

A LOFAR RFI detection pipeline and its first results

A.R. Offringa*

Kapteyn Astronomical Institute, University of Groningen, The Netherlands

E-mail: offringa@astro.rug.nl

A.G. de Bruyn,

Kapteyn Astronomical Institute and ASTRON, The Netherlands

E-mail: ger@astron.nl

S. Zaroubi,

Kapteyn Astronomical Institute, University of Groningen, The Netherlands

E-mail: saleem@astro.rug.nl

M. Biehl

Institute for Mathematics and Computing Science, University of Groningen, The Netherlands

E-mail: m.biehl@rug.nl

Radio astronomy is entering a new era with new and future radio observatories such as the Low Frequency Array and the Square Kilometer Array. We describe in detail an automated flagging pipeline and evaluate its performance. With only a fraction of the computational cost of correlation and its use of the previously introduced SumThreshold method, it is found to be both fast and unrivalled in its high accuracy. The LOFAR radio environment is analysed with the help of this pipeline. The high time and spectral resolution of LOFAR have resulted in an observatory where only a few percent of the data is lost due to RFI.

RFI mitigation workshop - RFI2010,

March 29-31, 2010

Groningen, the Netherlands

*Speaker.

1. Introduction

Now that radio telescopes of the next generation, such as the Low Frequency Array (LOFAR), the Murchison Widefield Array (MWA) and the Square Kilometer Array (SKA), are coming into operation, the dawn of software-driven telescopes producing terabyte sized data sets has begun. Because of that, data reduction in radio astronomy is entering a new era in which more emphasis is put on automated data processing and pipelining the various steps in the data reduction. One important step in the reduction process is dealing with radio frequency interference (RFI). Although trivial techniques performed by the data reducing scientist, such as manual baseline, time and frequency selection, have been sufficient albeit tedious for the last decades, more sophisticated automated flagging procedures are required for the next generation telescopes.

RFI mitigation can be performed pre- as well as post-correlation. Pre- and post-correlation techniques are mostly complimentary: they find or remove different kinds of RFI. Hence, the implementation of one does not make the other obsolete. The pre-correlation mitigation stage can detect RFI bursts of a sub-integration time changing nature with minimal loss of data, though has to be executed *on-line*, implying the requirement of real-time execution on small domains, in particular on a small time interval with respect to the entire observation. Examples of pre-correlation methods are based on thresholding [15, 9, 2, 10]; spatial filtering with eigenvalue decomposition [9, 13, 5]; and adaptive cancellation with a reference antenna [3].

To deal with RFI, the post-correlation phase is the final resort. Demonstrated techniques include the use of an independent RFI reference signal to subtract RFI [4]; an approach using singular value decomposition [11, 12]; and fringe fitting [1]. Since RFI comes in many forms [6, 8] not all contaminated samples can be recovered, despite the numerous existing techniques. Therefore, flagging remains an important final step [11, 17, 16].

The situation for RFI flagging strategies in modern observatories such as the Westerbork Synthesized Radio Telescope (WSRT), LOFAR and the Giant Metrewave Radio Telescope (GMRT) has changed. On one hand, the high time and frequency resolutions make accurate flagging of contaminated samples available, resulting in smaller loss of data. On the other hand, radio quiet zones are harder to achieve, and all of the above mentioned telescopes are situated in populated areas. Moreover, sensitivity requirements for telescopes are growing. For example, one of the LOFAR key science projects is the LOFAR Epoch of Reionization (EoR) project [7, 14], a very ambitious project with high demands on sensitivity and noise behaviour.

Automated flaggers can be compared on accuracy, i.e., the true/false-positive ratio; the speed of the algorithm; robustness; and technical requirements that it imposes. Constructing a flagger that performs good on all aspects is challenging. In this article, we will describe a novel flagging strategy that has been designed to take this challenge, and has been implemented in the LOFAR observatory pipeline.

2. Method

In this section we will explain the flagger step by step. An overview of the flow of execution is given in Figure 1.

2.1 Input

The flagger is executed on the amplitude information of one polarisation of a single sub-band of a baseline. In LOFAR’s common operation, a sub-band consists of 256 channels of 0.8 kHz resolution. The full band has 248 sub-bands. LOFAR can observe in two bands: the 10-80 MHz low band and the 110-240 MHz high band, which are observed by physically different antennae.

If speed is essential, the algorithm can be executed once on the Stokes-I values. Otherwise, if accuracy is more important than speed, the algorithm can be executed on the individual XX and YY or LL and RR polarisations, or on all polarisations individually. We do see some RFI that manifests in only one of the polarisations, or rotates through the polarisations, and some advantage is therefore seen when flagging all polarisations individually.

2.2 Iterations

A part of the algorithm is iterated a few times, depicted in Figure 1 by the “Continue iterating” block. This is necessary for finding low-level RFI, as will be explained in the thresholding paragraph, §2.3. Iterations, however, are costly in term of speed, and should be kept to a minimum. To do so, the fit should converge quickly. We do this by entirely ignoring channels and time steps in the first surface fit that superficially look bad, yet might only have been partially uncontaminated. The extra information that might have been added if the uncontaminated part of the channel or time step was added does not change the fit much, and therefore is not slowing down the convergence.

It was determined that performing the fit two times is enough for a stable, accurate fit. This is true for all data that was tested, in special for both WSRT and LOFAR data, and for both clean bands and strongly contaminated bands.

2.3 The SumThreshold method

The SumThreshold method detects series of samples with higher values than expected. Consider the orthogonal slices $\rho_d(x)$ and $\Omega_d(x)$ through the fit-subtracted visibilities $R(t, \nu)$ and the flag mask $W(t, \nu)$ respectively, where $W(t, \nu)$ evaluates to zero if the corresponding value in the mask at time t and frequency ν has been flagged and one otherwise. The function parameter x is scaled to have a unity step size in time or frequency, for respectively a slice through the frequency direction when $d = 1$ or time direction when $d = 2$. The SumThreshold method can now be defined by the following decision rule:

$$\Omega_d^{n+1}(x) = \begin{cases} 0 & \text{if } \Omega_d^n(x) = 0 \vee \exists i \in \{0 \dots M_n - 1\} : \\ & \sum_{j=0}^{M_n-1} |\rho_d \odot \Omega_d^n(x+i-j)| > \frac{\chi_{M_n}}{\sum_{j=0}^{M_n-1} \Omega_d^n(x+i-j)} \\ 1 & \text{otherwise,} \end{cases} \quad (2.1)$$

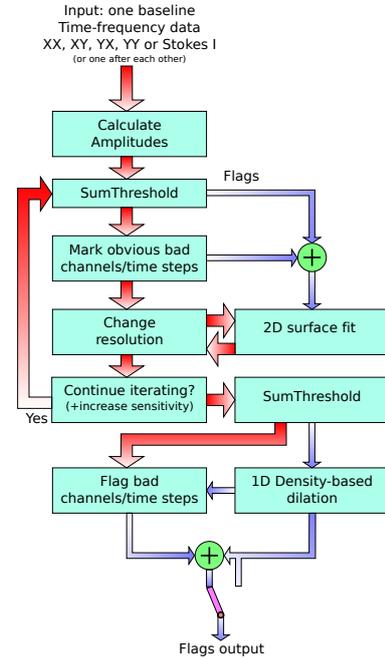


Figure 1: Overview of the RFI flagging strategy

with parameters $\mathcal{M} = \{M_1, M_2, \dots, M_{\max}\}$, the set of combination lengths and χ_{M_N} , the corresponding threshold levels for a specific combination length $M_N \in \mathcal{M}$. In the previous study [11], the `SumThreshold` was introduced and was shown to produce the highest accuracy of current post-correlation RFI detection algorithms. We refer to this study for detailed information about the `SumThreshold` method.

`SumThreshold` is performed in each iteration once, before the surface fit, in order to ignore RFI when fitting. It is performed one last time when the surface fit is expected to have been converged, to establish the actual flags. To increase the stability of the strategy, the sensitivity of the `SumThreshold` method starts low, i.e., it finds only the strongest RFI, and is exponentially increased each time it is executed.

2.4 Channel and time selection

After `SumThreshold` has found the contaminated samples, we observe especially after the first iteration, that some channels and time scans have not been fully flagged, even though they are fully contaminated. As explained in §2.2, this might slow down convergence, which is why a second step was implemented in order to completely, hence inaccurately but quickly, flag these channels and time steps before fitting.

In order to detect problematic channels and time steps, the values are compared based on their root mean square (RMS) values. The RMS series are Gaussian smoothed and if the difference exceeds 3.5 times the standard deviation of the sequence of differences, they are flagged completely. Optionally, this selection can be executed again as the last step in the algorithm.

2.5 Surface fitting

A surface fit is executed to subtract fringes caused by strong sources, in order to increase accuracy. Several fitting strategies have been tested, and all sliding window methods show similarly good results in terms of accuracy. In non-sliding window approaches such as the dimensional-independent polynomial fit described in [17], we have observed instability near the borders of the fixed windows.

A Gaussian kernel was found to produce the best average between speed, accuracy and stability. The accuracy is not significantly different from other sliding window fitting strategies, such as a dimensional-independent polynomial fit applied on a sliding window.

Since the fit is, relative to the other operations, a time-consuming operation, the input time-frequency matrix is rescaled before fitting. The time dimension is three times reduced before fitting, and the fitted Gaussians are interpolated to restore the original scale. No significant change in accuracy was observed, which underlines that the quality of the fit is, up to some point, not a crucial aspect of accurate detection.

2.6 Density-based dilation

It may be desirable to flag samples that are up to a few channels away from strong, continuous RFI. Thresholding does not flag these samples, if they are not significantly different in amplitude. Likewise, it may be desirable to flag more of a partially flagged channel, because a continuous transmitter might be recorded at different amplitudes, either because of different propagation of the

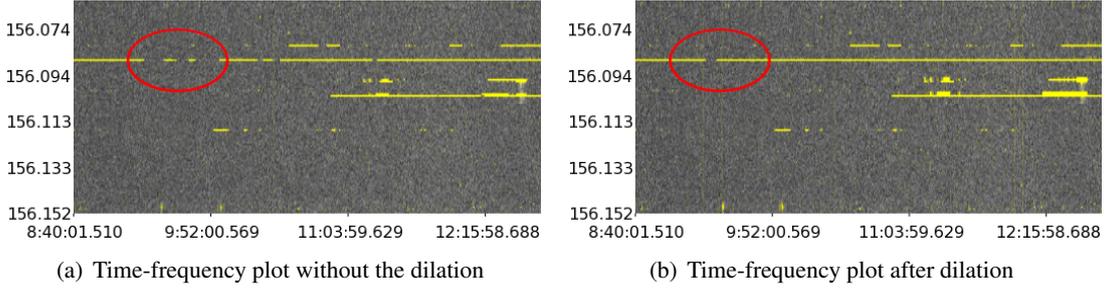


Figure 2: The result of a density-based dilation with $\eta = 0.9$: the flags in panel (a) are established by the *SumThreshold* method and dilated based on the flag density. The result is shown in panel (b). Noticeable differences are the small gaps in orthogonal lines that have been filled by the dilation, such as the area within the red ellipse. While this diagram displays over 6000 time steps, the algorithm also fills many invisible small holes: its behaviour is scale invariant.

signal, because of the transmitter moving in respect with the beam or because of a transmitter's intrinsically changing strength, and this might cause the received RFI not to trigger the threshold in some samples. To overcome this problem, we enlarge the flag mask after the apparent RFI has been flagged by the iterative procedure.

A typical approach in this problem is to perform a morphological dilation operation on the flag mask. For example, a dilation with a square mask of size $N \times N$ would enlarge each flag to a square of $N \times N$. Every sample, that has an orthogonal distance smaller than N samples from a flagged sample, would be flagged in this case. Although this technique is advantageous for its simplicity and establishment in the field of mathematical morphology, using this technique for the described purpose has the disadvantage of being inaccurate: it will typically flag too many samples when only a few samples are flagged in some area, while too few samples will be flagged when a channel or time step is almost completely flagged.

To correct for these problems, we introduce a dilation operation which mask size is related to the one dimensional flag density: the dilation mask is larger in dense areas and smaller in sparse areas, in respect to either the one dimensional time domain or frequency domain.

Consider an orthogonal slice $\Omega_d(x)$ through the flag mask as defined in §2.3. The following density-based decision rule is introduced:

$$\Omega'_d(x) = \begin{cases} 0 & \text{if } \exists Y_1 \leq x, \exists Y_2 > x: \sum_{y=Y_1}^{Y_2-1} \Omega_d(y) \leq \eta (Y_2 - Y_1) \\ 1 & \text{otherwise,} \end{cases} \quad (2.2)$$

where $\eta \in [0, 1]$ is the density ratio threshold. In words, this rule flags the samples that are in any constructable area $[Y_1; Y_2)$ with an unflagged sample ratio less or equal than η . Specifically, $\Omega'_d(x) = 0$ for all x if $\eta = 1$, while $\Omega'_d(x) = \Omega_d(x)$ for $\eta = 0$. Furthermore, since any element x with $\Omega_d(x) = 0$ will be in the single element area containing only itself, $\Omega_d(x) = 0 \implies \Omega'_d(x) = 0$. Consequently, the number of flags is increasing. Although a strict implementation of (2.2) will take $O(n^2)$ operations for n samples in the orthogonal slice $\Omega_d(x)$, by putting extra constraints on Y_1 and Y_2 , an $O(n \log n)$ implementation is possible without much loss of its accuracy. Figure 2 shows the result of such a dilation on actual data.

3. Computational requirements

Table 1 shows an estimate of the required floating point operations per sample for each individual step. The total number of operations required is on the order of 300 floating point operations per sample. In a typical full LOFAR observation, the correlator will output 4 polarisations \times 256 channels/sub-band \times 248 sub-bands \times $\frac{50^2}{2}$ baselines \times 1 sample/second \approx 0.3 gigasamples per second, yielding a computational requirement of \sim 0.1 TFLOP/s in the best flagging mode.

Although this is only a small fraction of the required computations for correlation, some simplifications can be made to lower the computational requirements. Techniques to improve the computational performance include: flagging on Stokes-I values; using a larger resizing factor before fitting; using a smaller window size; and determining the cross-correlation flag masks using autocorrelations.

The LOFAR flagging pipeline will be run on an off-line computing cluster. The flagging pipeline is parallelized by running each sub-band on a different computational node, and the flagging of the individual sub-bands is executed by a multi-threaded implementation. Concluding from the interpolation of the performance of the current implementation of the pipeline, which achieves processing 27 stations in a quarter of the observing time with its most computational expensive flagging strategy, real-time performance can be realised in a full 50 station LOFAR.

4. Input/output requirements

Processing baseline by baseline in a pipeline has implications for the software architecture of the observatory: since baselines are correlated simultaneously, the observed visibilities have to be written to disk before running the RFI pipeline, which is inefficient. After finishing observing, the flagging pipeline can read the data in the required order. However, flagging is normally followed by tasks such as calibration and source subtraction. These tasks expect time-sorted data, thereby requiring a second read of the data in its previously observed order. Since the architecture of LOFAR allows this flow of processing, and because of the advantages of baseline by baseline flagging in terms of accuracy and computational speed, the input/output-overhead caused by this deficiency is ignorable. This however might become a serious issue in even larger telescopes such as the SKA.

5. Flagging results

The implementation³ of the algorithm was tested on several LOFAR observations. Currently,

¹Floating point operations per sample

²These are actually integer operations, since this step uses the masks.

³The software implementation of the presented RFI pipeline has been made publicly available and can be downloaded from the following location: <http://www.astro.rug.nl/rfi-software/>

Table 1: Computational requirements of the RFI pipeline

Step	F/smp ¹	Count	Total F/smp ¹
Calculating amplitudes	4	1	4
Time/frequency selection	2	3	6
SumThreshold	20	3	60
Change resolution	4	2	8
Surface fit	50	2	100
Density-based dilation	100 ²	1	100
Total			278

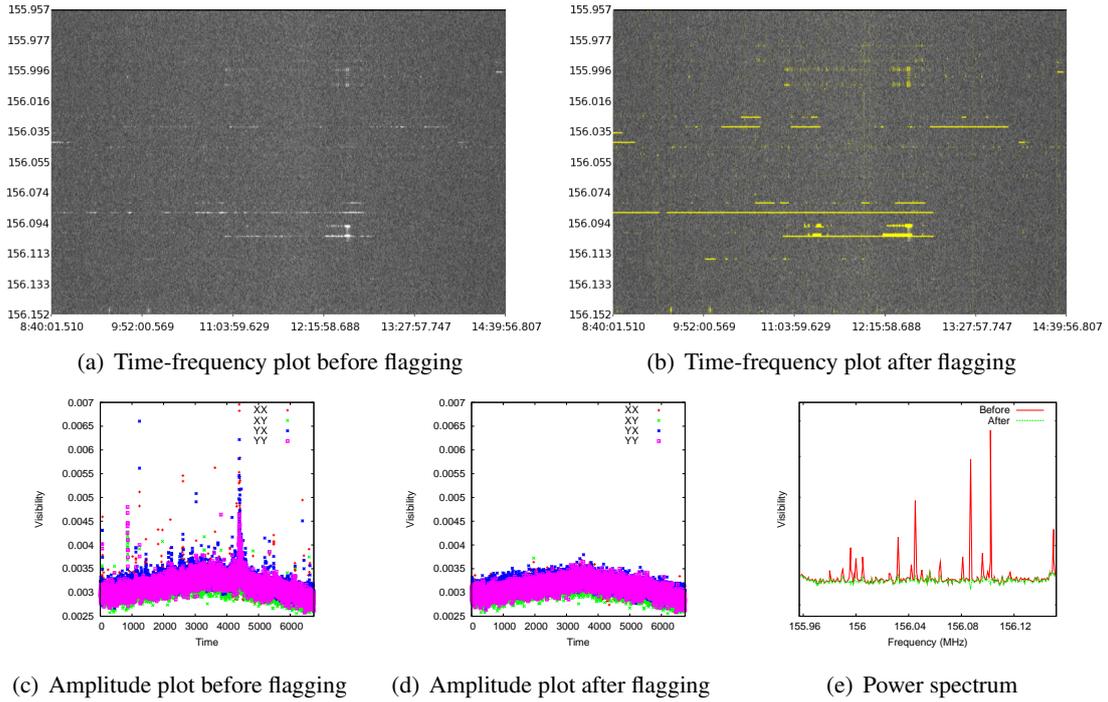


Figure 3: Flagging results of the 6 hour LOFAR observation L2010_07096 of April 24, 2010. All plots show the same randomly chosen sub-band around 156 MHz for a 1.5-km baseline ($CS302_HBA1 \times CS005_HBA0$) with three second integration time. The flagging pipeline was run with its default settings, and 1.8% of the data is flagged. As can be seen from panels (a), (c) and (e), this sub-band contains relatively many interfering transmitters, yet all of them are relatively weak. The panels (b), (d) and (e) show the cleaned band after flagging.

27 of the approximately 50 total LOFAR stations are ready. Flagging a single sub-band of a 6 hour observation with the 27 stations takes 90 minutes on a single cluster node. This implies real-time flagging speed for the full 50 station LOFAR that will produce four times more data. All RFI that can be found by visual inspection is typically flagged, thereby outperforming simpler methods such as a median absolute threshold filter in both accuracy and speed. An example result can be found in Figure 3.

The pipeline is, by the use of the `SumThreshold` method, very accurate. In some cases, the algorithm finds RFI which is invisible by eye on full scale time-frequency diagrams, but becomes only apparent when zooming in on the data and integrating certain cuts of the data cube. In the band shown in figure 3 an interferer is visible at approximately 156.03 MHz. Although it is visible as a small bump in the time integrated spectrum in Figure 3(e), it is not apparent in the time frequency plot of Figure 3(a). Nevertheless, the algorithm finds the samples that are contaminated by the interferer, and the particular bump at 156.03 MHz in Figure 3(e) is flattened.

On the other hand, if an interferer has a smooth time-frequency profile, it will be mistaken for astronomical data and will not be flagged. In these situations it might help to subtract a rough model for the celestial signal and increase the flagger's sensitivity.

6. LOFAR RFI environment

LOFAR breaks the tradition of building telescopes in sparsely populated areas, with its core being installed in the North-East of the Netherlands. Although the core is in a nature reserve, and therefore in a sparser populated part of the Netherlands, all the stations are relatively close to farms, roads and some nearby municipalities. Now that LOFAR is half-way ready and performing representable observations, we can start to evaluate the dynamic radio environment.

The first results of RFI mitigation show several promising characteristics of the LOFAR site. First of all, hardly any broadband RFI is observed. If observed, it is typically caused by electrical fences, lightning, power cables, hardware in situ, cars and trains. It can be concluded that the site is sufficiently remote and hardware on site is sufficiently shielded to prevent these interferences. Only one of the stations is close to an electrical fence that surrounds a farming meadow, causing broadband spikes every two or three seconds. The flagging pipeline flags 40% of the data in this station, and this station is therefore currently not useful. Options include negotiation with the farmer to switch off the electrical fence during observations or implementing an RFI nulling method in the station that nulls spikes on a high time resolution.

A second class of interferers are constant transmitters at a fixed frequency, such as FM radio. The FM range lies between the physically separated low and high bands. Transmitters in this range are therefore effectively blocked by the bandpass filters. Other constant sources that do transmit within the observing frequency often occupy only one or a few 0.8-kHz channels, which, after the pipeline has flagged these transmitters, cause only a minimal amount of data loss. While many of the sub-bands of 256-channels are completely clean of such constant transmitters, others have a few of such transmitters, such as the one shown in Figure 3.

A third class of interferers are transient sources with variable frequency. These occur mostly at random and their exact origin is often unknown. Some of these can be caused by moving objects, such as meteors or airplanes, that reflect a distant signal for a short period.

Considering all classes of interferers, typical observations with representable stations show only a few percent of data loss due to interference.

7. Conclusion and discussion

Radio astronomy is entering a new era with futuristic observatories such as LOFAR and the SKA. In this article we have presented a flagging technique that has shown the ability to operate accurately and efficiently on the LOFAR observations. Therefore, this technique is also a good basis for future observatories.

Because the computational costs of the RFI pipeline are only a fraction of the correlation costs, efficiently ordering the data before presentation to an RFI algorithm is the largest challenge, rather than optimising the computational costs. The pipeline also stipulates the importance of flexibility in an observatories' architecture, which adds freedom to design decisions. The LOFAR architecture also allows more sophisticated variations that include RFI mitigation at station level and different pipelines based on the observation mode. With the example of a complicated pipeline as described in this paper, it can be concluded that other algorithms such as transient detection and other pattern recognition techniques can be implemented in a similar manner in the pipeline.

Both the software and hardware of LOFAR are still under construction. The first observations of LOFAR nevertheless show very good prospects for the telescope, with only a few percent lost data due to interferers and, highly important, neither broadband nor in situ interference is commonly seen. The next step in RFI mitigation is to produce and analyse images on a maximum dynamic range, in order to analyse the effects of possible weak RFI that is undetectable in post-correlated time frequency domains. Prevention of new transmitters remains very important, and establishment of a radio-quiet zone, especially around the core, is recommended.

In order to improve data quality further, pre-correlation techniques might be added at station level or during correlation. An interesting improvement to the robustness of a correlator might be to execute the `SumThreshold` method prior to correlation. Considering the accuracy gain of the `SumThreshold` compared to normal thresholding, and considering the correspondence of RFI on small and large timescales, implementing this pre-correlation method on the highest time resolution data might improve blanking accuracy further.

Acknowledgments

We thank the LOFAR collaboration for providing the data on which the flagger was developed and tested.

References

- [1] R. Athreya. A new approach to mitigation of radio frequency interference in interferometric data. *AJ*, 696:885–890, April 2009.
- [2] W. A. Baan, P. A. Fridman, and R. P. Millenaar. Radio Frequency Interference Mitigation at the Westerbork Synthesis Radio Telescope: Algorithms, Test Observations, and System Implementation. *AJ*, 128:933–949, August 2004.
- [3] C. Barnbaum and R. F. Bradley. A new approach to interference excision in radio astronomy: Real-time adaptive cancellation. *AJ*, 115:2598–2614, November 1998.
- [4] F.H. Briggs, J. F. Bell, and M. J. Kesteven. Removing radio interference from contaminated astronomical spectra using an independent reference signal and closure relations. *AJ*, 120:3351–3361, December 2000.
- [5] Steven W. Ellingson and Grant A. Hampson. A subspace-tracking approach to interference nulling for phased array-based radio telescopes. *IEEE Trans. on Antennas & Propagation*, 50:25–30, January 2002.
- [6] P. A. Fridman and W. A. Baan. RFI mitigation methods in radio astronomy. *A&A*, 378:327–344, October 2001.
- [7] V. Jelić, S. Zaroubi, P. Labropoulos, et al. Foreground simulations for the LOFAR-epoch of reionization experiment. *MNRAS*, 389:1319–1335, September 2008.
- [8] John J. Lemmon. Wideband model of man-made hf noise and interference. *Radio Science*, 32:525–539, March 1997.
- [9] Amir Leshem, Alle-Jan van der Veen, and Albert-Jan Boonstra. Multichannel interference mitigation techniques in radio astronomy. *ApJS*, 131:355–373, November 2000.

- [10] N. Niamsuwan, J. T. Johnson, and S. W. Ellingson. Examination of a simple pulse-blanking technique for radio frequency interference mitigation. *Radio Science*, 40, June 2005.
- [11] A. R. Offringa, A. G. de Bruyn, M. Biehl, S. Zaroubi, G. Bernardi, and V.N. Pandey. Post-correlation radio frequency interference classification methods. *MNRAS*, 405:155–167, 2010.
- [12] Ue-Li Pen, Tzu-Ching Chang, Jeffrey B. Peterson, Jayanta Roy, Yashwant Gupta, Christopher M. Hirata, Julia Odegova, and Kris Sigurdson. The GMRT EoR Experiment: Limits on Polarized Sky Brightness at 150 MHz. *MNRAS*, 399:181–194, 2009.
- [13] Bart Smolders and Grant Hampson. Deterministic RF nulling in phased arrays for the next generation of radio telescopes. *IEEE Antennas & Propagation magazine*, 44:13–22, August 2002.
- [14] R. M. Thomas, S. Zaroubi, B. Ciardi, et al. Fast large-scale reionization simulations. *MNRAS*, 393:32–48, February 2009.
- [15] R. Weber, C. Faye, F. Biraud, and J. Dansou. Spectral detector for interference time blanking. *A&AS*, 126:161–167, 1997.
- [16] B. Winkel, J. Kerp, and P.M.W. Kalberla. Data reduction strategy of the Effelsberg-Bonn HI Survey (EBHIS). *Proc. Panoramic Radio Astronomy: Wide-field 1-2 GHz research on galaxy evolution, PoS(PRA2009)043*, June 2009.
- [17] B. Winkel, J. Kerp, and S. Stanko. RFI detection by automated feature extraction and statistical analysis. *AN*, 88:789–801, 2006.