PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Automatic data processing pipeline for 19-beam HI observation of FAST

Wang, Zhi-yue, Zhao, Yu-tong, Jia, Ming-hao, Zhang, Guang-yu, Jiang, Feng-xin, et al.

Zhi-yue Wang, Yu-tong Zhao, Ming-hao Jia, Guang-yu Zhang, Feng-xin Jiang, Jian Wang, Bo Zhang, Ming Zhu, "Automatic data processing pipeline for 19-beam HI observation of FAST," Proc. SPIE 11452, Software and Cyberinfrastructure for Astronomy VI, 1145212 (13 December 2020); doi: 10.1117/12.2560782

SPIE.

Event: SPIE Astronomical Telescopes + Instrumentation, 2020, Online Only

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 18 Mar 2021 Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

Automatic data processing pipeline for 19-beam HI observation of FAST

Zhi-yue Wang[a] , Yu-tong Zhao[a], Ming-hao Jia[a], Guang-yu Zhang[a], Feng-xin Jiang[a], Jian Wang[a,*],Bo Zhang[b, †], Ming Zhu[b]

[a]State Key Laboratory of Particle Detection and Electronics, Department of Modern Physics, University of Science and Technology of China, Hefei 230026, China

[b]National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China

ABSTRACT

In the field of radio astronomy, the 21cm absorption line of HI is an important way to explore the large-scale structure and evolution history of the universe. The working frequency of FAST's 19 beam receiver is 1.05GHz to 1.45GHz, and the main observation object is to conduct an accurate and rapid intensity mapping survey of extragalactic HI's signal. Aiming at the 21cm spectral line of the object, we designed a parallel data processing platform to mitigate the influence of foreground, instrument, radio frequency, standing wave and other noises on the spectral data, then generate the image data of the whole sky region. At present, we divide the process into flux calibration, bandpass and baseline correction, radio frequency interference marking and data gridding work, etc. The whole project was programmed in Python, and Cython was used for some projects to speed things up.

Keywords: Five-hundred-meter Aperture Spherical Telescope , astronomical data pipeline, Spark, H1 line

1. INTRODUCTION

The Five-hundred-meter Aperture Spherical Telescope (FAST) is the most powerful single-dish radio at this day (2020). The 19-beam receiver [1] developed by the Commonwealth Industrial and Scientific Research Organization (CSIRO) and the large reflector enables FAST to achieve many scientific goals, and one of the main scientific goal is to perform the extragalactic HI observation.

The data readout from the receiver need to be processed for further usage. Many factors, including Radio Frequency Interfere(RFI), instrumental bias, will influence the data, therefore it is important to process these data and reduce the influence.

The FAST generated extremely huge amount of data. Each beam of receiver readouts from 1024K channels, these channels are combined into 65536 channels, each channel has 4 polarization direction. The receiver normally readout once in a second, and the data of 2048 times readout will be recorded into a SDFITS, therefore this FITS file contain a data array of 4*65535*2048 float variables, size of this file can be up to 2GB. The observation results of each beam are recorded into separate FITS files. For one observation, 19 FITS files are written simultaneously. Overall, about 90TB of data will be generated in 8 hours. To process this data in time is a challenge. We design and implement a data processing pipeline so as to fulfill this demand.

*Email: wangjian@ustc.edu.cn;

†email:zhangbo@nao.ac.cn;

Software and Cyberinfrastructure for Astronomy VI, edited by Juan C. Guzman, Jorge Ibsen, Proc. of SPIE
Vol. 11452, 1145212 · © 2020 SPIE · CCC code: 0277-786X/20/$21 · doi: 10.1117/12.2560782

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 18 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

2. DESIGN OF PIPELINE

Since FAST has different observation modes, and each observation mode has serval different configurations, the data processing pipeline must have strong configurability and reusability. We made the following class design to satisfy such demand:

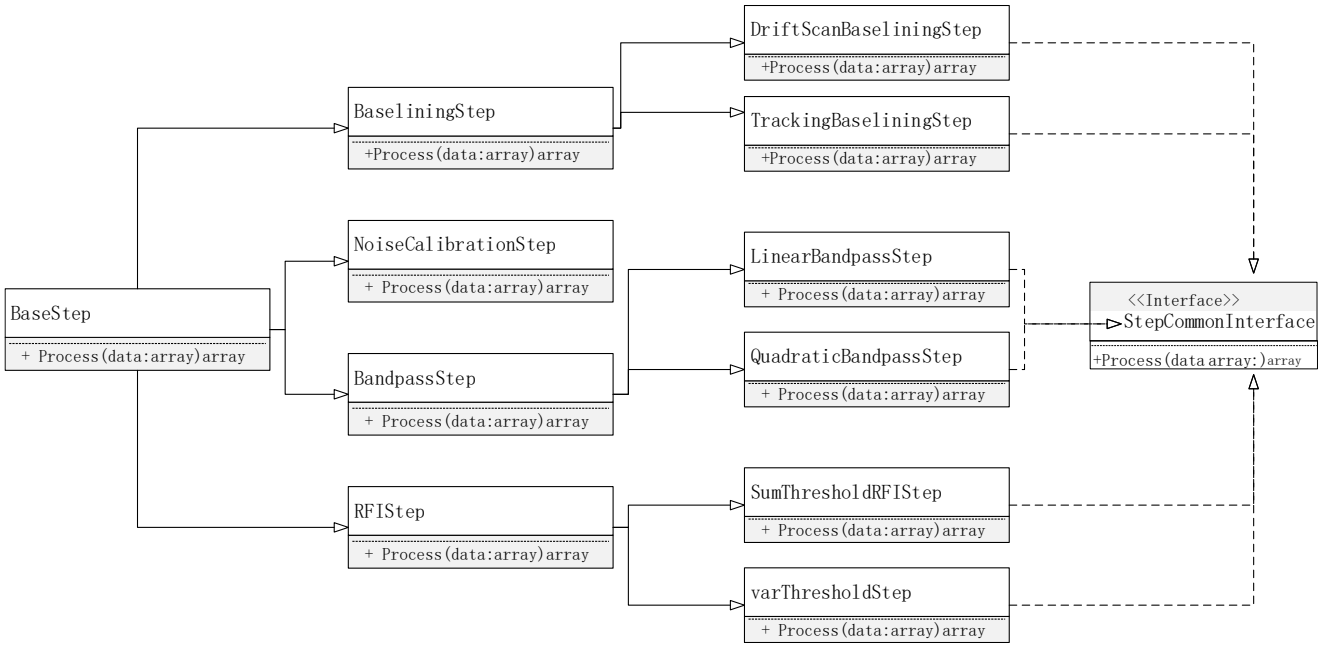

Figure 1 The Class Diagram for FAST Data Processing Pipeline

Each step in the data processing pipeline can be abstracted into a class with same interface and uses the same parameters and return types. The benefit is that each step of the pipeline can be easily rebuilt without affecting other steps. This design improves the reusability and develop efficiency.

The complete processing pipeline is formed by each separate steps by means of aggregation. The steps and configuration are managed uniformly with a configuration file. Since the pipeline usually work on the supercomputing platform, configuration files is suitable for the supercomputing platform. The different configuration files are passed to the pipeline executable with a command line parameter. Each process can be configured separately with different configuration file.

Since the pipeline is processing extremely amount of data, even the pipeline runs on high-performance platforms, it is still very important to optimizing the pipeline. We used Cython[2] to optimize the program. Cython is a library that compile python into the binary file to accelerate computing intensive tasks. Several steps, including the flux calibration, the band pass calibration and the baselining step of pipeline, need to perform a large number of matrix multiplication, traversal and interpolation. After the compiling with Cython, these steps can be many times faster than directly interpreting with python.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 18 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

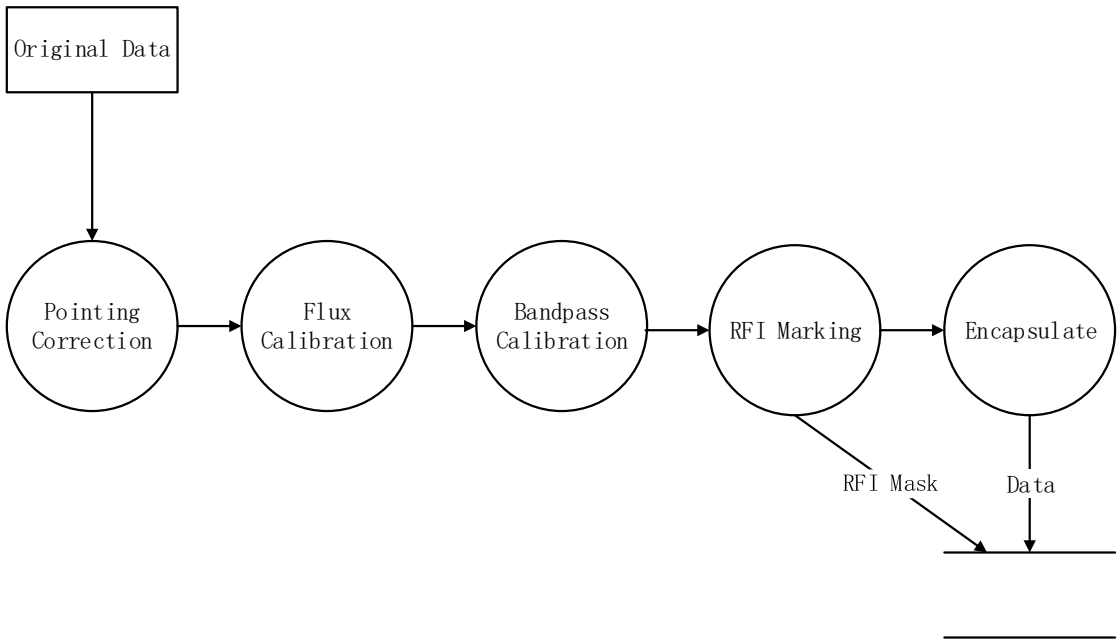

Figure 2 Processing Steps

As shown in Figure.2, the process of extragalactic radio data consists of four steps: pointing and speed reference conversion, flux calibration, band pass and baseline correction and radio frequency interference marking.

Pointing and speed reference. This step is used to get the direction information of the telescope in different frames of reference. Pointing conversion converts geodetic coordinates into horizontal coordinates (AZ, ZA) and equatorial coordinates (RA, DEC). The conversion of the velocity reference frame changes the velocity of the radio spectrum recorded in station reference to the earth central system and then to the solar mass central system. The above process calls the Time module function and the Coordinates module's *astropy.Transform_to* function to calculate it, as well as the Ephem library to do a simple but fast coordinate transformation. The wind speed, temperature and atmospheric pressure are also considered in this step.

Flux calibration. Flux calibration is to convert the raw readout count to actual brightness temperature of source. By adding a noise diode with known temperature T_{cal} to the observation system, we get two sets data, ON_{cal} and OFF_{cal} for each beam and each polarization direction, each frequency with the diode on and off periodically [3]. The brightness temperature T_{off} can be calculated with following equation:

$$T_{off}(f, beam, pol) = \frac{off_{CAL}}{on_{CAL} - off_{CAL}} T_{CAL}(f, beam, pol)$$

Furthermore, the brightness temperature can be converted to the flux with:

$$S = \frac{T_{source}}{Gain(az, za)}$$

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 18 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

During these step, we find that sometime the data is polluted by some stripe noise. The elimination of these noise is discussed in Section.III.

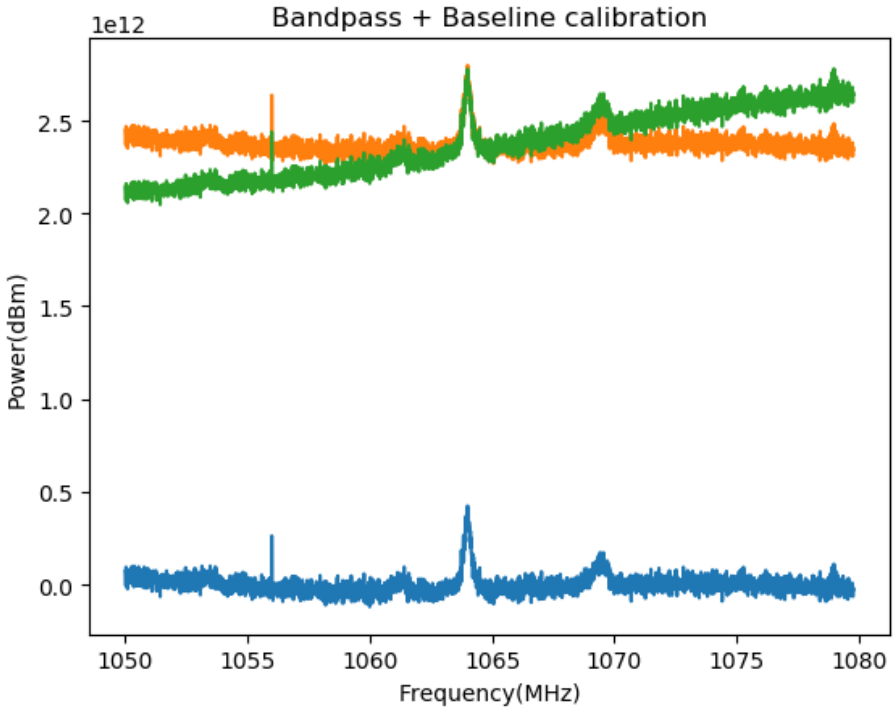

Figure 3 Green Line Shows the Original Data, Orange Line Shows Band-pass Data, the Blue Line shows the Result

Band pass and baseline correction. This step removes the effect of continuous spectrum signal of sky and the frequency response of the instrument. The data of each spectral line is fitted into polynomial curve, and remove large residual. Such step is repeated 3 or 4 times before the fitting curve become stable. The flux data is divided by fitting curve to correct the band-pass effect. The baseline correction process is very likely. To avoid the fit curve become diverge, the fitting order normally does not exceed the third. The results of this step is shown in the Figure.3.

RFI marking. The RFI marking needs to produce a mask of size as the original data. At present, the pipeline adopts the SumThreshold algorithm. This step will be further discussed in section III.

After these steps, the data is ready to be stored. Like the original data, the product data is storage in SDFITS format. The RFI mask is stored beside the data as another FITS file.

3. IMPORTANT ALGORITHM INTRODUCTION

3.1 Standing Wave Mitigation

When high temperature noise signal is injected, a stripe noise with width of about 144 channels is observed, as shown in figure below. This stripe noise can be explained by the standing wave of FAST antenna system. To eliminate this noise, we tried to process the data with several filters [4]. The filter we chose include Gaussian function, cube filter, and quadratic function. The width of filter is similar to the stripe. Beside these commonly used filters, we also tried several other band stop filters with multiple stopband, they show difference in performance with conventional filters in some aspects.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 18 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

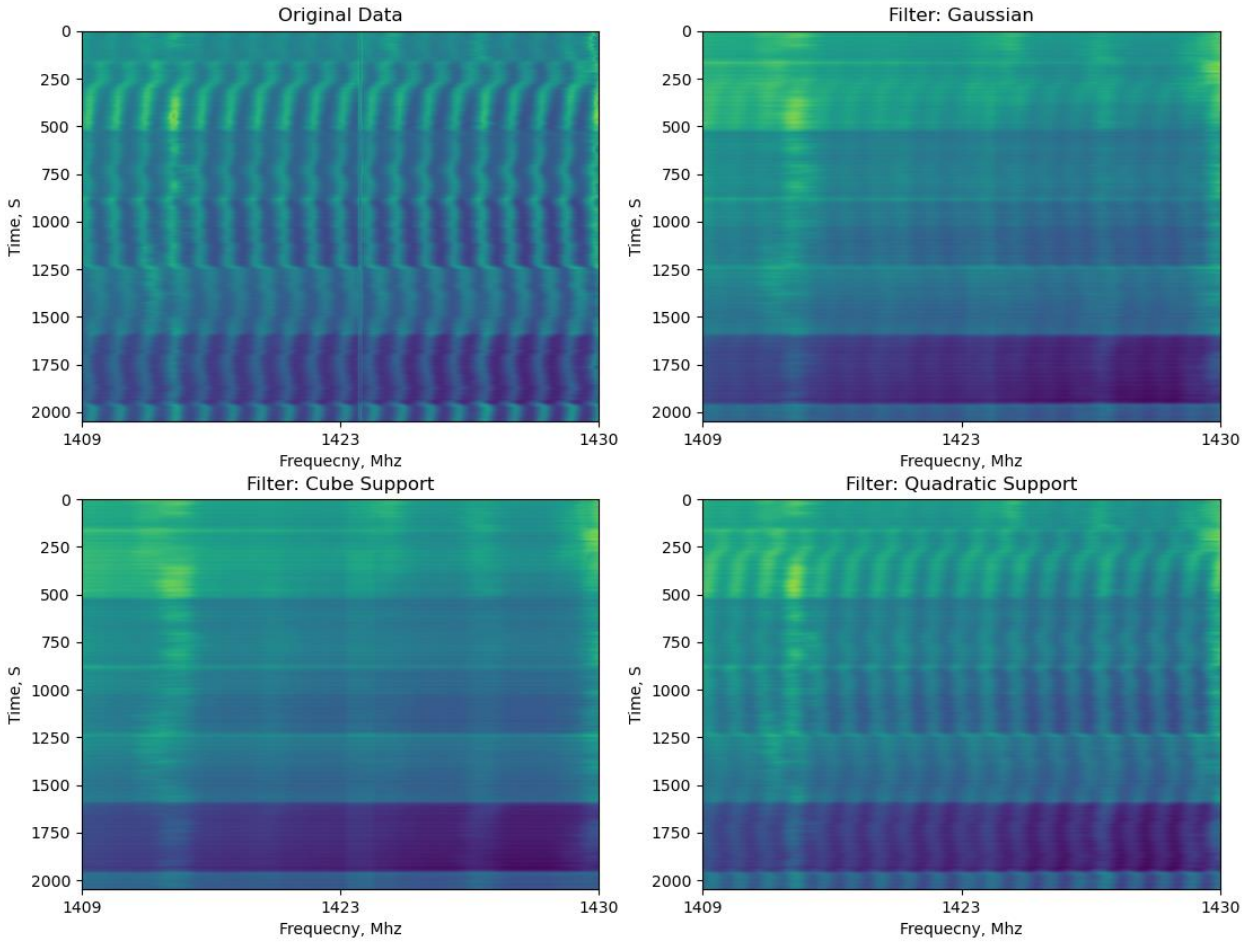

Figure 4 The Original Data and the Filtered Data

The effect of this process mainly depends on the filter we choose. As shown in Figure.4, cube filter support function shows the best effect among these filters. However, as a side effect of cube filter, the HI signal, which has similar width with stripe, is weaken. Gaussian and quadratic support can't fully eliminate the noise, but the Gaussian function can retain the HI signal better. Therefore, we use the Gaussian for general processing. User can specify which filter is to be used with configure file.

For the multiple-stopband filter we test, a filter which reject both 15~25Hz and 35~42Hz shows best performance during our test. Although it can't totally erase the standing wave for all the cases, the HI signal is retained during the process. But it doesn't show any advantage compare to Gaussian function.

3.2 SumThreshold & VarThreshold Algorithm

Although the FAST is sited in a radio quiet zone, the Radio Frequency Interference(RFI) still can't be shielded completely. The RFI must be marked and mitigated by data processing. We test the SumThreshold and VarThreshold algorithms. These algorithms both belong to combine threshold method, the main idea of combine threshold method is to combine nearby points to a group, and if the average of group exceed a threshold, the whole group is marked as RFI.

The core idea of this algorithm is very simple, but the implementation needs to select the size of each combination and adjust the threshold according to the actual data. We assess the performance of algorithm by the call back ratio and precision.

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 18 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use

We tested different set of configuration of combination size and threshold choosing strategy. The train set is marked manually, and the result of precision and recall ratio is shown in Table.1.

Table.1 The Best Effect of SumThreshold Under Several Combination Size

N_1	χ_1	Precision（%）	Recall Ratio（%）	RFI Cases
1	9 σ	93.26	15.87	727
2	13 σ	93.05	17.55	806
4	20 σ	93.58	16.38	748
8	27 σ	94.17	21.16	960
16	49 σ	94.60	13.11	592
32	60 σ	93.93	23.22	1056

We also use other methods to improve the RFI marking process. These methods include avoiding extreme value for segmentation, modifying the baseline, determine the threshold value with flat spectral line segment. Moreover, the process can be repeated for more than one times with different combination size and threshold configuration, in order to improve the precision for different kinds of RFI [5].

4. CONCLUSION

The data processing pipeline can process a large number of data of different observation modes. The pipeline effectively reduces the impact of noise and interference on the HI signal, and complete the two-dimensional intensity mapping survey for drift observation.

The pipeline has been deployed to the FAST data center cluster and we have obtained a batch of product data.

5. ACKNOWLEDGEMENTS

This work was supported in part by the Fundamental Research Funds for the Central Universities (WK2360000003, WK2030040064), in part by National Natural Science Funds of China under Grant No: 11603023, 11773026, 11903056, U1731125, in part by FAST project, in part by Research Funds of the State Key Laboratory of Particle Detection and Electronics, in part by CAS Center for Excellence in Particle Physics, in part by the Strategic Priority Research Program of CAS (Grant No. XDC07020200, XDA15020605).

6. REFERENCES

[1] Dunning, Alex , et al. "Design and laboratory testing of the five hundred meter aperture spherical telescope (FAST) 19 beam L-band receiver." 2017 XXXIInd General Assembly and Scientific Symposium of the International Union of Radio Science (URSI GASS) 2017.

[2] Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., & Smith, K. (2011). Cython: The Best of Both Worlds. Computational Science and Engineering, 13(2), 31–39.

[3] Liu, Bin , and S. Yu . Flux calibration for single-dish radio telescopes. Big Data in Astronomy. 2020.

[4] Gonzalez, Rafael C. , and R. E. Woods . "Digital image processing." Prentice Hall International 28.4(2008):484 - 486.

[5] Yang, Zhicheng , et al. "Deep residual detection of radio frequency interference for FAST." Monthly Notices of the Royal Astronomical Society (2020).

Downloaded From: https://www.spiedigitallibrary.org/conference-proceedings-of-spie on 18 Mar 2021
Terms of Use: https://www.spiedigitallibrary.org/terms-of-use