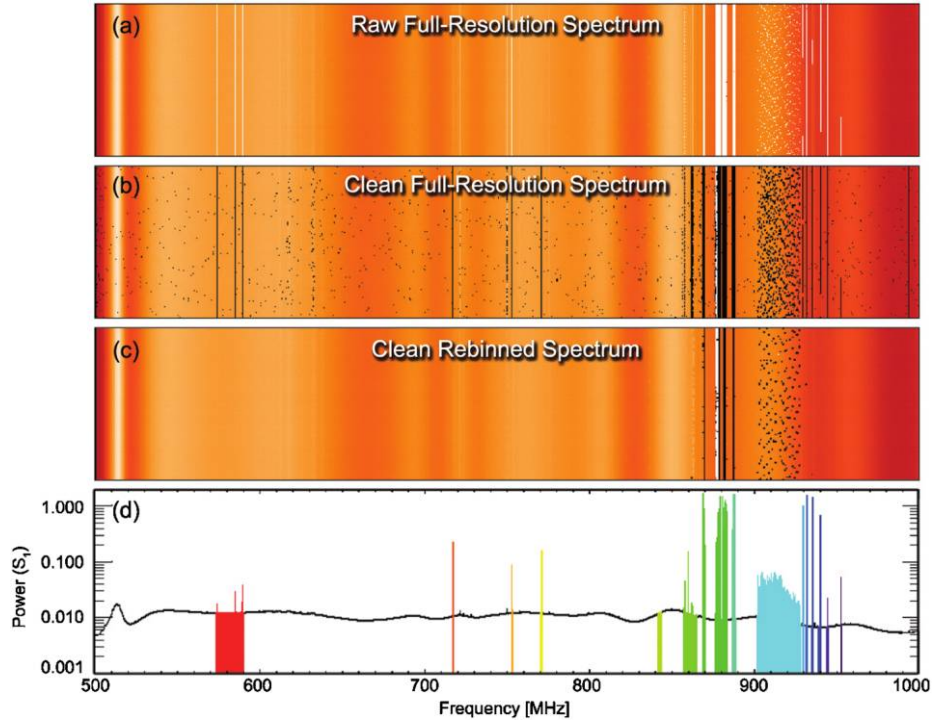
FIG. 8.—Same as in Figure 6 but now for a band (500–1000 MHz) with considerable RFI. Spectral data in selected frequency ranges containing RFI are color coded in (*d*) for later reference.
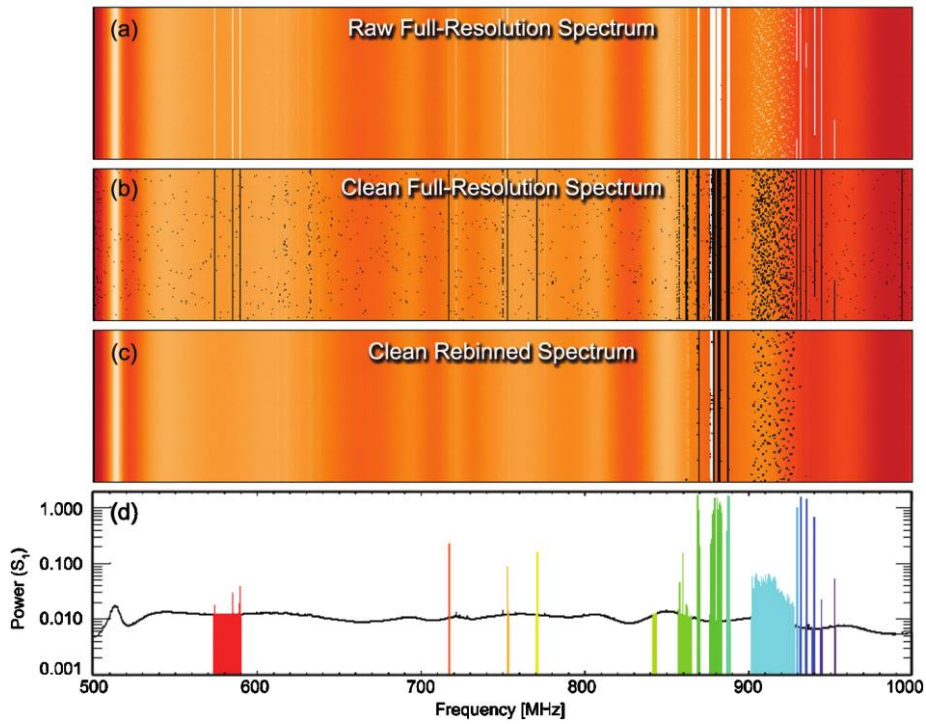

FIG. 9.—Same as Figure 5 but for the 500–1000 MHz band containing considerable RFI. Each dot represents a single frequency–time bin in Figure 8, and the dots for selected frequency ranges are color coded with the same colors as in Figure 8*d*. The colors in this and later figures represent frequency range only, whether or not RFI is present.

values. The histogram shown by the thick line was taken on 2010 January 7 after the bit-select register was lowered to 0. The gray curve shows the theoretical PDF derived by Nita & Gary (2010), which matches well with the thick-line histogram.

This illustrates the effect of inappropriate scaling of the $\langle P \rangle$ and $\langle P^2 \rangle$ output, affected by the scaling parameters indicated by the light gray blocks in Figure 2. It should be clear that the dynamic range of $S_2 = M\langle P^2 \rangle$ is generally larger than that of $S_1 = M\langle P \rangle$; hence, the 32 bit integer output values for each must be scaled to avoid both excessive loss of precision (truncation) at the LSB and overflows (clipping) at the most significant bit (MSB). In practice, the overflows occur in spectral channels with strong RFI, so some overflow in S_2 may be permissible because those channels are to be flagged anyway. More subtle is the effect of inappropriate scaling that causes truncation at the LSB. It is clear from equation (2) that excessive truncation of S_1 caused by setting register P bit-select too high will decrease the magnitude of S_1 and hence will cause the SK estimator level to be biased to slightly greater than unity. Alternatively, excessive truncation (decrease) of S_2 will cause the SK estimator to be biased to slightly less than unity. This latter case is the cause of the shift of the entire PDF curve shown by the thin line in Figure 3. Experimentation with the bit scaling registers showed that the register labeled bit-select in Figure 2 was set too high, thus, truncating the power too much prior to squaring. Setting it to 0 (i.e., selecting the lowest 16 bits) results in the curve shown by the thick line in Figure 3, which matches almost exactly the theoretical curve (*gray line*). The values of the mean, variance, skew, and kurtosis excess for the three curves are shown in the inset table in the figure. It is clear that the higher moments that characterize the shape of the PDF are not affected by minor S_2 truncation, but the shift of the mean by 0.0062, while seemingly small, has a significant impact on RFI excision. Many more samples of the shifted curve are incorrectly flagged below the lower threshold (*left vertical line*) while fewer samples are flagged above the upper threshold (*right vertical line*). Once the scaling parameters are determined for the system, they do not change with time so long as the IF power level is held fixed by the preceding analog electronics. It is common practice to use ALC to ensure constant IF level to the digitizer, and indeed this is the case for KSRBL, as described in Dou et al. (2009).

We note in passing that second-generation hardware, the ROACH (reconfigurable open architecture computing hardware) board, has since become available with more FPGA resources. We have implemented the SK spectrometer on ROACH with full-resolution calculation of S_2, and initial tests show that it does not suffer from the scaling issues illustrated in Figure 3. We conclude that the KSRBL SK spectrometer matches the theory of Nita et al. (2007) and Nita & Gary (2010) to high accuracy when properly scaled. It remains only to demonstrate the utility of SK for real-world RFI excision.

5. USING SK TO CLASSIFY RFI

5.1. Example of a Band with no RFI

To introduce some concepts we will need in our discussion of the performance of the SK spectrometer, let us first look at some examples of data with no RFI. As described in Dou et al. (2009), each of the four spectrometer boards of the KSRBL instrument samples a 500 MHz band in the range 0.5–18 GHz. Although most of these bands contain some RFI, a few bands appear to be completely RFI-free. In Figure 4 we show a single, 25 ms accumulated spectrum snapshot of one such band, the band from 3000 to 3500 MHz. The upper panel shows the accumulated power, $S_1 = M\langle P \rangle$, as a function of frequency, taken while the dish is pointing at blank sky. The spectrum is relatively flat, with smooth variations due to the nonuniform response of the system over the bandpass. Each individual PSD estimate is obtained from 4096 1 ns time samples, and $M = 6104$ such estimates are accumulated for a total duration of $6104 \times 4096 = 25,001,984$ ns. The bottom panel shows the SK estimator, equation (2), along with horizontal lines representing the thresholds of equation (6). It is clear that the SK estimator indeed falls nicely between the limits given by equation (6), with only a few spectral channels statistically falling slightly outside the thresholds.

We will find it extremely useful to show the behavior of the SK estimator for larger blocks of data containing many such snapshot spectra by plotting the SK estimator versus S_1. Figure 5 shows this for a set of 150 instantaneous spectra, together with horizontal dashed lines showing the limits given by equation (6). This log–log plot shows that the SK estimator is independent of power level, which is a key property formally proven by Nita & Gary (2010). The plot is shown with the same axis ranges as SK versus S_1 plots we will show later, which do include RFI, but for improved visibility we provide a zoomed view in the inset to the figure. As expected for a band with no RFI, almost all of the points fall within the thresholds. The percentage of points exceeding the upper threshold in this example is 0.157%, while that below the lower threshold is 0.150%. These are close to the target values of 0.135% for which the thresholds are set from the theory of Nita & Gary (2010).

Figure 6 shows one additional representation of the data, the dynamic spectrum (or waterfall plot) corresponding to a block of data. Three versions of the dynamic spectrum are shown, along with a line plot, Figure 6*d*, showing the average spectrum. The top panel, Figure 6*a*, is the raw dynamic spectrum at the full 244 kHz resolution, including any RFI, if present (there is none in this example). The second panel, Figure 6*b*, shows the full-resolution clean spectrum, after spectral channels exceeding the SK thresholds have been blanked. This is done by forming the SK estimator, comparing with the fixed thresholds, and forming a flag array that is then applied to S_1. Finally, the third panel, Figure 6*c*, shows the clean spectrum at the target resolution of ≈1 MHz, formed by averaging (rebinning) in frequency by a
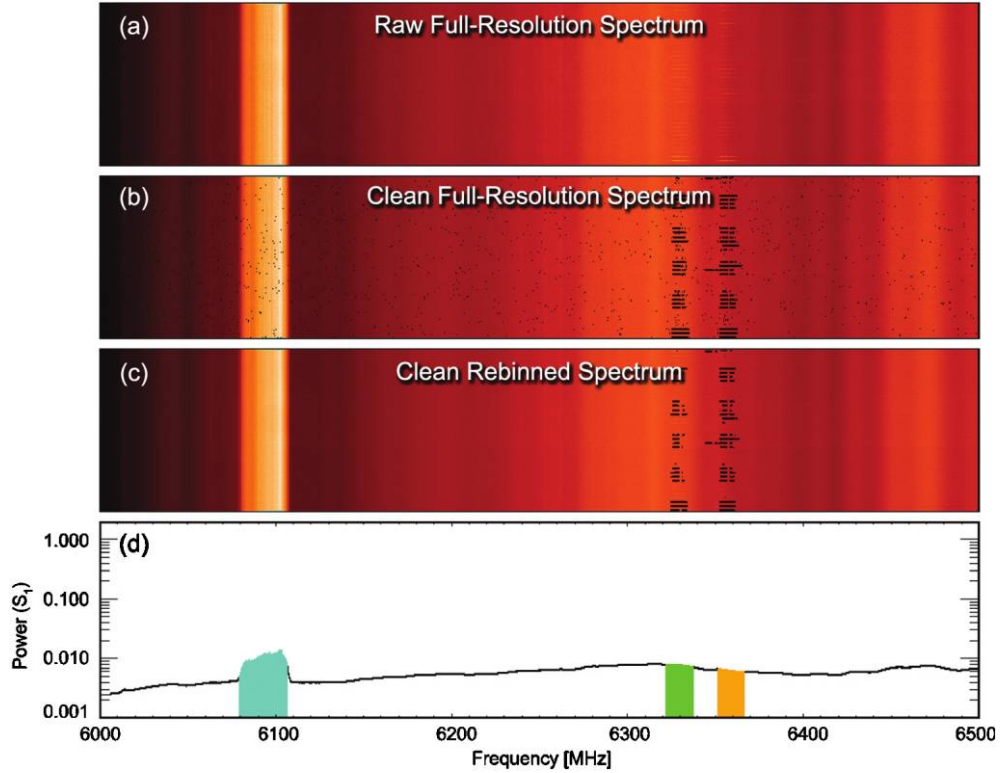
FIG. 10.—Same as Figure 6 but now for the band 6000–6500 MHz. Strong RFI centered on 6093 MHz is not detected, while weaker but intermittent RFI above 6300 MHz is detected.
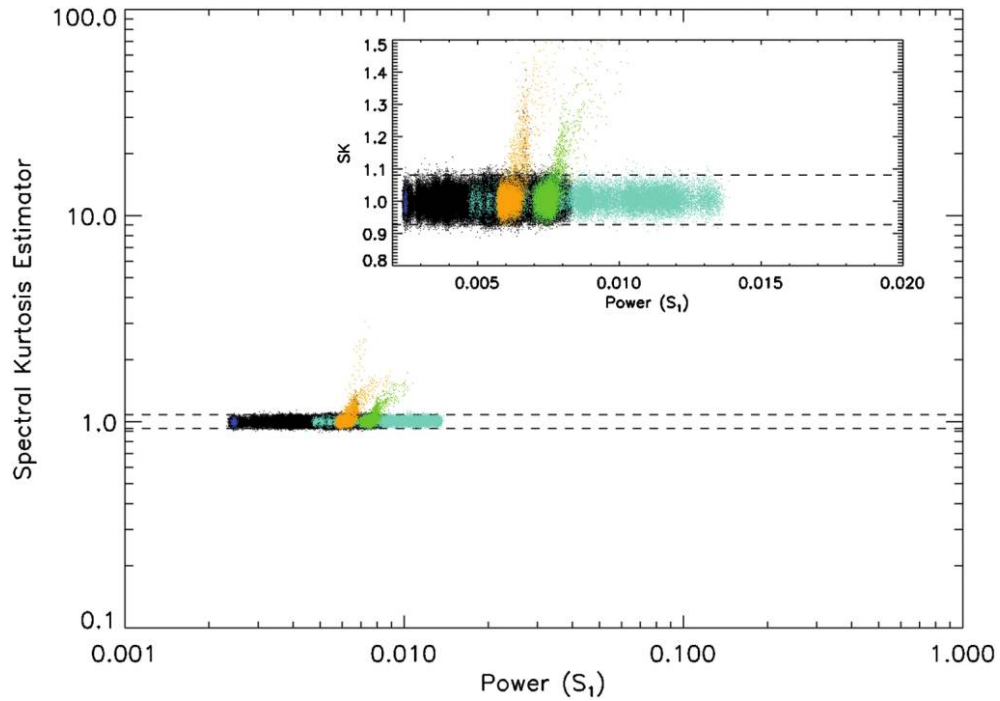

FIG. 11.—Same as Figures 5 and 9 but color coded for the band shown in Figure 10. The points color-coded orange and green correspond to the RFI near 6330 and 6365 MHz, respectively, and are well detected despite their relative weakness. However, the points color-coded cyan, corresponding to a digital data link, fall exactly in the window and are not flagged.

factor of 4. In this procedure, four adjacent frequency bins are added, and then the sum is divided by the number of unflagged bins. For this example band that contains no RFI, the application of SK flags is trivial and obviously not needed, but later examples will show the importance of the algorithm for removing RFI.

5.2. SK Estimator for Non-Gaussian Signals

To understand the behavior of the SK estimator for non-Gaussian signals (i.e., RFI), we examine a few more results from the theory introduced by Nita et al. (2007), slightly modified to conform to the new definition in equation (2). One type of RFI we have simulated is a continuous-wave (CW) signal of constant amplitude, which can be used to simulate transient RFI by considering its presence or absence with some duty cycle d. Consider M contiguous PSD estimates of which R are contaminated by RFI at signal-to-noise ratio (S/N), η_k, where k is the spectral channel index. The duty cycle is therefore defined as $d = R/M$. For this case, the SK estimator expectation value (Nita et al. 2007) is

$$\widehat{SK} = \frac{M+1}{M-1}\left[1 + \frac{(1/d - 2)\eta_k^2}{(1/d + \eta_k)^2}\right]. \qquad (7)$$

Note several interesting properties of the SK estimator indicated by this equation:

1. For $d = 1/2$ (50% duty cycle), the estimator is always 1;
2. For $d < 1/2$ (highly intermittent RFI), the estimator is above 1;
3. For $d > 1/2$ (more continuous RFI), the estimator is below 1.

This behavior is illustrated in Figure 7, adapted from Nita et al. (2007), which shows SK estimator versus duty cycle for various $1 < \eta_k < 50$, for our case of $M = 6104$. This shows that the S/N of the contaminating RFI in our case must be more than about unity, and that there is a window near 50% duty cycle where the SK algorithm (and the related algorithm based on TDK; De Roo et al. 2007) is blind no matter how great the S/N. This window is small for strong RFI but grows larger with smaller S/N. For our case of $M = 6104$, e.g., the blind window range is $0.42 < d < 0.57$ for S/N ≈ 2.

We are now ready to examine the performance of the SK algorithm for identifying RFI and to understand the types of RFI most amenable to excision by this method. Figure 8 shows the band from 500 to 1000 MHz, which has considerable RFI. The band contains strong fixed-frequency RFI, as well as RFI from roughly 903 to 928 MHz that skips rapidly in frequency and time. Some channels above 928 MHz are not continuously present but turn on and off. Selected frequency ranges are color coded in Figure 8d for later discussion. Close inspection of Figure 8b shows that although most of the RFI is successfully flagged, some parts of the RFI near 877 MHz escape detection. Although not apparent in the figure, a few isolated points in the 903–928 MHz range also slip through.

The SK estimator versus S_1 plot introduced in Figure 5 is shown for the 500–1000 MHz band in Figure 9, where now the power of this visualization is apparent. We see that more or less continuous RFI of constant amplitude (*various shades of blue and red*) results in clusters of points below the SK ≈ 1 range (*dashed lines*), as expected from Figure 8. Some RFI channels are continuously present but vary in amplitude to form diagonal traces or chains of points. Meanwhile, the highly intermittent RFI between 903 and 928 MHz (*cyan color*) forms a fountain of points lying above the SK ≈ 1 range. A few isolated points of the intermittent RFI just happen to fall between the dashed lines and hence escape detection. Note, however, the green points with $S_1 \approx 0.2$–0.3, which correspond to the unflagged RFI at 877 MHz in Figure 8. These points fall within the SK ≈ 1 range, indicating that this RFI is somehow mimicking Gaussian noise statistics.
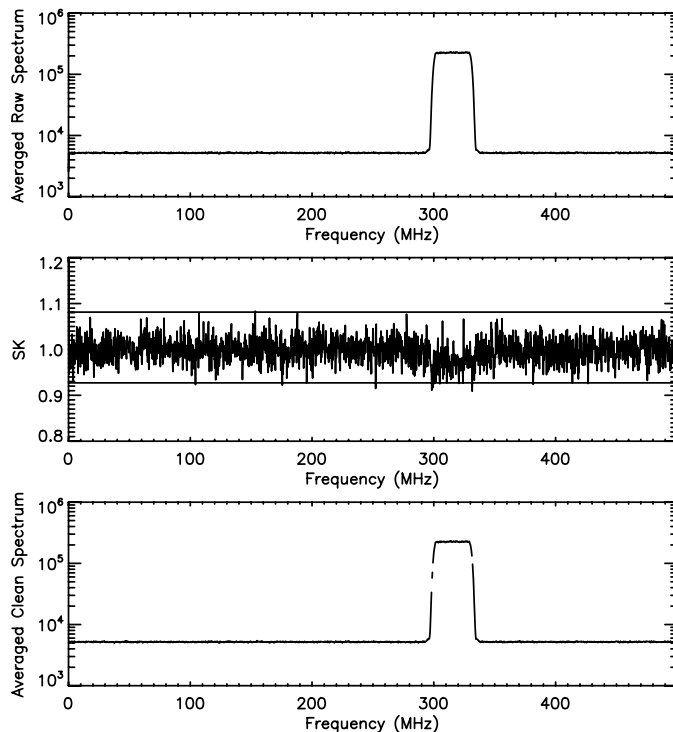

FIG. 12.—Results of the LabVIEW simulation of QPSK-modulated digital data representing 32 channels separated by 1 MHz. The spectrum in (*a*) shows the digital data from 300 to 330 MHz and that in (*b*) shows the SK estimator derived from an accumulation of $M = 6104$ spectra. In this simulation, the SK estimator is slightly biased below unity but not enough to result in its elimination, as shown in the clean spectrum in (*c*).
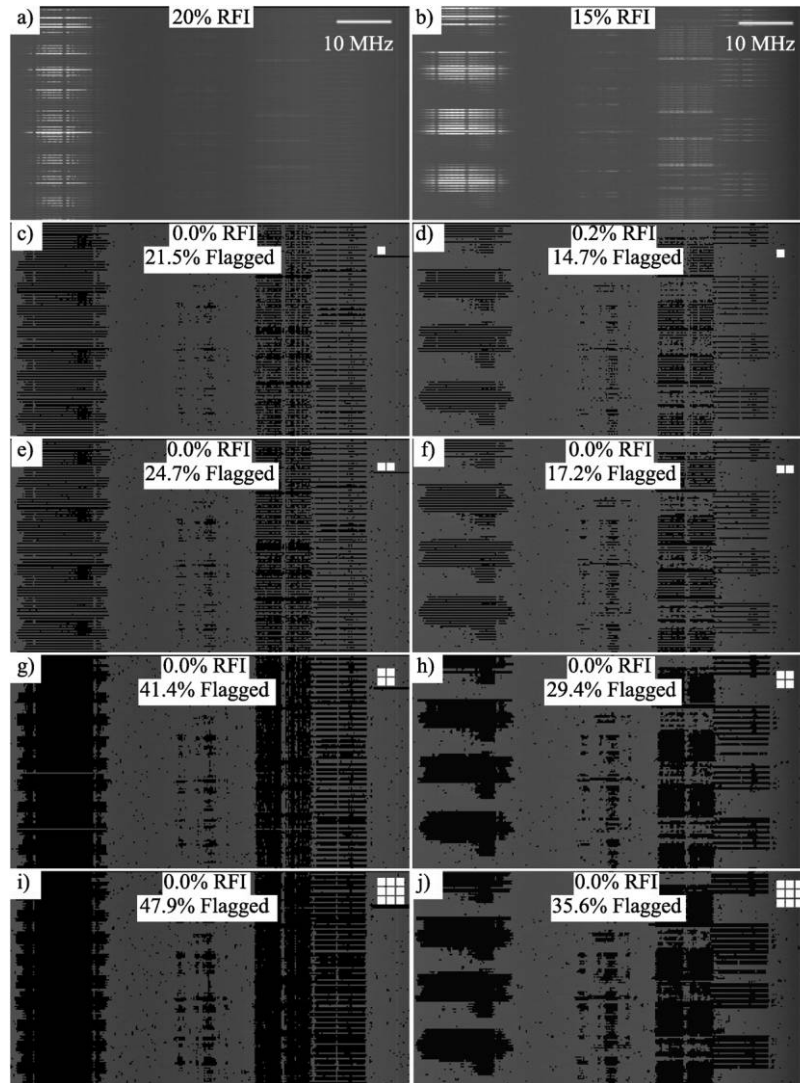
FIG. 13.—A comparison of quality of RFI excision for various choices of parameters, using 73 MHz bandwidth from 2402 to 2475 MHz. The left column shows data taken with an IBOB with 4096 channels, while the right column shows identical data taken with an IBOB with 2048 channels. The top row shows raw data with no flagging and includes a bar representing a 10 MHz frequency range. The second row shows data with SK applied, and the remaining rows show data with various multiscale SK applied. The multiscale pattern for each column is shown by the icon in the upper right of each panel, representing $m = n = 0$; $m = 1$, $n = 0$; $m = n = 1$; and $m = n = 2$.

5.3. Digital RFI

The SK versus S_1 plot allows us to classify and investigate those types of RFI that somehow are not detected by the SK algorithm. The 877 MHz RFI signal has not been identified, but a clue to what might be happening is provided by the band 2000–2500 MHz, which contains the digital radio signals from the XM and Sirius satellites. The band has both intermittent and continuous strong RFI from other sources that is well detected and removed, as we will show later in § 6.1. However, the digital radio RFI near 2330 MHz is not detected, and indeed we find that its $\widehat{SK}$ values fall exactly within the Gaussian noise thresh-

olds. This suggests that digital signals may be a problem for the SK algorithm.

This supposition is confirmed in Figures 10 and 11, which show the band 6000–6500 MHz. RFI color-coded green and orange is well identified and removed, even though it is extremely weak in the integrated total power plot (Fig. 10d), but the stronger, cyan-colored, 30 MHz-wide RFI signal centered at 6093 MHz is completely untouched. This signal is identified as a digital data link between Lone Pine and Big Pine, CA. Another data link between the OVRO site and the CARMA (Combined ARray for Millimeter Astronomy) high site at 5800 MHz (not shown) is also found to escape the SK algorithm. Such data
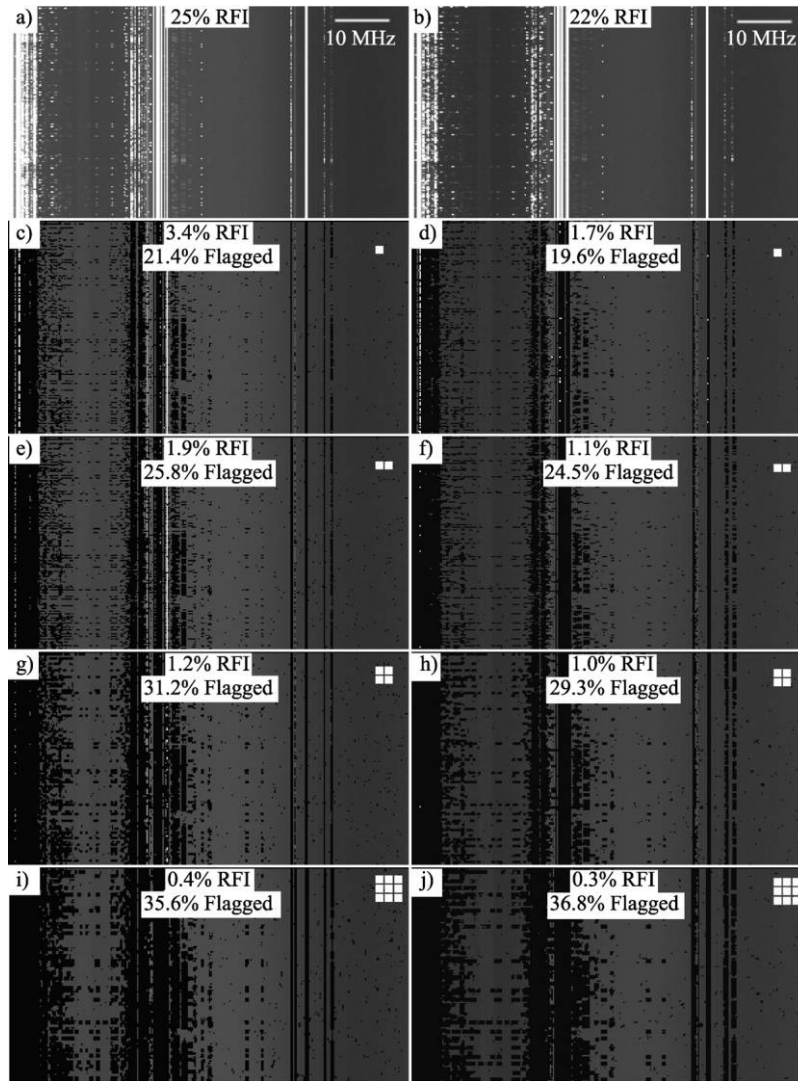
FIG. 14.—Same as Figure 13 but now showing 73 MHz bandwidth from 2000 to 2073 MHz.

links use various types of digital modulation, e.g., amplitude shift keying (ASK) and/or phase shift keying (PSK), to encode the data (see, e.g., Proakis 2001).

We examined the SK algorithm's response to such signals by simulating in LabVIEW a simple QPSK (quadrature PSK) modulated signal with 32 carriers separated by 1 MHz, encoded with randomly generated bits. QPSK works by applying a phase modulation to the carrier that places bits in a set of predefined regions of a phasor diagram (the so-called constellation; four regions, or symbols, were used in our simulation). The results of our simulation, shown in Figure 12, reveal that, indeed, the SK estimator is only slightly affected by digital RFI. Presumably this phase shift modulation gives the data-laden carrier noiselike characteristics that distribute power in a manner similar to a normal distribution. Commonly used larger constella-

tions of 8, 16, or more regions will likely make detection even more difficult.

6. STRATEGIES TO IMPROVE RFI ELIMINATION

6.1. Multiscale Spectral Kurtosis

One of the advantages of retrieving both S_1 and S_2 information from the hardware (as opposed to calculating SK in hardware and retrieving only the flags) is that these sums can be combined in alternative ways to address the algorithm's blindness to RFI with a duty cycle near 50% (e.g., the fountain of points in Fig. 9 include a few high-power points that fall between the dashed lines). Consider a frequency–time bin b_{ik} in a dynamic spectrum at full resolution that contains RFI at duty cycle $d_{ik} = 0.5$ and an adjacent time bin that contains

RFI at some other duty cycle $d_{i+1,k}$. Because of the linearity of S_1 and S_2, new values $S_1 = S_{1,ik} + S_{1,i+1,k}$ and $S_2 = S_{2,ik} + S_{2,i+1,k}$ can be formed, and a new SK can be calculated for the combined bins, whose duty cycle is now $d = (d_{ik} + d_{i+1,k})/2$. If $d_{i+1,k} = 0$ (no RFI) or $d_{i+1,k} = 1$ (continuous RFI), then the combined duty cycle is 0.25 or 0.75, respectively. In either case, the combined SK is likely to exceed the thresholds and could be used to flag bin b_{ik}. This is likely to be a good strategy in the case of highly intermittent RFI such as that shown in Figures 8 and 9. Alternatively, for more continuous narrowband RFI one might use instead an adjacent frequency channel, i.e., $S_1 = S_{1,ik} + S_{1,i,k+1}$ and $S_2 = S_{2,ik} + S_{2,i,k+1}$. In general, we define multiscale SK for bin i, k as the determination of alternative values of SK over multiple macrobins of size $m \times n$, i.e.,

$$S_{1,ik}^{mn} = \sum_{j=0}^{m} \sum_{l=0}^{n} S_{1,i+j,k+l}, \qquad S_{2,ik}^{mn} = \sum_{j=0}^{m} \sum_{l=0}^{n} S_{2,i+j,k+l}, \quad (8)$$

each of which will have duty cycle

$$d_{ik}^{mn} = \frac{1}{(m+1)(n+1)} \sum_{j=0}^{m} \sum_{l=0}^{n} d_{i+j,k+l}. \quad (9)$$

These multiple values of SK, which we refer to as multiscale SK, can be thought of as votes that bin b_{ik} contains RFI, and various strategies can be used to tally the votes and decide whether bin b_{ik} should be flagged. Perhaps the simplest is to flag b_{ik} if any of the SK values exceed the thresholds. This strategy will result in more bins being flagged than necessary, but this may be acceptable in order not to permit some RFI from corrupting the data.

In Figure 13, we use the RFI in the 2402–2475 MHz band to illustrate the effectiveness and trade-offs of multiscale SK, for m and n ranging from 0 to 2, and for two spectrograph resolutions, 2048 and 4096 channels, respectively, over 500 MHz. The 73 MHz bandwidth corresponds to 300 or 600 channels, encompassing a region of complex, intermittent RFI. To gauge the effectiveness, in each panel we indicate both the percentage of the displayed spectrum that contains RFI (determined from a simple amplitude threshold) and the percentage that has been flagged at each stage. A panel with 0.0% RFI means that all bins with RFI have been eliminated. The RFI in Figure 13 occupies about 20% of the displayed 4096 channel spectrum (Fig. 13a) and about 15% of the 2048 channel spectrum (Fig. 13b). After application of SK (Figs. 13c and 13d, respectively), nearly all of the RFI is detected and removed. Application of progressively larger multiscale SK in subsequent panels (see icon in upper right of each panel for multiscale SK pattern used) unnecessarily flags more data due to the influence that an RFI-containing

bin has on adjacent bins with good data. For this RFI signal, multiscale SK is actually a disadvantage.

Figure 14 shows the same parameter choices as Figure 13 but now for a case of more continuous RFI occurring in the range 2000–2073 MHz. The RFI occupies about 25% of the 4096 channel spectrum (Fig. 14a) and 22% of the 2048 channel spectrum (Fig. 14b). After application of SK (Figs. 14c and 14d, respectively), most of the RFI is detected and removed, but 3.4% remains in Figure 14c and 1.7% in Figure 14d. Even after application of multiscale SK with $m = n = 2$ (Figs. 14i and 14j), with 10%–12% of good data removed, a small fraction (0.3%–0.4%) of RFI remains. Thus, for the characteristics of this particular RFI, multiscale SK appears to be needed if the goal is to ensure that there is as little surviving RFI as possible. Note that the results of Figures 13 and 14 show little advantage to having higher resolution (4096 channels versus 2048 channels), as we remarked earlier.

For the KSRBL system, we currently apply multiscale SK with $m = n = 1$, as shown in Figures 13h and 14h, which we find eliminates nearly all RFI without flagged excessive good data. This is done in pipeline software for the KSRBL system, enabled by the fact that we dump both S_1 and S_2 from the hardware, but we note that multiscale SK could be implemented in FPGA firmware, given sufficient resources, by simply including additional accumulator buffers, doing a parallel calculation of SK, and performing the appropriate logic to generate (and optionally apply) the flags.

6.2. Occupancy Statistics and Always-On RFI

It is clear that no single algorithm can eliminate all sources of RFI and multiple strategies are needed for practical RFI mitigation. Because the SK algorithm detects RFI in each frequency–time bin independently of others, it is ideal for determining occupancy statistics, i.e., what fraction of time RFI is present in frequency channel k. One simply sums the SK flags. Depending on the science being addressed using the spectrometer data, it may be that frequency channels with too high an occupancy fraction should be permanently flagged. This avoids problems with RFI that may be nearly always present but at an amplitude level where the SK estimator may occasionally fall within the allowed thresholds, thus, intermittently letting RFI through. In other cases, always-on RFI may be present at low S/N, or may be due to digital signals as in Figure 10, such that SK is not a reliable algorithm for detection. Because the RFI in these frequency channels is nearly always present, it makes sense to have a permanent channel mask, determined by the SK algorithm or by other means (e.g., Wang et al. 2009). This strategy then leaves only RFI that changes with time, such as highly intermittent RFI, RFI due to mobile transmitters, overflying aircraft and spacecraft, or RFI due to transmitters that are turned on and off irregularly.

7. CONCLUSIONS

We have described the design and performance of the first operational SK spectrometer, implemented for the KSRBL (Dou et al. 2009). The spectrometer differs from other spectrometers in accumulating both power (S_1) and power-squared (S_2) spectra, which are then used to calculate a SK estimator, $\widehat{SK}$. This represents the first field test of a practical implementation, based on the theory of the SK algorithm developed by Nita et al. (2007) and Nita & Gary (2010) and on CASPER IBOB hardware (Parsons et al. 2008). A key part of the design is to provide multiple levels of scaling of power and power-squared in order to maintain sufficient precision. To aid in classification of RFI types, we have developed an important visual representation of the data, the SK versus $S1$ plot. We find distinct signatures in such plots for continuous and intermittent RFI and have demonstrated that the SK algorithm is particularly well suited to identify and excise intermittent RFI but also works well for certain types of continuous RFI. We have also shown that the QPSK-based encoding techniques for digital communication signals cause the power and power-squared statistics to act very much like Gaussian noise, and hence, such RFI is not easily detectable by the algorithm. It may be that alternative statistical techniques (e.g., Fridman 2001; Fridman & Baan 2001; Fridman 2008) could be devised that would be able to distinguish digital signals from Gaussian noise.

We have examined the performance of our hardware implementation of the algorithm and demonstrated the effect of incorrect scaling of the power and power-squared. In particular, we find that the scaling registers in our implementation must be set carefully to ensure sufficient precision of power-squared, in order for the distribution of SK to match the theoretical PDF, and we have shown how the histogram of calculated SK agrees precisely with theory when optimally set. We note that an alternative implementation that maintains S_1 and S_2 to full precision, and is therefore immune to scaling issues, has been tested using the ROACH hardware platform with its greater FPGA resources.

We have also examined strategies for improving the performance of the algorithm using multiscale SK. We find that multiscale SK can be used effectively as a trade-off because it eliminates additional time–frequency bins containing RFI but at the cost of also eliminating more uncontaminated data. We also suggest the use of a permanent channel mask for removing always-on RFI that may not be entirely eliminated by the SK algorithm.

Given the success of the SK algorithm for the highly problematic intermittent RFI, coupled with its relative ease of implementation, we believe that designers should consider implementing SK in any future spectrometer based on FPGA hardware as a first line of defense against RFI. It can easily be combined with other strategies to combat RFI. In addition to spectrometers, of course, the algorithm can be applied in the F engine of any FX correlator, where the generated flags would be used to flag all correlations involving each affected antenna and channel.

REFERENCES

De Roo, R. D., Misra, S., & Ruf, C. S. 2007, IEEE Trans. Geoscience Remote Sensing, 45, 1938

Dou, Y., Gary, D. E., Liu, Z., Nita, G. M., Bong, S.-C., Cho, K.-S., Park, Y.-D., & Moon, Y. J. 2009, PASP, 121, 512

Fridman, P. A. 2001, A&A, 368, 369

———. 2008, AJ, 135, 1810

Fridman, P. A., & Baan, W. A. 2001, A&A, 378, 327

Nita, G. M., & Gary, D. E. 2010, PASP, 122, 560

Nita, G. M., Gary, D. E., Liu, Z., Hurford, G. J., & White, S. M. 2007, PASP, 119, 805

Offringa, A. R., de Bruyn, A. G., Biehl, M., Zaroubi, S., Bernardi, G., & Pandey, V. N. 2010, preprint (arXiv:1002.1957)

Parsons, A., et al. 2008, PASP, 120, 1207

Proakis, J. G. 2001, Digital Communications (4th ed.; NY: McGraw-Hill)

Ruf, C. S., Gross, S. M., & Misra, S. 2006, IEEE Trans. Geoscience Remote Sensing, 44, No. 3, 694–706

Wang, X., Ge, H.-Y., Gary, D. E., & Nita, G. M. 2009, PASP, 121, 1139