D ATA PROCESSING PIPELINE FOR DECAMETER PULSAR/TRANSIENT SURVEY

Vasylieva I.Y.[1,2], Zakharenko V.V.[1], Zarka P.[2], Ulyanov O.M.[1], Shevtsova A.I.[1], Seredkina A.A.[1]

[1] Institute of radio astronomy of NAS of Ukraine
Kharkiv, Ukraine
[2] LESIA, Observatory of Paris of CNRS of France
Meudon, France
iaroslavna.vasylieva@gmail.com

ABSTRACT. The study of the low-frequency radio sky at short time scales provides the insights into the nature of transient radio sources as well as gives a sensitive tool for interstellar medium sensing. The decameter survey of the northern sky is aimed to discover new sources of transient and repetitive emission within more than 100 Tb of the high time resolution data. The efficient pipeline of data cleaning and processing is developed and successfully tested.

Key words: transient emission, pulsars, cleaning procedure

1. Introduction

The sources of rapidly varying pulsed radio emission (pulsars) have been investigated in the decameter range for more than 40 years. It was revealed that this range is very promising in terms of detecting new phenomena in their behavior. For example, the phenomenon of the anomalously intense pulses was discovered in the decameter range [1]. The sporadic behavior of certain pulsars is more noticeable at low frequencies [2]. The precision of interstellar medium sensing is highest in decameter range, as the dispersion measure and rotation measure increase as a square of frequency. Transient radio emission which can also be generated by neutron stars may also exhibit many interesting peculiarities at low frequencies.

The study of transient sources of radio emission is one of the key projects for the new generation of radio telescopes such as LOFAR and SKA [3,4]. The SKA mainly explores the higher band of radio spectrum, whereas LOFAR has the operating frequency range 10-240 MHz but is still optimized to work above 30MHz. Therefore the lowest part of radio spectrum, observable from the ground (10-30 MHz), will not be covered enough by investigation of transient radio sources.

The most sensitive radio telescope in this frequency range is still the UTR-2 (Ukrainian T-shaped Radio telescope, 2nd modification). It has shown the significant success in the detection of pulsed and transient radio emission [5,6]. Until recent time the sensitivity of the telescope was restricted by the receiving system, and the limited capacity of data storage prevented the massive study of all pulsar sample. Now, with the introduction of digital backend it became possible to measure the pulsar characteristics precisely and to draw statistical conclusions from the large sample of observed pulsars [5].

The low-frequency range provides special benefits for the study of neutron stars which are known to be sources of pulsed and transient emission. These benefits are connected with expansion of their beaming fraction at lowest frequencies [5], and they allow discovering new sources which are not reachable for higher-frequency observations. This fact and the statistics of the pulsar behavior at decameter wavelengths have revealed the possibility to detect up to 100 new pulsars and/or transients in the decameter range. This was the motivation to commence a new survey of northern sky with relatively high time resolution (8 ms). In this survey we expect to discover nearby slow pulsars which haven't been detected so far due to the orientation of their narrow radiation cones, which get broader at low frequencies as well as the bursts from RRAT-like sources and other transients.

2. The data cleaning and processing pipeline

The survey strategy is related to the peculiarities of UTR-2 and is a compromise between the integration time for each source (which must be large) and the total observation time (must be the least possible). This contradiction was resolved in the technique using the Earth's rotation and the wide beam pattern of antenna. Thus, the entire northern sky can be surveyed in 40 days, implying the storage and processing of ~100 terabytes of the data. To confirm the detections, the re-observations of candidate-sources will be needed, and to detect more transients, the multiple records of the sky are essential. It will lead even to larger data amounts to be processed. The fast an efficient data processing pipeline is therefore needed to deal with all the observational data in the automatic mode.

The processing of survey data includes the following stages: (i) the interference excision, (ii) procedure of dispersive delay compensation with different trial values of dispersion measure (DM) constant (hereafter, dedispersion), (iii) search for individual dispersed pulses and (iv) search for periodicity in the processed data.

2.1. The interference excision

The most tricky and important task, the data cleaning procedure, is needed for both pulsar and transient pipelines. Due to the large area, occupied by the beam on the sky, there are much terrestrial interference and signals from other cosmic sources captured by the beam. All these artificial or natural signals may show up as the false-positive detection of the spurious sources, giving certain periodicities in the Fourier domain or bright spots on the individual bursts search diagrams.

The level of the interference varies from time to time, depending on the time of the day, season etc. Therefore the different depth of cleaning should be used during the processing. In our case the most reasonable way is to perform an adaptive procedure, consisting of several stages, which adds or removes the additional stages depending on the percentage of data samples, contaminated by the interference. The thresholds used in the cleaning should be more severe when searching for periodic weak pulsar signals, which are normally under the noise level, and less severe when identifying the transient signals, which are supposed to be rather strong.

The implementation, best fitting to the criteria above, is made up of several stages of cleaning in the time and frequency direction.

Initially we calculate iteratively the mean and standard deviation (σ) in each frequency channel, clipping the values above 3 σ at each iteration. Then we normalize the data, to eliminate the amplitude-frequency characteristic unevenness of the telescope and to make data zero-mean. The normalized data are shown in Fig. 1, a). The intense (black) features are all from terrestrial origin, and are the interference.

After that we make two stages of cleaning the 'bad' (interference) samples. Fig 1 shows the cleaning stages in the same order as in the pipeline. The underlying principle of first stage is described in [7], in section 'The SumThreshold method', and adapted to UTR-2 data. The main point is that the different thresholds are applied to the data, and the connected shapes of samples exceeding each threshold are found. A combination of N samples (either in time or in frequency direction) is flagged as interference if its average exceeds the threshold T_N (given in the absolute value or relative to the standard deviation σ). The sliding window size N is increased from iteration to iteration, and the larger is N, the lower is the threshold, applied to the average of the samples. The values that have been flagged in the previous iteration are not taken into account in the subsequent one. For UTR-2 data after normalization (zero-mean), the sizes of N and the corresponding thresholds differ for the time and frequency direction. They are listed in the table 1.

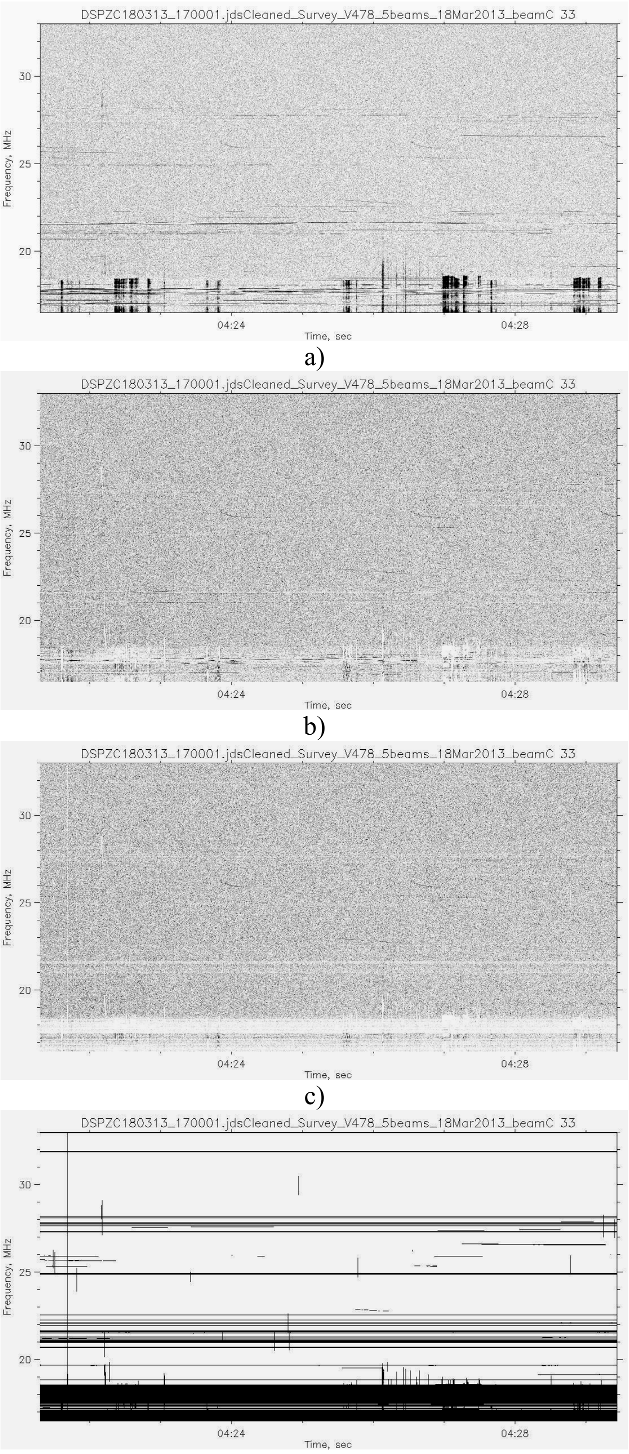
Table 1

Time direction:					
N	2	8	16	128	256
T_N	$6.67\cdot\sigma$	$2.96\cdot\sigma$	$1.97\cdot\sigma$	$0.58\cdot\sigma$	$0.4\cdot\sigma$
Frequency direction:					
N	1	2	4	8	64
T_N	$10\cdot\sigma$	$6.67\cdot\sigma$	$4.44\cdot\sigma$	$2.96\cdot\sigma$	$0.88\cdot\sigma$

The result of the performance of the described can be seen in Fig. 1, b).

The last stage is the modified version of cleaning procedure, described in [5], without examining of single samples (it is moved to the 1^{st} stage). To eliminate the low-level wideband and narrowband interference, we accumulate the data by frequency or time, correspondingly. After that the σ of integrated time series (or of average spectrum) is re-calculated and the spikes exceeding the 4σ-level are flagged (Fig. 1, c)).

All the flagged samples are stored in binary 'bad samples map' (Fig. 1, d)), where 1 is a good sample, 0 is a bad one. They are replaced by the median, calculated over all 'good' samples of the dynamic spectrum.



a)

b)

c)

d)

Figure 1: The successive stages of data cleaning procedure

When the data are rather clean, only with short features along the time or frequency direction, the first stage of the cleaning algorithm is applied. If the data contains numerous wideband interference or continuously broadcasting radio stations' signals, we add the remaining two stages of the cleaning to the pipeline. At this step, the whole frequency channels or time intervals are removed from the data.

This sophisticated algorithm shows sufficient performance visible by naked eye and is well adapted to UTR-2 data. The average value of data, lost due to interference during one observational night, is about 4%. Mostly it is the data of the first and the last hours of observations, whereas the middle of the night tends to be almost interference-free.

2.2. Dispersive delay compensation

The clean data are then dedispersed with different dispersion measure (DM) trials. Initially we restrict the upper value of DM as 30 pc·cm^{-3}, but later this value will be raised to 60 pc·cm^{-3} for pulsars, and 100 pc·cm^{-3} for transients. We have chosen the step of trial DMs to be 0.01 pc·cm^{-3}. The precision could be higher, but we have selected this value as an optimum between the precision and the processing time. The precision reached in the higher-frequency surveys is less, because the propagation effects are not so influential there. Due to this, the decameter range is the unique probe of the interstellar medium via its dispersion, scattering and the Faraday effect.

2.3. Search for individual dispersed pulses

The further step of the pipeline is a search for individual transient events, which implies examining the data for dispersed pulses, exceeding certain threshold. At this stage we potentially can detect the giant pulses or anomalously intense pulses of pulsars, RRAT-like signals etc. The nearby pulsar will show up as a series of the intense pulses at the same dispersion measure trial value. To reduce the fluctuations of the background noise and to distinguish the useful signal, we normally apply the filtering of the low-frequency components of the data with a cut-off frequency of high-pass filter about 0.2 Hz. To increase the sensitivity, we also use the integration of 4 time samples into 1 (32 ms).

The detection algorithm denotes the suspected transient events by the circles, with the diameter proportional to the signal-to-noise ratio of the event. To reject the spurious events, we return to previous stages of processing and check the event visually. We check whether the cleaning has removed only obstructive signals, whether after dedispersion with the corresponding trial DM the signal is aligned into a straight vertical line on the dynamic spectrum and whether there are several events at the same dispersion measure (then we assume the detection of a pulsar). Fig. 2 shows the result of algorithm performance at the strong pulsar B0809+74 with the dispersion measure 5.755 pc·cm^{-3}. We see the repetitive events of different intensity, with intervals multiple of pulsar period.

2.4. Search for periodicity

The search for weak pulsar signals is possible due to their periodicity. We make the Fourier analysis and fold the integrated time series with the trial periods of pulsar. As we restrict the lower value of pulsar periods in the survey down to 0.1 s, we should consider the harmonics less than 10 Hz in the Fourier domain. The higher harmonics are filtered out. The Fourier transform is applied to each integrated time series, obtained by summing the shifted frequency channels with respect to one another according to the DM trial value. To increase the signal-to noise, we sum several FFT outputs of neighboring DM trials. To increase the power of the first harmonic, we add the 2, 4, 8 and 16 harmonics to it. The result is the peaks of certain shape, outstanding above the noise. Each of these peaks corresponds to 1/trial period, which we will depict on the period-dispersion measure plane, looking for characteristic shape of average dispersed pulses.

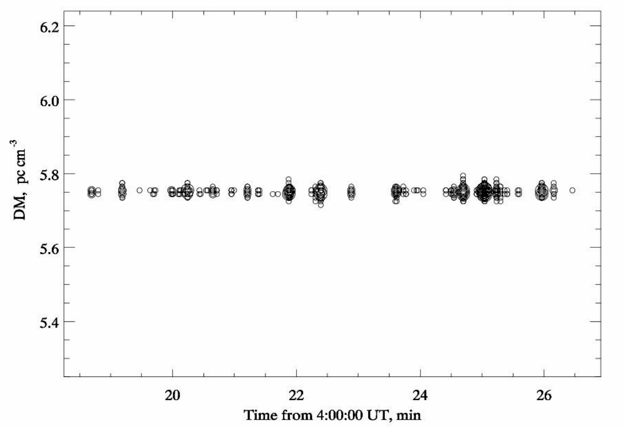


Figure 2: The result of individual pulses search algorithm with pulsar B0809+74

3. Results and perspectives

Up to now, by means of the described pipeline we have processed about 25% of survey data. A large number of transient-candidates are obtained, and we are excluding the spurious events from them.

The upper value of DM=100 pc·cm^{-3} will lead into 10 000 different shifts to be made between the frequency channels in the data. For 100 Tb of data, this will take a lot of processing time. To accelerate the data processing, we are introducing the techniques of parallel and distributed computing to the pipeline. The GPU-based dedispersion and processing of the different data fragments at the available nodes of Ukrainian Academical Grid will speed-up the processing up to 200 times. Another alternative is the usage of volunteer distributed computing system BOINC [8]. The Einstein@Home project, powered by BOINC, was used to discover 24 new pulsars in Parkes Multibeam Pulsar Survey data [9].

References

1. Ulyanov O. et al.: 2006, *Radiofizika and Radioastronomia*, **11**, 113.
2. Ulyanov O., Zakharenko V.: 2012, *Astron. Rep.*, **56**, 417.
3. Stappers B.W. et al.: 2011, *A&A*, **530**, id.A80.
4. Colegate T.M., Clarke N.: *PASA*, **28**, 4, 299.
5. Zakharenko V. et al.: *MNRAS*, **431**, 4, 3624.
6. Popov M. V. et al.: 2006, *Astron. Rep.*, **50**, 562.
7. Offringa A.R.: 2012, University of Groningen, *diss.*
8. *http://boinc.berkeley.edu/*
9. Knispel B. et al.: 2013, *ApJ*, **774**, 93.