4^{th} URSI AT-RASC, Gran Canaria, 19 – 24 May 2024

Flagging strategy for the DSA-2000

Gregory Hellbourg
California Institute of Technology, Pasadena, CA, USA

Abstract

The DSA-2000 radio telescope is a next generation interferometer with stringent requirements in terms of real-time processing, sensitivity, and data quality. Corrupted data need to be identified and excised in real-time from the data processing chain as the system architecture does not support the storage and accumulation of the data. This paper addresses the problem of efficient and fast data flagging for DSA-2000.

1 Introduction

The Deep Synoptic Array DSA-2000 is a future survey radio interferometer made of 2,000×5-m dish antennas, and aiming at producing deep radio images of the sky in near-real-time over the 0.7-2 GHz frequency range. The main time allocation will be dedicated to a cadenced sky survey imaging the full accessible sky 16 times over 5 years through $\sim$ 6,000 $\times$ 15 mins pointings.

The signal chain of the telescope follows a classic FX correlator design [1, 2]. The radio frequency signals captured at each dual-polarized antennas feeds are first amplified, filtered, then transported to the main computing building over fiber optics. They then enter the digital front-end, responsible for the digitization by direct sampling, channelization into 130 kHz-wide channels, equalization, requantization, and packetization. The data packets are then sent to GPU-accelerated nodes over an ethernet network responsible for the "cornerturn". The GPU nodes will receive the signals from all antennas (2,000×2 polarizations) for a small number of frequency channels ($\sim$4). These nodes will then perform the cross-multiplication between all signals, the integration over $\sim$1 second to form "high-time-resolution" visibilities, the data flagging, calibration, gridding, accumulation over $\sim$15 mins windows, and finally the fast fourier transform to obtain final images that will later be agglomerated over all frequency channels. Real-time imaging is enabled by the large number of baselines that nearly fill the UV grid, alleviating the need for computationally intensive deconvolution stages.

The real-time nature of the system requires an efficient on-the-fly flagging scheme to accurately discard corrupted data without data storage and accumulation. This paper presents the needs in data flagging in radio astronomy, and describes the data flagging schemed proposed for DSA-2000.

2 Data flagging

Data corruption may have several origins:

- *Radio Frequency Interference* (RFI): signals of human origin may be captured in addition of astronomical signals by the telescope. Without full prior knowledge of the RFI contribution, the recovery of the underlying astronomy signal is impossible. It is accepted by the astronomical community that discarding the data affected by RFI is the safest interference mitigation protocol. The detection of RFI-affected data relies on power detection, but more advanced approaches exist to alleviate low-power RFI scenarios [3, 6]. RFI usually spans narrow frequency bandwidths (up to a few MHz), may be temporally structured (with a given duty cycle or of impulsive nature), and will usually non-uniformly affect baselines of the telescope due to their local, terrestrial, or near-field origin, and the phase tracking operated on a radio interferometer [4]. We note however that strong RFI may affect wide bandwidth when the total received power in the receiver drives them into non-linear regimes (due to digitizer clipping, or due to the saturation of analog components).
- *Misbehaving antennas*: The antennas of an array telescope have their individual radio frequency environments, and all components of individual antennas exhibit statistical fluctuation leading to potential misbehaviors. It is recommended to discard all contributions of a misbehaving antenna signal. Such misbehavior may be due to strong local RFI driving the antenna receiver into non-linear regime, shadowing, mechanical drivers failure, electronic components failure, interfering presence of fauna or flora, weather artifacts such as accumulated snow or rain in the reflector. Antenna misbehavior can be detected at the F-engine by tracking the overall collected data quality.
- *Maintenance*: preventive and restorative maintenance may occur at any time, and possibly even when the telescope is operating. Maintenance may affect individual antennas, groups of antennas, or the full array depending on the system level being assessed or repaired. In the two first cases, it is necessary to discard

antennas that are either completely unavailable or have a varying and unpredicted performance. Maintenance is planned or logged, and the data originating from the affected subsystem should be preventively flagged and discarded.

3 F-engine flagger

The F-engine is responsible for data digitization, channelization, quantization, and packetization. The data available at the F-engine level provide an assessment of the global behavior of individual antennas. Deploying a flagger at the F-engine level allows the detection of data corruption that otherwise would remain undetected after the first corner-turn where only narrow portions of the spectrum is available for all antennas.

The engine can compute an instantaneous autocorrelation spectrum made of the sample variance for individual polyphase filter bank (PFB) channels that can be compared to a reference "standard" spectrum measured in a controlled environment. Deviation from the standard spectrum is quantified by a mean squared error, and excursion of this error above an empirically set threshold will result in the full flagging of the data transmitted by the F-engine. Such deviation is typically due to analog electronics failures or saturation in case of strong RFI.

Flagging will also occur on individual PFB channels. The rationale for this feature is that some channels will overlap with allocated sources of RFI that are intermittent or bursty in nature. An example of such RFI is the 4G LTE cellular downlinks sparsely allocated at frequencies below 960 MHz. During low-occupancy periods, i.e. when only a low number or no cellular phones are connected to a cellular tower, the tower only transmits short bursts of energy with a low duty cycle. A significant portion of time can then be recovered by only discarding time samples affected by the RFI burst. The channel flagger operates by simple thresholding. Thresholds are also set empirically after calibration measurements (i.e. measuring signals in the absence of RFI).

Both the "total spectrum" and the "individual channel" flaggers are based on thresholding of a quantity : a deviation metric for the former, and sample variance for the latter. The choice of threshold is associated with a given probability of non-detection which reduces the detector sensitivity to avoid over-flagging. Data corruption that remains undetected at the F-engine level can however still have a significant impact on the telescope data product, and more sensitive data quality assessment is required at the X-engine level.

Flags generated by both F-engine flaggers are embedded in the output data stream of the system as an invalid sample value. That value, when detected by the X-engine (i.e. GPU node), is understood as a missing value, and is discarded before further processing.

4 X-engine flagger

We assume in this section that the data received at the X-engine level have not been significantly flagged at the F-engine level (i.e. $\leq 1\%$). The contradictory scenario is out of the scope of the present paper.

Under the null hypothesis (i.e. in the absence of RFI), we assume that the digital signal received from each antenna is made of a contribution of both uncorrelated and correlated noise. The uncorrelated noise models the system thermal and electronic noise. Each sample is modeled as an independent and identically distributed (i.i.d) noise following a centered complex normal distribution. The correlated noise represents the astronomical information, and each sample is also modeled as an i.i.d noise following a centered complex normal distribution. We neglect mutual-coupling, although such effect could simply be included as an additional source of correlated noise. The signal coming from antenna i can therefore be written as:

$$x_i[n] = \alpha_i x_a[n] + x_{i_n}[n] \tag{1}$$

where n is the digital sample number, α_i is a complex scalar relating the spatial information sampled by antenna i, $x_a[n]$ is the astronomical contribution, and $x_{i_n}[n]$ is the system noise unique to antenna i. The visibility of baseline (i, j) is defined as:

$$V_{i,j} = \frac{1}{N}\sum_{n=1}^{N} x_i[n] \cdot x_j^*[n] \tag{2}$$

where N=200,000 is the number of samples the visibility is calculated over and is roughly equivalent to a 1.5 seconds integration time, and $(.)^*$ is the complex conjugate operator. Considering the independence between samples, we have:

$$x_i[n] \cdot x_j^*[n] = \underbrace{\alpha_i \alpha_j^* x_a[n] x_a^*[n]}_{\text{correlated contribution}} + \underbrace{x_{i_n}[n] x_{j_n}^*[n] + \alpha_i x_a[n] x_{j_n}^*[n] + \alpha_j^* x_a^*[n] x_{i_n}[n]}_{\text{uncorrelated contribution}} \tag{3}$$

We define the 4 independent products:

$$V_{a_{i,j}} = \frac{\alpha_i \alpha_j^*}{N}\sum_{n=1}^{N} x_a[n] x_a^*[n]$$

$$V_{n_{i,j}} = \frac{1}{N}\sum_{n=1}^{N} x_{i_n} x_{j_n}^*$$

$$V_{a_j,n_i} = \frac{\alpha_j^*}{N}\sum_{n=1}^{N} x_{i_n} x_a^*[n]$$

$$V_{a_i,n_j} = \frac{\alpha_i}{N}\sum_{n=1}^{N} x_a[n] x_{j_n}^*[n]$$

And we can therefore write:

$$V_{i,j} = V_{a_{i,j}} + V_{n_{i,j}} + V_{a_j,n_i} + V_{a_i,n_j} \tag{4}$$

The correlated contribution $V_{a_{i,j}}$ follows a χ^2 distribution, such that [5]:

$$\frac{N \cdot V_{\alpha_{i,j}}}{\alpha_i \alpha_j^* \sigma_a^2} \sim \chi^2(N) \tag{5}$$

where σ_a^2 is the variance of $x_a[n]$. The uncorrelated noise contributions all follow the same distribution, which is a variation of the Student t distribution:

$$\sigma_{n_i}^2 \sigma_{n_j}^2 V_{n_{i,j}} \sim \frac{(1-\nu^2)^{\frac{N-4}{2}}}{B\left(\frac{1}{2}, \frac{1}{2}(N-2)\right)} \tag{6}$$

where $B(.,.)$ is the Beta function, and ν is the visibility real or imaginary value.

By application of the Central Limit Theorem, $V_{i,j}$ is approximately normally distributed with zero mean, as a sum of independent random variables. When subjected to RFI, the visibility between antennas i and j noted $\widehat{V_{i,j}}$ is expressed as:

$$\widehat{V_{i,j}} = V_{i,j} + \Gamma_{\text{RFI}}(\tau) \tag{7}$$

where $\Gamma_{\text{RFI}}(\tau)$ is the autocorrelation function at time-lag τ of the RFI waveform. $\widehat{V_{i,j}}$ is also normally distributed with a deterministic mean following the deterministic function $\Gamma_{\text{RFI}}(\tau)$.

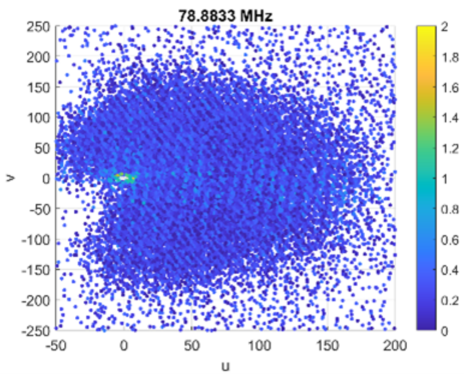

Figure 1. LWA visibilities at 78.9 MHz (no RFI)

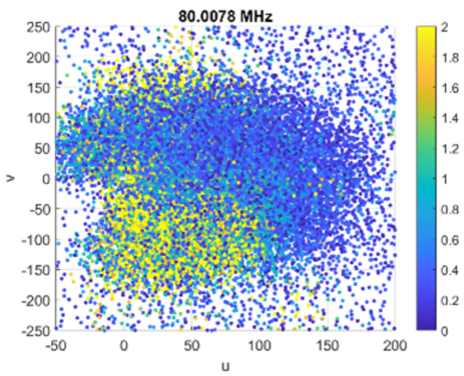

Figure 2. LWA visibilities at 80 MHz (with RFI)

Figure 1 shows an example of a set of visibilities collected with the Long Wavelength Array (LWA) at 78.9 MHz projected in the UV domain. The color of each dot indicates the magnitude of the visibilities. The data set does not contain any contribution from RFI. The diagonal streaks are due to a strong astronomical sources (e.g. Sun) up in the sky, which dominate the system noise and any other astronomical source. This structure is expected with a telescope such as the LWA because of its wide field of view but will be negligible with the DSA-2000. The visibilities statistics are not uniform over the whole UV domain but are “locally” stationary. This statement indicates that RFI-free visibilities will follow almost-identical distribution over a small region in the UV domain. Stronger variance can for instance be observed in proximity of the (0,0) point of the UV plane due to a stronger contribution of correlated noise due to the cross-coupling between antennas over shorter baselines. A similar behavior is observed along the streaks representing constructive spatial interference of the astronomical source.

Figure 2 shows a similar plot when RFI is present. The RFI in this context is a VHF TV transmitter (channel 5 or 6). The visibilities are this time dominated by the RFI which corrupts patches of the UV plane. These patches are due to strong correlation at given baseline lengths, which are related to the RFI bandwidth, center frequency, and direction-of-arrival. The presence of the RFI changes the statistics of the visibilities, which are no longer locally normally distributed due to the presence of the deterministic RFI autocorrelation.

We propose the use of the spectral Kurtosis [6] to detect the non-Gaussianity of the data which is a direct consequence of the presence of a deterministic contribution to the data model. The spectral Kurtosis is an adaptation of the Kurtosis estimator which applies to the squared magnitudes of a random process. The spectral Kurtosis is expressed as:

$$S_K = \frac{M+1}{M-1}\left(\frac{M \cdot S_2}{S_1^2} - 1\right) \tag{8}$$

with $S_1 = \sum_{i,j}^{M} |V_{i,j}|^2$ and $S_2 = \sum_{i,j}^{M} |V_{i,j}|^4$ are the sum of visibilities magnitudes and squared visibilities magnitudes. M is the number of visibilities processed to extract the S_K metrics, which corresponds to M independent pairs of antennas (i, j). The expected value of S_K conveniently equals one when the underlying distribution of $V_{i,j}$ is normal, i.e., in the no-RFI case.

The correlated astronomical and system noise contributions vary locally in the UV domain (the visibilities distribution is not “spatially” stationary), similarly to the RFI contribution. We propose to break the visibilities into local “clusters” of size M. The current distribution of clusters assumes L=3,000 clusters with M=1,000, although these values may have to be optimized and should remain parameters of the algorithm so as to optimize the flagger performances with real telescope data. We note that some visibilities may be associated to multiple clusters without loss of performance. An example of visibilities clusters is shown in Figure 4. The distribution of clusters is set once for all, does not have to be recalculated, and holds for any observing parameters (e.g. Figure 4 and Figure 5 show the clusters for observations at

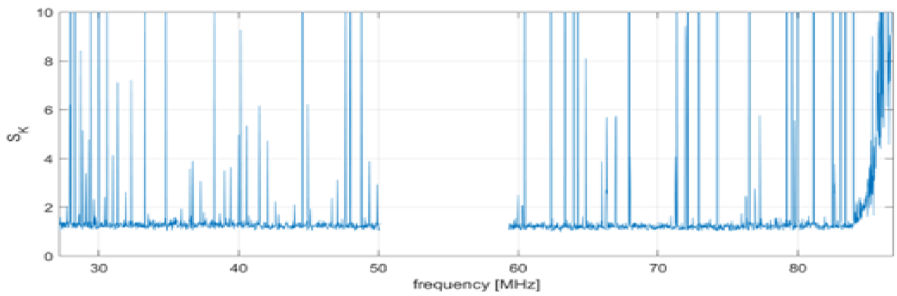

Figure 3. Spectral Kurtosis as a function of frequency (LWA data)

declinations 88 degrees and 5 degrees, respectively).

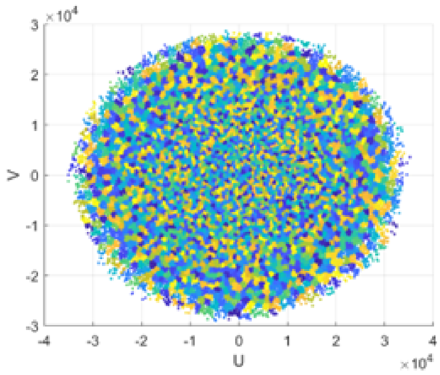

Figure 4. Clusters of visibilities at Dec = 88 degrees

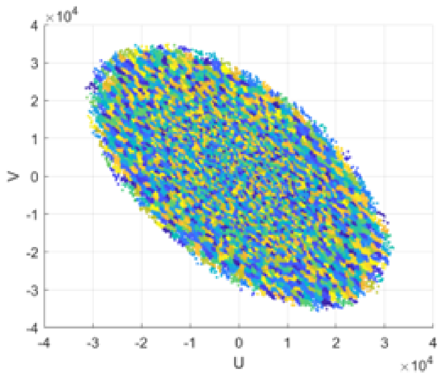

Figure 5. Clusters of visibilities at Dec = 5 degrees

An additional step involving the subtraction of contiguous channels is included in the algorithm. This subtraction significantly improves the sensitivity of the flagger to RFI by attenuating the contribution of correlated astronomical signals that exhibits similar statistical characteristics over neighboring channels. The flagger computes then S_K values for individual clusters or frequency-channel-subtracted visibilities. S_K values exceeding a given parametrized threshold above or below the expected value "1" follow non-Gaussian distribution, which is a consequence of the presence of RFI. All visibilities belonging to the flagged cluster in both frequency channels considered are flagged as erroneous and set to zero. A binary visibility flags array is stored for log keeping.

Figure 3 shows the SK values for one center cluster made of 1,000 visibilities of the previous LWA data set centered around (U,V) = (56 λ, 46 λ). A small offset from the center value "1" is noted and is related to variations of the actual data distribution from an actual Gaussian distribution due to correlated noise. A threshold can then be set to identify outlier frequency channel pairs and flag them for that one cluster.

5 Conclusion

We present in this paper the real-time flagging concept for the DSA-2000 telescope. After introducing the telescope architectue and the need for real-time flagging, we describe the strategy for flagging corrupted at data at the F-engine and X-engine levels. Further work will evaluate the performances of the proposed approaches with real and simulated telescope data, and conclude on the effectiveness of the method and the expected detection and false alarm rates.

6 Acknowledgements

The author wishes to thank Schmidt Sciences for supporting this work.

References

[1] Romney JD. Cross correlators. In Synthesis Imaging in Radio Astronomy II 1999 (Vol. 180, p. 57).

[2] Thompson AR, Moran JM, Swenson GW. Interferometry and synthesis in radio astronomy. Springer Nature; 2017.

[3] Lynch RS, Smith ET, Harrison M. Radio Frequency Interference Excision Using Cyclostationary Signal Processing. InThe RFI2022 Workshop at ECMWF 2022 Sep.

[4] Perley R. EVLA Memo 49 Attenuation of Radio Frequency Interference by Interferometric Fringe Rotation.

[5] https://online.stat.psu.edu/stat414/lesson/26/26.3

[6] Nita GM, Gary DE. The generalized spectral kurtosis estimator. Monthly Notices of the Royal Astronomical Society: Letters. 2010 Jul;406(1):L60-4.