The Astronomical Journal, 166:146 (11pp), 2023 October https://doi.org/10.3847/1538-3881/acf12a

A Search for Extraterrestrial Technosignatures in Archival FAST Survey Data Using a New Procedure

Yu-Chen Wang[1,2,3], Zhen-Zhao Tao[1,4,5], Zhi-Song Zhang[6], Cheqiu Lyu[2,3,4], Tingting Zhang[7], Tong-Jie Zhang (张同杰)[1,4,5], and Dan Werthimer[8,9]

[1] Institute for Frontiers in Astronomy and Astrophysics, Beijing Normal University, Beijing 102206, People's Republic of China; tjzhang@bnu.edu.cn
[2] Kavli Institute for Astronomy and Astrophysics, Peking University, Beijing 100871, People's Republic of China
[3] Department of Astronomy, School of Physics, Peking University, Beijing 100871, People's Republic of China
[4] Department of Astronomy, Beijing Normal University, Beijing 100875, People's Republic of China
[5] Institute for Astronomical Science, Dezhou University, Dezhou 253023, People's Republic of China
[6] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, People's Republic of China
[7] College of Command and Control Engineering, Army Engineering University, Nanjing 210017, People's Republic of China; @seu.edu.cn
[8] Breakthrough Listen, University of California Berkeley, Berkeley, CA 94720, USA; danw@ssl.berkeley.edu
[9] Space Sciences Laboratory, University of California Berkeley, Berkeley, CA 94720, USA
Received 2023 May 1; revised 2023 July 31; accepted 2023 August 14; published 2023 September 6

Abstract

The search for extraterrestrial intelligence (SETI) commensal surveys aim to scan the sky to find possible technosignatures from an extraterrestrial intelligence (ETI). The mitigation of radio frequency interference (RFI) is an important step, especially for the most sensitive Five-hundred-meter Aperture Spherical radio Telescope (FAST), which can detect more weak RFI. In this paper, we propose several new techniques for RFI mitigation and use our procedure to search for ETI signals from the archival data of FAST's first SETI commensal survey. We detect the persistent narrowband RFI by setting a threshold of the signals' sky separation and detect the drifting RFI (and potentially other types of RFI) using the Hough transform. We also use the clustering algorithms to remove more RFI and select candidates. The results of our procedure are compared to the earlier work on the same FAST data. We find that our methods, though relatively simpler in computation, remove more RFI (99.9912% compared to 99.9063% in the earlier work) but preserve the simulated ETI signals, except for those (5.1%) severely affected by the RFI. We also report more interesting candidate signals, about a dozen of which are new candidates that were not previously reported. In addition, we find that the proposed Hough transform method, with suitable parameters, also has the potential to remove the broadband RFI. We conclude that our methods can effectively remove the vast majority of the RFI while preserving and finding the candidate signals that we are interested in.

Unified Astronomy Thesaurus concepts: Search for extraterrestrial intelligence (2127); Astronomy data analysis (1858); Radio astronomy (1338)

1. Introduction

The search for extraterrestrial intelligence (SETI; Cocconi & Morrison 1959; Tarter 2001) aims to answer one of the most profound questions: are we alone in the Universe? Unlike other approaches that search for biosignatures (including but not limited to the products of biological processes; e.g., Roth et al. 2014; Webster et al. 2015), SETI searches for technosignatures of intelligent civilizations. Since the 1960s (Drake 1961), SETI has mainly been carried out in radio observation, striving to find possible signatures of radio emission produced by civilizations that can communicate via electromagnetic signals (e.g., Lebofsky et al. 2019; Sheikh et al. 2020; Zhang et al. 2020; Smith et al. 2021; Gajjar et al. 2021; Ng et al. 2022; Tao et al. 2022; Luan et al. 2023; Ma et al. 2023). Though no rigid evidence of extraterrestrial intelligence (ETI) signals has been confirmed so far, efforts to answer this profound question will not stop.

Though we cannot rule out the possibility of broadband ETI signals, most works on SETI radio observations mainly focus on searching narrowband signals. This is based on the fact that the narrowest known natural radio emission is ∼500 Hz (e.g., Cohen et al. 1987), while narrowband radio signals are commonly used in human communications, are easy to distinguish from astrophysical sources, and can be produced with relatively low energy (Li et al. 2020).

There are generally two observation modes for the SETI radio observations, namely, commensal surveys and targeted observations. Targeted observations focus on preselected objects, usually nearby stars, while the SETI commensal sky surveys observe large sky areas to find candidate ETI signals for later confirmation. For fixed single-dish telescopes, commensal surveys usually use the "drifting scan" observation mode, which utilizes the rotation of the Earth to scan in R.A. An example of a SETI commensal survey is the SERENDIP program (Werthimer et al. 2001), which spent decades searching for narrowband ETI signals at the 305 m Arecibo Observatory in Puerto Rico. Thanks to the expanding data sets of known exoplanets in recent years, much work has been done on targeted SETI observations (e.g., Sheikh et al. 2020; Gajjar et al. 2021; Smith et al. 2021; Tao et al. 2022; Luan et al. 2023). Nevertheless, the SETI commensal surveys are still complements to the targeted observations, since commensal surveys have a few advantages: (1) they search ETI signals in larger sky areas, (2) they have orders-of-magnitude longer observation times, and (3) they are target-agnostic (and therefore might avoid anthropocentric biases in target selection).

As the largest single-aperture radio telescope so far, the Five-hundred-meter Aperture Spherical radio Telescope (FAST; Nan 2006; Li & Pan 2016) provides us with great opportunities to search for extraterrestrial technosignatures (Li et al. 2020; Chen et al. 2021). With its 19 beam receiver, FAST is one of the most sensitive and efficient telescope for the multibeam SETI observation. The first SETI observation with FAST was a drift-scan survey, observed during its commissioning in 2019 July, and the preliminary results were reported in Zhang et al. (2020). FAST's first targeted SETI observation has also been conducted toward 33 selected exoplanet systems (Tao et al. 2022; Luan et al. 2023). FAST will conduct more SETI observations in the future, both targeted searches and commensal surveys.

One of the challenging tasks in the data analysis of the SETI radio observations is the identification and mitigation of radio frequency interference (RFI), signals generated by humans rather than ETI. Considering the large sensitivity of FAST, we can expect more weak RFI in its observational data. Though in common astronomical observations, one can usually directly remove the narrowband RFI, in most SETI campaigns, the expected signal morphology for the ETI signals is also narrowband. In previous work (Zhang et al. 2020) analyzing the first SETI commensal survey of FAST, the Nebula[10] and kNN pipeline were used to mitigate the RFI. Several types of RFI, i.e., zone, drifting, and multibeam RFI, were removed by the pipeline, and the k nearest-neighbor (kNN) algorithm was also used to further mitigate the RFI. Though most RFI was successfully removed in Zhang et al. (2020), it is still meaningful to further improve the algorithms for the RFI mitigation, as well as the candidate selection, e.g., removing more RFI to reduce the work of visual inspection and finding more candidates for later investigation.

In this paper, we present a novel procedure of RFI mitigation for the SETI commensal multibeam radio survey that searches for narrowband ETI candidates. We propose removing the "persistent narrowband RFI" by rejecting frequency bins that contain signals distributed in a large sky area (larger than a threshold); we also propose detecting and removing the "drifting RFI" (i.e., the narrowband RFI that drifts in frequency) by detecting lines on the time–frequency waterfall plot with the Hough transform method. We apply these methods to the same FAST SETI commensal survey data as analyzed in Zhang et al. (2020) and then use the kNN and candidate selection methods to complete the full procedure of RFI mitigation and candidate selection.

The rest of this paper is organized as follows. We describe the RFI removal methods proposed and used in this paper in detail in Section 2, then briefly introduce the data and report the results of RFI removal and candidate selection in Section 3. In Section 4, we explore and discuss more possibilities of our method based on the Hough transform when different parameters are set. We finally conclude and discuss the results in Section 5.

2. Methods for RFI Mitigation

In a SETI commensal survey, the original input data set for RFI removal programs is the record of hits. A "hit" here refers to the information about a potentially interesting signal that has a high signal-to-noise ratio (S/N) in its frequency channel at each moment (see Section 3.1 and Zhang et al. 2020 for details). Each hit consists of information including the time, frequency channel, telescope pointing,[11] etc., which can be used for detecting RFI. Most hits belong to the RFI, while some may belong to interesting candidates. In contrast to targeted observations, we typically expect the commensal survey to extend over a very long period of time, and the recorded data are usually only these hits instead of the complete spectra at each moment. Therefore, we need to analyze the hit data, which requires software different from that usually used for the "filterbank" data.

In this paper, the RFI is removed in three steps, namely, persistent narrowband RFI removal, drifting (narrowband) RFI removal, and removal of RFI using the clustering algorithm.

2.1. Persistent Narrowband RFI Removal

Although narrowband signals are searched in SETI programs, many RFI signals produced on Earth are also narrowband signals. The sources of these RFI signals include near-ground radar, television and radio broadcasts, artificial satellites, cell phone signals, etc. Unlike narrowband ETI signals, narrowband RFI signals in a specific frequency channel are usually observed in different sky areas (as they are not actually of astronomical origin) and can persist for a relatively long period of time. For the SETI commensal survey observations considered in this paper, the persistence of signals is to some extent equivalent to observing signals in a large sky area, since the telescope usually scans along R.A. due to the rotation of the Earth.

To remove this kind of RFI, we divide the whole frequency range (1000–1500 MHz for the FAST observation) into small bins and define a threshold of angular separation on the sky. If the hits in a frequency bin are found to spread over a sky area larger than the threshold (i.e., the angular separation of at least one pair of hits is larger than the threshold), we call the signals in this frequency bin as affected by the persistent RFI and remove all hits in the bin. The spirit of this method is similar to the "on–off strategy" applied in targeted SETI observations, which observe both the target (on-source) and reference (off-source) locations and reject RFI signals that are observed on both on- and off-source locations.

In practice, we process the hits one by one in order of time. For each frequency bin and value of decl.,[12] we only record hits with the maximum or minimum R.A. Whenever we are processing another hit in this frequency bin, we calculate the distances between this hit and all of the recorded hits in this frequency bin. Since the sky area observed in a reasonable time period is not too large, the maximum distance to the recorded hits is equal to the maximum distance to all previous hits in this bin. If the maximum angular distance is larger than the threshold, we mark this frequency bin as persistent narrowband RFI and ignore all subsequent hits in this frequency bin.

The persistent narrowband RFI removed with the aforementioned method is to some extent similar to the zone and multibeam RFI described in, e.g., Zhang et al. (2020). In Zhang et al. (2020), the zone RFI was defined as the frequency bins with numbers of hits larger than a threshold (set according to the Poisson statistics), and the multibeam RFI was defined as

[10] http://setiathome.berkeley.edu/nebula

[11] The offsets of the pointings of different beams are taken into account.

[12] For a SETI commensal survey using the "drifting scan" mode, the decl. for each beam is constant, so there are only several possible values of decl.

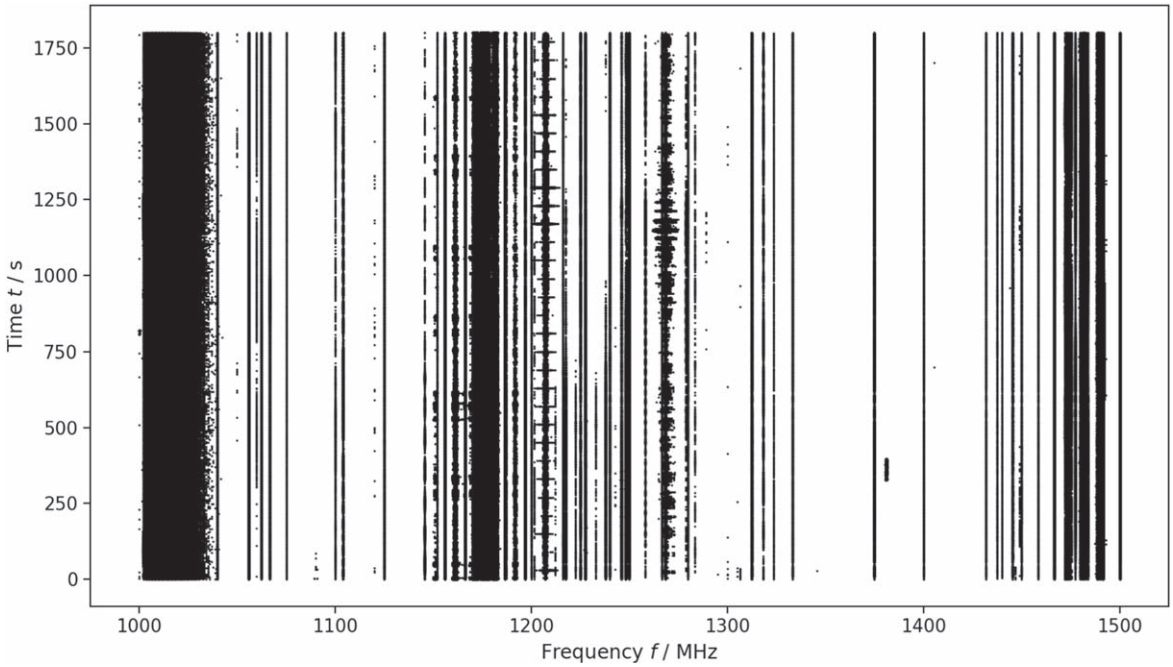

Figure 1. Waterfall (frequency–time) plot showing the raw observational data (i.e., hits, marked with black dots) in the first 1800 s of observation. The narrowband RFI is prominent, appearing as vertical lines; some broadband RFI (appearing as horizontal line segments) also exists. Throughout this paper, $t = 0$ corresponds to JD = 2,458,682.209155.

the hits received at a similar time and frequency but in nonadjacent beams (see Zhang et al. 2020 for details). We note that a frequency bin with a great number of hits usually means a large angular separation on the sky, and hits received by nonadjacent beams have relatively large angular separations.

Thus, our method in this part mainly deals with generally similar (though not identical) types of RFI as the zone/multibeam RFI defined in Zhang et al. (2020). However, we only record and compare several R.A. and decl. coordinates for each frequency bin, and we do not need to find hits with a similar time and frequency for each given hit. As a result, our method is simpler and usually needs less computation. We also note that our method processes hits sequentially in chronological order and removes a band immediately when a large sky separation is found. This means that the method can be potentially used for the real-time RFI detection, and we do not need to wait for a long time until the accumulated number of hits is large enough for the Poisson statistics.

On a waterfall plot, i.e., a plot of time t versus frequency f of the hits, a typical group of persistent RFI is a vertical line, as can be seen in, e.g., Figure 1.

2.2. Hough Transform for Drifting RFI Removal

The drifting RFI is a special kind of narrowband RFI that drifts in frequency; some of these signals are not fully understood. Possible origins include local oscillator malfunctions near the telescope, satellites, moving objects (e.g., cell phones), etc. Since this kind of RFI drifts rapidly in frequency (relative to the frequency resolution), it can be missed when using the methods of rejecting frequency bins (like the persistent RFI removal method in this paper). To remove the drifting RFI, one needs approaches to detect this kind of drifting feature.

In SETI commensal surveys, drifting RFI removal methods can be designed based on the waterfall plot, i.e., the time–frequency plane. In, e.g., Zhang et al. (2020), the drifting RFI is detected by defining symmetrical triangular bins for each hit and counting signals in the bins. If the number of signals in a triangular bin and its opposite three bins is above a threshold, the signals in these bins are defined as the drifting RFI (see, e.g., Figure 5 in Zhang et al. 2020 for an illustration of this method). Similar methods are also described in, e.g., Cobb et al. (2000).

In this paper, we propose a method of removing the drifting RFI based on the Hough transform. As a commonly used method in the image analysis, the Hough transform can robustly detect straight lines or any parameterized curves in images. This method has been used to detect several kinds of signals in the time–frequency plane, e.g., fast radio bursts (Zuo & Chen 2020) and sinusoidal ETI signals (Monari & Montebugnoli 2018). We note that the drifting RFI, although it may have no well-defined patterns, also consists of curves that can be clearly seen and detected with the Hough transform. Since short segments of a curve can be approximated as straight lines, we use the Hough transform to simply detect straight lines.

The Hough transform detects patterns in images by "voting" on the parameters of a family of curves. For straight lines, the commonly used parameterization, suggested by Duda & Hart (1972), is

$$\rho = x\cos\theta + y\sin\theta, \tag{1}$$

where (x, y) are the coordinates of the points on a line, and (ρ, θ) are the normal parameters that specify the line. Here ρ is the distance of the line from the origin, and the angle θ specifies the direction of the line's normal. Given a binary image, each (nonzero-valued) figure point (x_i, y_i) specifies a curve,

$$\rho = x_i\cos\theta + y_i\sin\theta, \tag{2}$$

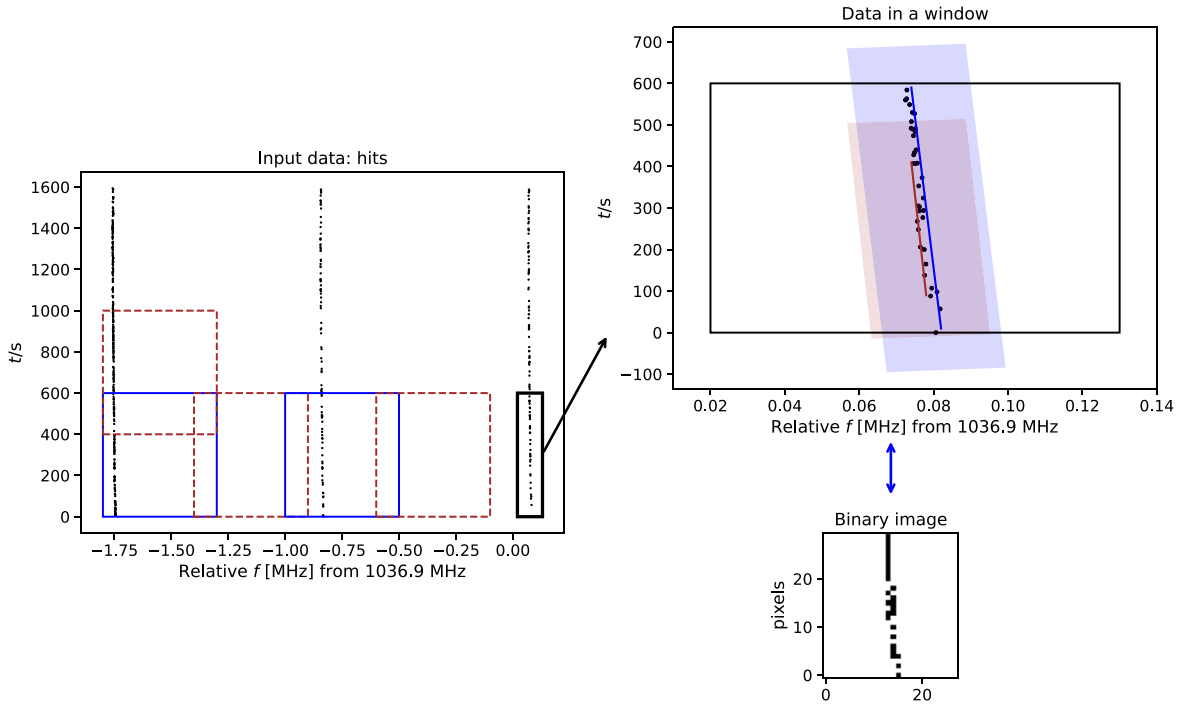

Figure 2. Illustration of the drifting RFI removal method proposed in this paper. Left panel: frequency *f*–time *t* waterfall plot showing an example of several groups of drifting RFI, where the signals called "hits" are marked with black dots. Note that although the signals look like vertical lines, they are drifting significantly in frequency compared to the resolution, 3.725 Hz. The *f*–*t* plane is divided into windows with overlaps, as shown with blue and brown rectangles (different colors and line styles are used for clarity). Right panels: a small window (also marked with a thick black rectangle in the left panel) that consists of a group of drifting RFI for demonstration. The upper right panel shows the waterfall plot of the small window (the thick black box), which is converted to a binary image in the lower right panel. Lines are detected with the binary image using the Hough transform and shown in the upper right panel with colored lines. Hits in a "corridor" around the detected lines, shown with the shaded regions, are marked as the drifting RFI. Note that all of the parameters used to make this figure, except the width and overlap in frequency of the windows, are the same as those adopted in the implementation in Section 3.2.

in the (ρ, θ) parameter space. After discretizing the parameter space into cells (usually $N_\rho \times N_\theta$ rectangular cells), Equation (2) gives a "vote" on the possible parameters in the (ρ, θ) space. Accumulating votes from all of the figure points, one can find the local maxima of votes and get the parameters (ρ, θ) of the detected lines.

To reduce the computation time, the probabilistic Hough transform was proposed (Kiryati et al. 1991). Rather than considering all of the figure points, this method uses small subsets of figure points to detect patterns. In this paper, we use the *OpenCV*[13] implementation of the probabilistic Hough transform, *HoughLinesP*, which is based on the progressive probabilistic Hough transform (PPHT) proposed by Matas et al. (2000). The parameters of the PPHT include accuracies of ρ and θ, the vote threshold (i.e., the minimum vote of a detected line; in other words, the number of points on a line), the minimum length of a line, and the maximum allowed gap in a single line. The algorithm can give the endpoints of all detected lines. For details, see Matas et al. (2000) and the *OpenCV* documentation.

In this paper, we first set pixel sizes for frequency and time and convert the scatter plot of hits[14] in the time–frequency plane into binary images, whose pixels in which there are hits are set to 1. To reduce the computation, the complete time–frequency plane is divided into an array of overlapping windows, and each window is converted into a binary image, as illustrated in Figure 2. For each image, line segments are detected with PPHT. Then, each line segment is extended by several pixels (so as not to miss the hits near the ends of the lines), and the hits in a "corridor" with a width of several pixels are marked as the drifting RFI. The process of removing the drifting RFI using the Hough transform is illustrated in Figure 2.

2.3. Clustering for RFI Removal

After removing the persistent and drifting RFI as described above, there are still RFI signals in the data. One example is the broadband RFI, as shown in, e.g., Figure 10 in Zhang et al. (2020). This can be due to lightning, power transmission cables near the surface of the Earth, electric fences, sparks, etc. While the narrowband RFI tends to persist in a range of time but is restricted to a narrow range of frequency (though drifting may be present), the broadband RFI is characterized by transient surges in relatively wide frequency ranges but small time durations. This feature makes it difficult to detect broadband RFI with the aforementioned methods.

Following Zhang et al. (2020), we use the kNN algorithm to remove more RFI from the data that pass the persistent and drifting RFI removal steps. Considering the nature of a SETI commensal survey, a set of ETI signals can neither persist for a long time duration (because of the drift of the telescope pointing) nor spread over a wide frequency range (because of the commonly assumed narrowband nature). Thus, ETI signals would cluster on a scale smaller than that of RFI clusters. To remove large clusters that we consider as RFI, we calculate the

[13] https://opencv.org/

[14] As mentioned above, these hits refer to the recorded information of the high-S/N signals in the frequency channels at each moment; thus, the input of the Hough transform is not the traditional waterfall plot of complete spectra (the matrix of signal powers at each frequency channel and time).

mean distances to the k nearest hits for each hit, and those with a distance below a threshold are considered as RFI. The number k should not be too small, so that hits in small clusters also have large mean distances, and only larger clusters are marked as RFI.

Though we mainly use the kNN algorithm for removing the broadband RFI (and other residual RFI), we also note that the Hough transform method described in Section 2.2 has the potential to detect broadband RFI. We explore the details in Section 4.

3. Data and Results

3.1. The Observed and Simulated Data

In this paper, we apply our method to the same observational data as in Zhang et al. (2020). The 5 hr data were collected during a drift-scan survey performed by FAST during commissioning in 2019 July. The decl. of FAST's pointing was a constant, but the accurate real-time pointing information of FAST was not yet available during the time of that observation (as discussed in Section 4.3 of Zhang et al. 2020). To make it possible for us to remove the persistent narrowband RFI according to the sky separation (Section 2.1), we approximately calculated the R.A. and decl. for the hits of each beam according to the scan velocity. Although there can be errors relative to the real pointings, it is enough for the work of RFI removal. We expect that accurate pointing information will be available in the future data of FAST's SETI commensal surveys.

The data set of hits was generated from the results of a real-time SETI spectrometer, SERENDIP VI (Cobb et al. 2000; Archer et al. 2016), which was used to process the raw observational data. With the SERENDIP VI system, the power spectrum was calculated at each signal time, covering frequency bands from 1000 to 1500 MHz with a resolution of about 3.725 Hz. The power of each frequency channel, normalized with respect to the baseline, was compared to an S/N threshold. At each signal time, the frequency channels that exceed the threshold ($S/N > 30$) were recorded as hits. The information of each hit includes the signal power, time, frequency, telescope pointing at the moment, etc. Normally, only these hits, rather than the complete spectra, are recorded for commensal sky surveys, which is different from common targeted observations. The reason is that sky surveys are typically long-term observations, and recording the full spectra ($\sim$38 billion spectral points per second) is expensive and difficult. For more information on the real-time data processing, see Zhang et al. (2020).

The hits generated as described above are the initial data for an RFI removal program and thus the input of our procedure. For reference, we plot the original data for the first 1800 s, i.e., the hits on the frequency–time (f–t) plane, in Figure 1.

To check the methods proposed in this paper, we also add the same set of mock ETI signals, called "birdies," generated and described in Zhang et al. (2020). Each group of simulated signals was generated assuming a source that is randomly located on the moving trajectory of FAST and only has emission in one frequency channel. When a beam goes through the location, some hits are generated in the beam. As shown in Figure 8 (see also Figure 4 in this paper and Figure 13 in Zhang et al. 2020), there are a total of 20 simulated "ETI signal sources," generating 20 groups of birdies with 294 hits.

Table 1
Ratios of Persistent and Drifting RFI

	Persistent RFI	Drifting RFI	Both[a]	Total[b]
Number of hits	538,956,414	538,112,874	536,574,702	540,494,586
Ratio	99.7067%	99.5506%	99.2660%	99.9912%

Notes.
[a] Hits marked as both persistent and drifting RFI.
[b] Hits marked as any of the two types of RFI.

3.2. RFI Removal

As described in Section 2, our RFI removal procedure includes (1) persistent narrowband RFI removal, (2) drifting RFI removal, and (3) RFI removal with the kNN algorithm. Since a hit can be part of the persistent narrowband RFI and the drifting RFI simultaneously, we only mark them as RFI rather than actually removing them in the first two steps. Then, we remove the two types of RFI together, after which the clustering analysis (step 3) is performed.

During the removal of the persistent narrowband RFI, we divide the frequency range (1000–1500 MHz) into bins with sizes of 7.45 Hz, twice the frequency resolution of the data. A frequency bin with hits spreading over an angular distance of $0.^{\circ}14$ is marked as the persistent RFI. This threshold is $\sim$1.5 times the distance between the centers of adjacent beams and about two to three times the beam size (FWHM; as reported in, e.g., Jiang et al. 2019, 2020), so we can safely determine that the signals that exceed this threshold are in an extended area of the sky and unlikely to be ETI signals. After this RFI removal step, 2,747,835 (4.094%) of the frequency bins, consisting of 99.7067% of all hits, are marked as RFI.

During the process of drifting RFI removal using the Hough transform, we divide the total frequency–time plane into windows with sizes of 20 MHz in frequency and 600 s in time. The overlaps of the windows are 1 MHz and 200 s, and the pixel sizes when converting the hits in windows into images are 0.004 MHz and 20 s. For the PPHT parameters (described in Section 2.2), we set the vote threshold to five, the minimum line length to 10 pixels, and the maximum allowed gap to 8 pixels. Then, we extend the line segments by 5 pixels and mark all hits with a distance of less than 4 pixels from the line as RFI. As shown in Figure 2, these parameters are appropriate such that the typical drifting RFI can be resolved and detected with the Hough transform method. Note that all of the parameters used to make Figure 2, except the width and overlap in frequency, are the same as the adopted parameters mentioned above. Since the frequencies of the RFI drift with scales much smaller than the width of the windows, changing the width and overlap in frequency has little effect on the RFI detection. After this step, 99.5506% of the hits are marked as drifting RFI.

After the first two steps, a total of 99.9912% of the hits are removed as RFI signals, most of which are marked as both persistent and drifting RFI; there are only 47,485 hits left for subsequent steps. The numbers of the two types of RFI signals and their ratios to all recorded signals are summarized in Table 1. The data for the first 1800 s after the removal of persistent and drifting RFI are shown in Figure 3. As can be seen, the narrowband RFI is significantly mitigated by our method. By comparing to the results in Zhang et al. (2020;

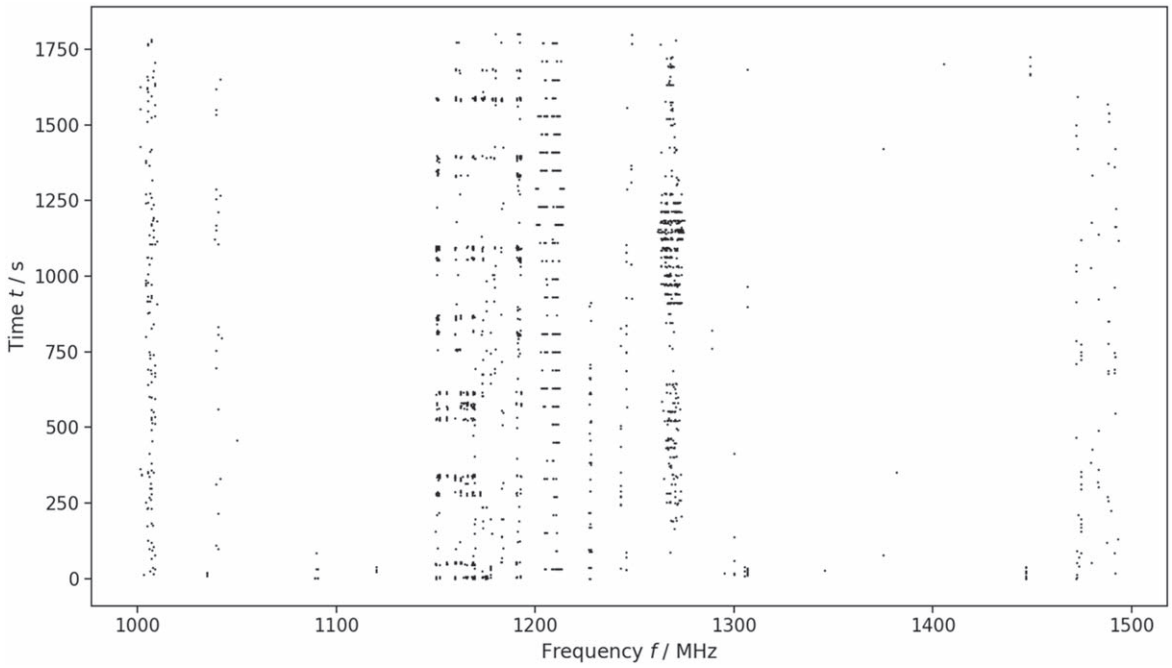

Figure 3. Same as Figure 1 (with the same size dots), but the hits marked as persistent or drifting narrowband RFI are removed. The narrowband RFI is greatly mitigated, though some broadband RFI still exists (and is mitigated later with the kNN method).

e.g., 98.1976% zone RFI, 99.9063% zone+drifting+multibeam RFI in Table 1 of Zhang et al. 2020), we find that we remove more RFI. We can also see in Figure 3 that the signatures of narrowband RFI are weaker, as compared to, e.g., Figure 14 in Zhang et al. (2020).

The birdies are added to the data set before the RFI removal process, and only one group of 15 birdies (5.1020% of the 294 birdies) are marked as (both) persistent and drifting RFI. As shown in Figure 4, this group of birdies happens to be in an RFI-affected region (the frequency bin that this group is in is marked as RFI even without adding birdies). Thus, it is not surprising that the algorithm marks these birdies as RFI, since it is hard to distinguish birdies and RFI in this case. We can conclude that our method, with the parameters set above, removes most of the RFI while preserving all (mock) ETI signals except those severely affected by the RFI.

The hits that pass the first two steps go through the last step of RFI removal based on the kNN algorithm. Following Zhang et al. (2020), we rescale the frequency and time of the data into the range of [0, 1] (a common data preprocessing step for machine-learning algorithms) and calculate the mean distance of each hit to the nearest $k = 100$ (excluding itself) hits. The process is mainly implemented with the *Python* package *scikit-learn* (Pedregosa et al. 2011). As shown in Figure 5, the mean distances for birdies are relatively larger than most observed hits. We set a threshold and remove 70% of the hits whose mean distances are below the threshold, while no birdie is lost with this threshold. The hits removed with kNN are shown in Figure 6. As can be seen in the right panel, some of the RFI removed in this step tends to be broadband RFI, which is hard to detect in the first two steps, since they mainly focus on the narrowband RFI.

In summary, our method of RFI removal effectively removes the vast majority of RFI while preserving the simulated ETI signals (birdies). In the step of persistent and drifting narrowband RFI removal, a total of 99.9912% of the hits are removed; in the kNN step, we can further remove 70% of the remaining hits, many of which belong to the broadband RFI. We remove more hits than the result in Zhang et al. (2020), which removed RFI with four steps (zone, drifting, multibeam RFI removal, and the kNN method). Thus, in our test, our simple method removes more RFI without loss of more birdies.

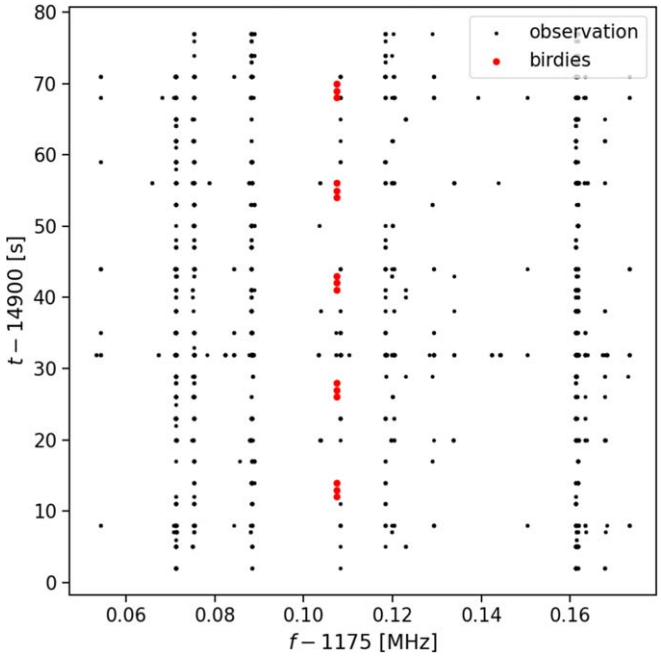

Figure 4. A group of birdies (artificially simulated ETI signals) marked as drifting/persistent RFI is shown in red. The raw observational data are shown in black. These birdies are significantly affected by the RFI, so it is reasonable that the algorithms mark them as RFI.

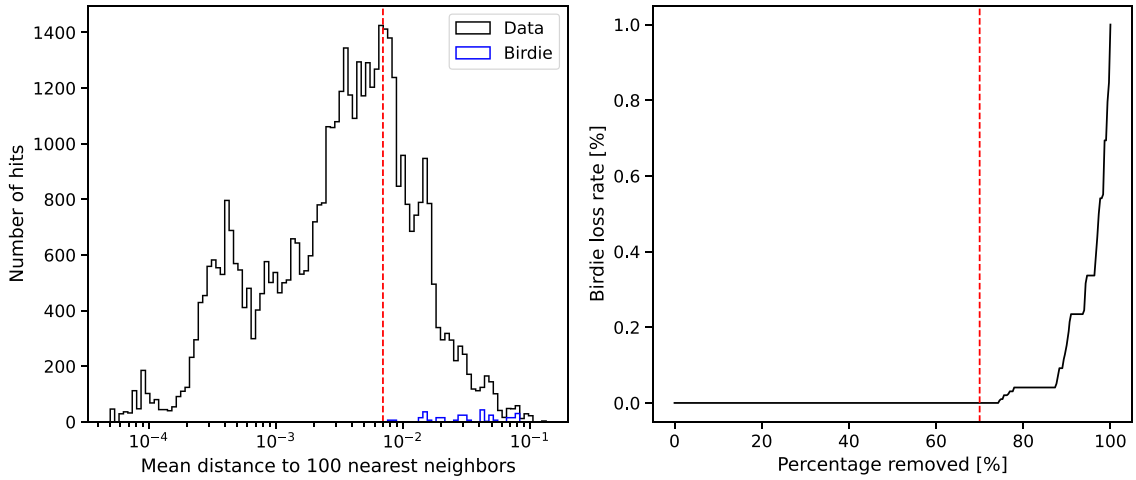

Figure 5. Selection of the threshold for the RFI removal step with kNN after removing the persistent and drifting RFI. Left panel: distribution of the mean distance between each hit and the 100 nearest neighbors on the normalized frequency–time plane. The distribution of all data (real observation and birdies) is shown in black, and the distribution for the birdies is shown in blue. The red vertical line shows the threshold for RFI removal (hits below it are removed). Right panel: birdie loss rate (the percentage of birdies that are removed as RFI) as a function of the percentage of all data removed (black curve). We set the threshold to remove 70% of all data, as marked with the red vertical line.

3.3. Candidate Selection

After the three RFI removal steps in Section 3.2, we are left with only 14,040 hits. Though most remaining hits should still be RFI, the result is good enough for us to perform the subsequent candidate selection steps. Similar to Zhang et al. (2020), we assume that a group of meaningful candidate signals should be a small cluster of hits and find the clusters using Density-based Spatial Clustering of Applications with Noise (DBSCAN; Ester et al. 1996; Schubert et al. 2017).

The DBSCAN model detects clusters with two parameters (thresholds), *minPts* and *Eps*, as introduced in, e.g., Ester et al. (1996) and Schubert et al. (2017). It defines the core points in the sample as those with more than *minPts* neighbors within the radius of *Eps*, and the neighbors that are not core points are called border points. A cluster is a group of core and border points that are close to each other decided with the above thresholds, while the points that are neither core nor border points are considered as noise. In our test, we rescale the frequency and time to [0, 1] (as for the kNN step) and set *minPts* and *Eps* to 5 and 9×10^{-4}, respectively. These thresholds are set for the scaled values of frequency and time and chosen to preserve all of the birdies, as shown in Figure 7.

With the above parameters, we use the DBSCAN implemented in *scikit-learn* (Pedregosa et al. 2011) and find 546 clusters (including 20 birdie clusters, which is consistent with the real number of birdie groups mentioned in Section 3.1). We then select candidate clusters by requiring the following.

1. The maximum sky separation is less than 1.5 times the distance between the centers of adjacent beams, thus selecting clusters that do not distribute over a large area. (This automatically rejects clusters with signals simultaneously detected by nonadjacent beams while allowing signals to be detected simultaneously by two adjacent beams or successively by different beams as a fixed point source on the sky drifts across the beams.)
2. The time duration is smaller than 100 s, and the frequency bandwidth is smaller than 500 Hz, thus selecting narrowband signals that do not persist for too long.

All 20 birdie clusters satisfy the above criteria, and 31 candidate clusters of real data pass the selection, as shown in Figure 8. We find fewer candidate clusters than Zhang et al. (2020; where 83 groups were selected), which is expected because we remove more RFI, and fewer hits are left for the candidate selection.

We visually inspect the candidate clusters to check whether they have obvious features of RFI. As shown in Zhang et al. (2020), some selected clusters may actually be very close to other hits removed as RFI and can be regarded as parts of the RFI that were missed in the previous steps. We extract and check the raw data within distances of 0.1 MHz, 1000 s to the candidate groups, and preliminarily find 14 promising groups that do not seem to be parts of large clusters of RFI.

The 14 interesting candidate groups are shown in Figure 9. As can be seen, the candidates are indeed narrowband small clusters, which also look similar to the birdies. Thirteen of the groups (all except that at ∼(1055 MHz, 4280 s)) are newly found in this paper, as they were not reported by the previous work (Zhang et al. 2020), which reported two groups of interesting candidates. Another group at ∼(1055 MHz, 4430 s) was also reported in Zhang et al. (2020) but is marked as RFI by the kNN step in this paper. As shown in the lower panel of Figure 18 of Zhang et al. (2020), this group of candidates belongs to a relatively larger cluster. Thus, the hits in the cluster have smaller mean distances to the nearest neighbors and are removed as RFI according to the threshold. This reminds us that the interesting candidates can still be parts of RFI that are missed during the removal procedure. Further verification and follow-up observations should be done before they can be considered as real "signals of interest."

By visual inspection, we also find that there are a few selected candidate groups that belong to the drifting narrowband RFI. Most of the cases are line segments of the drifting RFI, where the points are so sparse that they may be missed by the Hough transform line detection. However, there is no candidate group that is part of a "wide" (simultaneously affecting several frequency channels at each moment) segment of the drifting RFI (as found in, e.g., Figure 16 in Zhang et al.

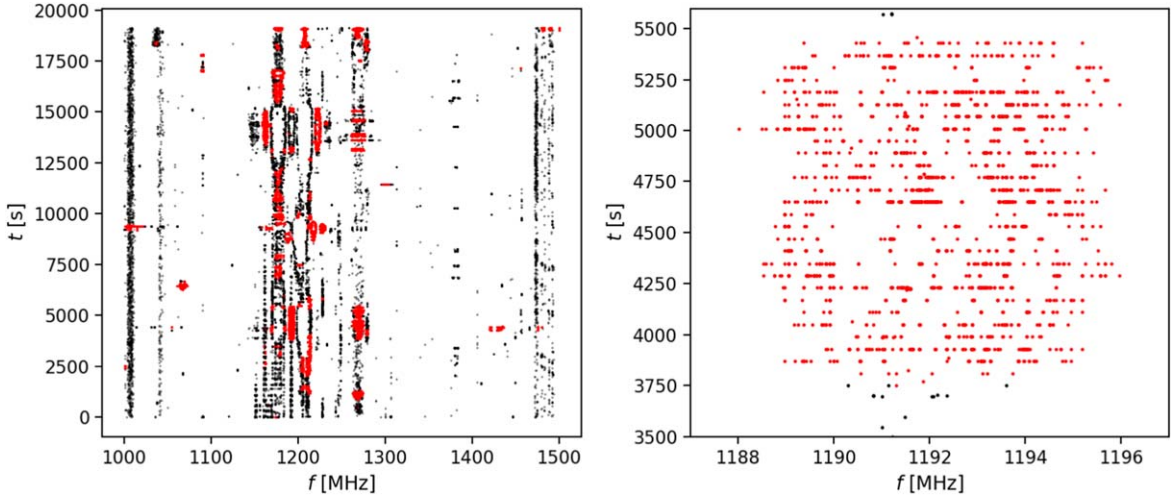

Figure 6. The hits removed by the kNN are shown in red, and those not removed are shown in black. Left panel: all hits after removing the persistent and drifting RFI. Right panel: example of a region where most hits of the broadband RFI are not removed before using kNN. The kNN effectively removes most of these hits.

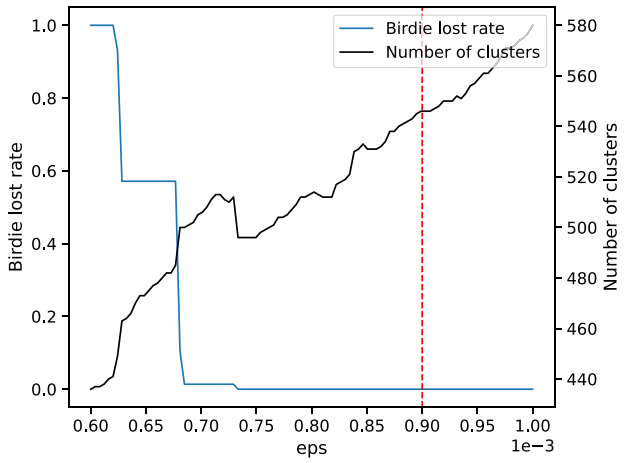

Figure 7. Selection of the parameter *Eps* for the DBSCAN algorithm (red vertical line). The birdie loss rate (in blue) and the number of clusters (in black) detected with DBSCAN are also shown as functions of *Eps*.

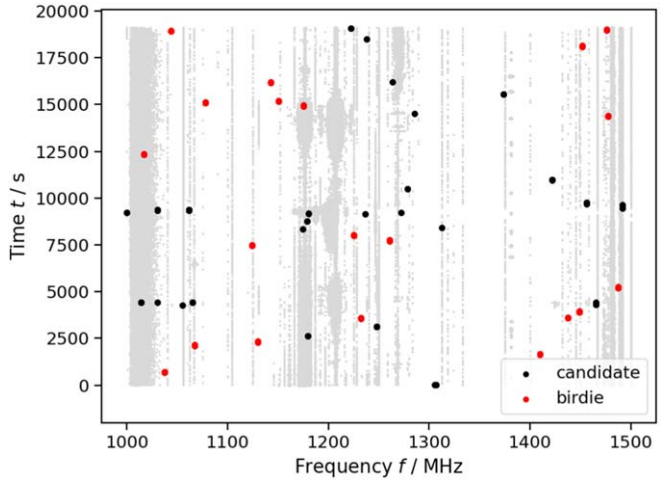

Figure 8. Frequency–time plane showing candidate clusters that are detected by DBSCAN and pass the selection criteria of candidates. Candidates of real observational data are marked in black, and birdies are marked in red. The background shows a randomly drawn sample (∼1/1000 of all) of the raw observational data, since the raw data are too large to be plotted simultaneously. Note that each dot consists of a cluster (or even several clusters) of hits, but the distances between them are too small to be resolved in this figure.

2020). This should result from the fact that we detect drifting RFI as hits within corridors (as described in Section 2.2 and illustrated in Figure 2), rather than removing the hits in shapes like triangles (as in, e.g., Zhang et al. 2020). The triangle areas may miss some hits near the edge of the drifting RFI, which may be detected later by DBSCAN and pass the selection criteria, since the missed points can be very small clusters. On the other hand, our method tends to remove the drifting RFI more cleanly as long as the line segment of the drifting RFI can be detected.

In summary, we find about a dozen groups of interesting candidates that do not seem to be members of large clusters of RFI. Zhang et al. (2020) found two groups of interesting candidates, and we report more new candidates that are very similar to the birdies, which represent the features of ETI signals that we are trying to detect. In addition, since we remove more RFI in the previous steps, this leaves us with fewer candidate groups, reducing the work of the visual inspection.

4. Discussion: Pixel Sizes for the Hough Transform

When removing the drifting RFI, we converted the scatter plots on the frequency–time plane into binary images on which we performed the Hough transform to detect lines. The parameters of the process, especially the pixel sizes of the frequency and time, need to be reasonably chosen such that the RFI feature can be resolved in the images. As mentioned in Section 3.2, we set the pixel sizes to 0.004 MHz and 20 s when detecting drifting RFI. The sizes are set considering the fact that the drifting (narrowband) RFI tends to persist for a relatively long time while drifting in a relatively small frequency scale, as shown in Figure 2 (and, e.g., Figure 4 in Zhang et al. 2020). With these parameters, we effectively remove most RFI while preserving all birdies except those affected by RFI.

However, as shown in the left panel of Figure 10 (and also Figure 3), some broadband RFI is not detected by the drifting

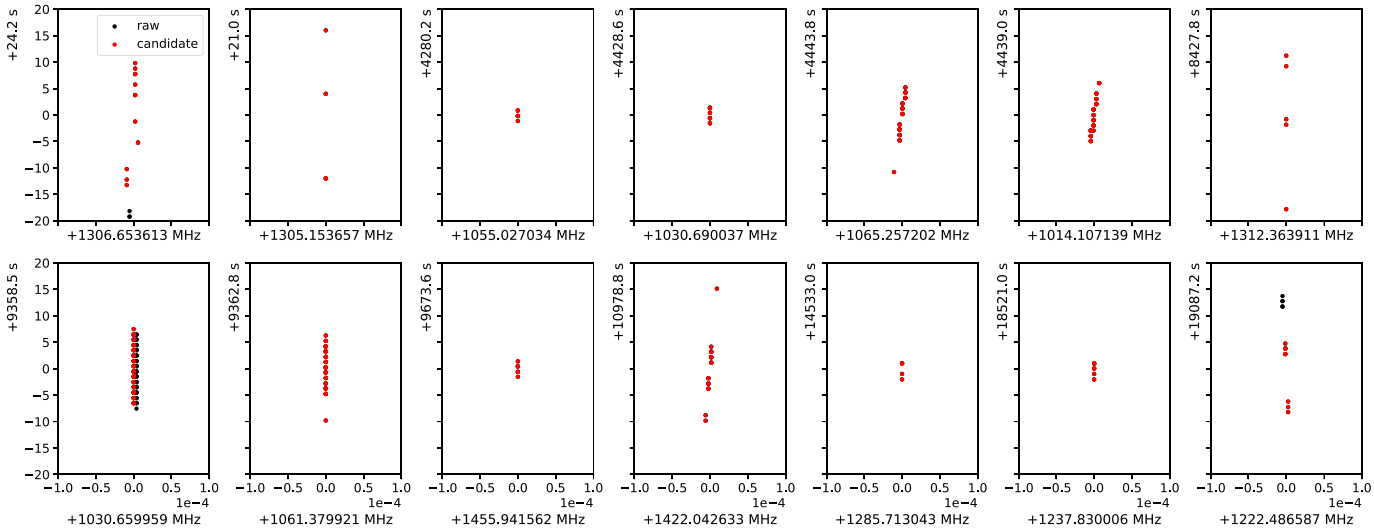

Figure 9. Some interesting candidate groups found in this paper. The groups are detected with DBSCAN, selected with the criteria described in Section 3.3, and no obvious evidence of a large cluster of RFI is found near the candidate by visual inspection. The elements of the groups are shown in red, and other hits from the raw data that do not belong to the detected group are shown in black. All panels are shown in the same scale, and the positions of the centers of the panels are labeled on the axes.

RFI removal method. A group of broadband RFI is typically a horizontal line on the frequency–time waterfall plot, which can be detected with the Hough transform, in principle. However, the frequency bands are so broad that, if the aforementioned pixel size of 0.004 MHz is used, the hits within a group of broadband RFI are too far away from each other to be considered as a line.

We note that the Hough transform can also be used to detect the broadband RFI, provided that the pixel sizes (and other parameters, if necessary) are properly set to match the scale of the broadband RFI. We make a test using part of the raw data that contains both narrowband (persistent/drifting) and broadband RFI, as shown in Figure 10. By changing the pixel sizes to 0.06 MHz and 0.2 s, our method with the Hough transform removes most of the broadband RFI, as shown in the right panel of Figure 10. However, with these parameters, some sparse narrowband RFI is missed, as expected. We find that the pixel sizes chosen for drifting and broadband RFI complement each other.

In summary, though the Hough transform is mainly used to detect drifting (narrowband) RFI in this paper, we show that it is also capable of detecting broadband RFI. However, broadband RFI appears to be horizontal lines on the *f*–*t* plot that are significantly different from the nearly vertical lines of the drifting RFI. Thus, different pixel sizes need to be set for the broadband RFI. Considering that the broadband RFI is much less dominant in the raw data, we use the parameters suitable for the drifting RFI in previous sections and remove the undetected broadband RFI with kNN.

5. Conclusions

In this paper, we propose a new procedure of RFI mitigation and removal for the FAST multibeam SETI commensal survey. Our methods remove the RFI in three steps, i.e., persistent narrowband RFI removal, drifting (narrowband) RFI removal, and kNN RFI removal. By applying our new procedure to the same FAST data and birdies analyzed in Zhang et al. (2020), we find that our methods can effectively remove the vast majority of the RFI while preserving the birdies (the simulated ETI signals). None of the birdies are detected as RFI, except for one group that is severely affected by the RFI. We detect and find about a dozen new interesting candidate groups, many of which look similar to birdies and do not obviously seem to be part of large RFI clusters. Thus, we conclude that our methods successfully mitigate the RFI for the SETI commensal survey and help us find signals similar to our simulated ETI signals.

Compared to a previous work (Zhang et al. 2020) on the first SETI multibeam observation with FAST (whose data are used to test our method), we use relatively simpler methods, remove more RFI in our test, and report some more interesting candidate groups. We use the simple threshold of sky separation to remove RFI (called persistent RFI in this paper) similar to that removed as zone and multibeam RFI in Zhang et al. (2020); for the drifting RFI, we use the Hough transform method to detect lines directly from windows consisting of many hits and remove RFI signals in "corridors," rather than checking each hit for the number of other hits in triangular bins and removing hits in triangles. With these methods, we effectively remove the RFI, even more than the result reported in Zhang et al. (2020), and do not see more loss of birdies.

We also explore the effect of different parameters for the Hough transform method, especially the pixel sizes, on the performance of RFI removal. We find that, though mainly used for removing the drifting RFI in this paper, the Hough transform is actually capable of detecting the broadband RFI, provided that proper parameters are chosen. This suggests that our method based on the Hough transform is flexible and has the potential of being applied to more tasks.

The RFI mitigation and candidate selection algorithms, either real-time or offline, are important in the work of SETI surveys. In the future, we also plan to continue the study of improving these algorithms, utilizing the strength and characteristics of different methods. A better understanding of the properties of both RFI and ETI signals may help us improve the data analysis pipelines. For example, if we know the characteristics of some sources of RFI, it might be possible

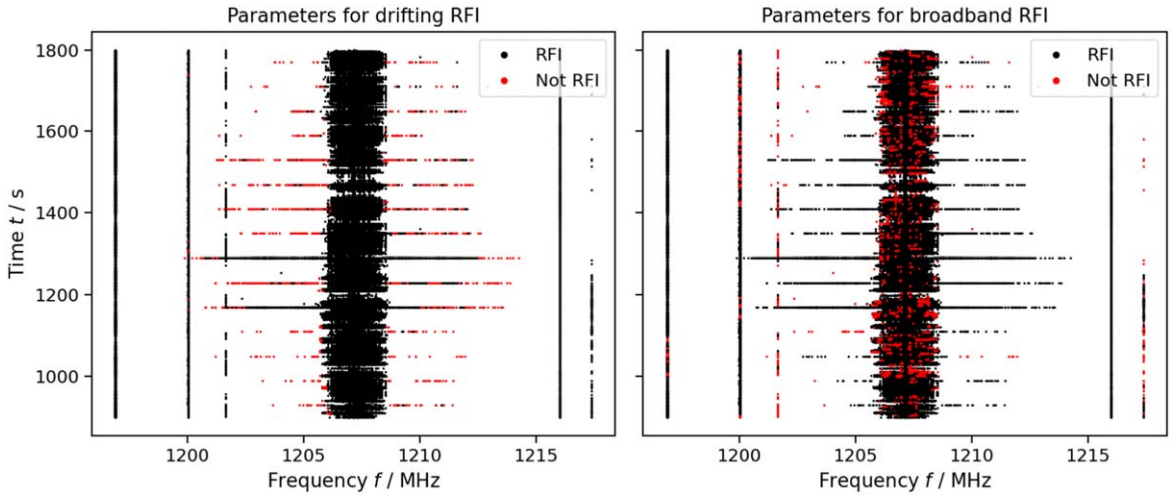

Figure 10. Example area where both narrowband and broadband RFI can be seen, and the Hough transform is used to detect RFI, as described in Section 2.2. The hits detected as RFI are marked in black, and those not detected are marked in red. In the left panel, the parameters suitable for removing drifting RFI (pixel sizes are 0.004 MHz and 20 s; as given in Section 3.2) are used. Most hits of the narrowband RFI are detected, while many of the broadband RFI are missed. In the right panel, the parameters suitable for removing the broadband RFI (pixel sizes 0.06 MHz and 0.2 s) are used. The broadband RFI is better detected, while some hits of the narrowband RFI are missed.

to search for candidate signals that are not consistent with the features of the known RFI, even in regions with a mixture of both ETI signals and RFI (as in, e.g., Figure 4).

On the other hand, further studies of the properties of potential ETI signals are also useful. We note that the parameters of our methods, especially those for the kNN and DBSCAN methods, are set with reference to the birdies. Some parameters are chosen to make sure birdies are not removed by the algorithm. However, this process relies on a fully representative library of simulated ETI signals, including more possible patterns of birdies. For example, the pattern may be different for different brightness, drift rate (e.g., Li et al. 2022), and relative position to the beams. Some interesting candidates may be missed (removed with RFI) because there is no kind of birdie that could guide our parameter selection. Thus, future work should make a more representative and diverse library of simulated ETI signals.

Acknowledgments

We are grateful for the referee's insightful and useful comments, which helped us improve our manuscript. T.-J.Z. (张同杰) dedicates this paper to the memory of his mother, Yu-Zhen Han (韩玉珍), who passed away 3 yr ago (2020 August 26). We sincerely thank Pei Wang, Bo-Lun Huang, and Jian-Kang Li for useful discussions. This work was supported by the National Science Foundation of China (grant Nos. 61802428 and 11929301). This work was finished on the workstation in Dezhou University.

Software: Python, NumPy (Harris et al. 2020), pandas (The pandas development team 2020; McKinney 2010), Astropy (Astropy Collaboration et al. 2022), Matplotlib (Hunter 2007), scikit-learn (Pedregosa et al. 2011), OpenCV (Bradski 2000).

ORCID iDs

Yu-Chen Wang https://orcid.org/0000-0002-8429-7088
Zhen-Zhao Tao https://orcid.org/0000-0002-4683-5500
Zhi-Song Zhang https://orcid.org/0000-0001-9294-0363
Cheqiu Lyu https://orcid.org/0009-0000-7307-6362
Tong-Jie Zhang (张同杰) https://orcid.org/0000-0002-3363-9965

References

Astropy Collaboration, Price-Whelan, A. M., Lim, P. L., et al. 2022, ApJ, 935, 167
Archer, K., Siemion, A., Werthimer, D., et al. 2016, in 2016 United States National Committee of URSI National Radio Science Meeting (Piscataway, NJ: IEEE), 1
Bradski, G. 2000, DDJ, 25, 120
Chen, Y.-X., Liu, W.-F., Zhang, Z.-S., & Zhang, T.-J. 2021, RAA, 21, 178
Cobb, J., Lebofsky, M., Werthimer, D., Bowyer, S., & Lampton, M. 2000, in ASP Conf. Ser. 213, Bioastronomy '99 - A New Era in Bioastronomy, ed. G. A. Lemarchand & K. J. Meech (San Francisco, CA: ASP), 485
Cocconi, G., & Morrison, P. 1959, Natur, 184, 844
Cohen, R. J., Downs, G., Emerson, R., et al. 1987, MNRAS, 225, 491
Drake, F. D. 1961, PhT, 14, 40
Duda, R. O., & Hart, P. E. 1972, Commun. ACM, 15, 11
Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. 1996, in Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining, KDD'96, ed. E. Simoudis, J. Han, & U. Fayyad (Washington, DC: AAAI Press), 226, https://dl.acm.org/doi/10.5555/3001460.3001507
Gajjar, V., Perez, K. I., Siemion, A. P. V., et al. 2021, AJ, 162, 33
Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Natur, 585, 357
Hunter, J. D. 2007, CSE, 9, 90
Jiang, P., Tang, N.-Y., Hou, L.-G., et al. 2020, RAA, 20, 064
Jiang, P., Yue, Y., Gan, H., et al. 2019, SCPMA, 62, 959502
Kiryati, N., Eldar, Y., & Bruckstein, A. M. 1991, PatRe, 24, 303
Lebofsky, M., Croft, S., Siemion, A. P. V., et al. 2019, PASP, 131, 124505
Li, D., Gajjar, V., Wang, P., et al. 2020, RAA, 20, 078
Li, D., & Pan, Z. 2016, RaSc, 51, 1060
Li, J.-K., Zhao, H.-C., Tao, Z.-Z., Zhang, T.-J., & Xiao-Hui, S. 2022, ApJ, 938, 1
Luan, X.-H., Tao, Z.-Z., Zhao, H.-C., et al. 2023, AJ, 165, 132
Ma, P. X., Ng, C., Rizk, L., et al. 2023, NatAs, 7, 492
Matas, J., Galambos, C., & Kittler, J. 2000, Comput. Vis. Image Underst., 78, 119
McKinney, W. 2010, in Proc. of the 9th Python in Science Conf., ed. S. van der Walt & J. Millman (Austin, TX: SciPy), 56
Monari, J., & Montebugnoli, S. 2018, MmSAI, 89, 342
Nan, R. 2006, ScChG, 49, 129
Ng, C., Rizk, L., Mannion, C., & Keane, E. F. 2022, AJ, 164, 205
The pandas development team 2022, pandas-dev/pandas: Pandas, v1.5.2, Zenodo, doi:10.5281/zenodo.7344967
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, JMLR, 12, 2825
Roth, L., Saur, J., Retherford, K. D., et al. 2014, Sci, 343, 171

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. 2017, ACM Trans. Database Syst, 42, 1
Sheikh, S. Z., Siemion, A., Enriquez, J. E., et al. 2020, AJ, 160, 29
Smith, S., Price, D. C., Sheikh, S. Z., et al. 2021, NatAs, 5, 1148
Tao, Z.-Z., Zhao, H.-C., Zhang, T.-J., et al. 2022, AJ, 164, 160
Tarter, J. 2001, ARA&A, 39, 511
Webster, C. R., Mahaffy, P. R., Atreya, S. K., et al. 2015, Sci, 347, 415
Werthimer, D., Anderson, D., Bowyer, C. S., et al. 2001, Proc. SPIE, 4273, 104
Zhang, Z.-S., Werthimer, D., Zhang, T.-J., et al. 2020, ApJ, 891, 174
Zuo, S., & Chen, X. 2020, MNRAS, 494, 1994