Detection and visualisation of radio frequency interference

Philippa Hillebrand
Gerard Nothnagel

June 17, 2014

1 Project description

The MeerKAT radio telescope array is currently being built in the Karoo, South Africa, and will be among the most sensitive radio telescopes in the world[16]. Despite the site's remote location, radio frequency interference (RFI) is an ongoing issue. RFI, a type of electromagnetic interference can cause errors in recorded astronomical data by overloading the telescope receivers, or causing ripples in the observed spectrum[7]. This means that it is difficult to form a clear picture of the sky and make accurate observations on the astronomical data. A declaration of an area as radio-quiet (no radio transmissions allowed) is not sufficient to remove RFI from a region[3], so post-correlation (after the data collected by multiple dishes has been put together) techniques are required. RFI can be characterised by type, giving a clear idea of the RFI environment surrounding the telescope[21, 7].

RFI data is collected by an on-site antenna. Due to the volume of data collection at MeerKAT, it is not possible to store raw data permanently. In situ visual analytics can be used to extract useful information while the data is in memory [23]. Another challenge is that large data sets are often ambiguously represented in lower dimensions; different data instances might map to identical positions in screen coordinates [22]. It is therefore important to intelligently reduce raw data, but without losing important information. A prototype RFI management system has been implemented, which performs both RFI detection and radio frequency visualization. In this project, we will further develop the prototype, investigating alternative visualizations and detection algorithms applicable to RFI environments.

2 Related Work

2.1 Detection

Basic thresholding methods such as spectral kurtosis[1, 17] and combinatorial methods (designed for the Low-Frequency Array (LOFAR) by Offringa et al[19]) form the basis for most RFI detection algorithms as they are the simplest type. Further work has been done, again at the LOFAR telescope by Offringa et al, on designing flagging techniques which accurately flag as much of the RFI as possible. These are the AOFlagger [18], and the morphological algorithm[20]. Other more specific algorithms have also been designed, such as PEACE[14] which is aimed specifically at finding Pulsar data, and Spatial Filtering[4, 12] which is a post-correlation mitigation algorithm.

The thresholding techniques and morphological algorithm are of special interest as they show both the base line and the possibility of extending algorithms.

2.2 Visualization

The imMens project [15], from the Stanford Visualization Group, was designed to support user interaction with large data sets. Data sizes ranged from thousands to billions of records and maintained a steady frame rate of 50 frames per second on commodity hardware. The project addressed perceptual and interactive scalability through the use of data reduction techniques. The visualization limits scalability based on the resolution of the visualized data rather than the number of records, and preserves both global and local data features (e.g. densities and outliers respectively).

Keim et al. [11] address the same problems of mapping large data sets to screen coordinates, but focus on reducing over plotting in image space and utilising unused screen real estate. Their solution of Generalised Scatter Plots use distortion techniques accompanied with clustering algorithms and Voronoi tessellation. A number of real-world applications have successfully made use of this technique, which demonstrates its effectiveness.

Similar scalability challenges are likely to arise for the MeerKAT project, such as visualizing an hour's worth of data compared to visualizing data collected over two weeks. The radio frequency data sets contain more points than pixels, and the two related projects provide insights into techniques which are able to exploit limited image space.

3 Problem Statement

3.1 Detection

RFI and astronomical signals (radio waves produced by a source) both come in many different forms, which makes detection of RFI difficult. Also, the amount of data recorded by a radio telescope is very large, so any detection algorithm is required to be as efficient as possible. As such the following question will be investigated:

Is it possible to adapt an existing detection algorithm to the supplied data, and add any form of characterization to that algorithm? As seen in the past work, Offringa et. al.[18, 19, 20] have worked extensively on detecting RFI in array type telescopes. The data for this project, however, is collected, formatted, and stored differently. The challenge is therefore to apply existing methods to the new data. The characterization of a particular signal has not been researched in as great depth, and so to design an algorithm to appropriately characterize the signals may be beyond the time scale of this project.

3.2 Visualization

There are both technical and perceptual challenges which affect visualization design: only a limited amount of pixels are available on display devices and humans have a limited cognitive ability [5]. In this project, we evaluate visualization techniques applicable to MeerKAT and the following two research questions are explored:

Which visualization methods are most effective for RFI detection?
The visualization prototype incorporates techniques such as line graphs, waterfall charts (similar to heatmaps) and bar charts. These techniques are helpful for managing large RFI data sets at the site, but there is room for further improvement. Alternative visualization methods which exploit modern computation exist [11, 15], and it is worth investigating how RFI detection can benefit therefrom.

How can large amounts of data be mapped to screen coordinates while giving a fair representation of the data?
Data reduction techniques such as sampling might miss important structures within the data [15], as different types of sampling preserve certain features while losing others [2]. It is important to experiment with different methods of data aggregation, and to determine which features are relevant to RFI detection. A challenge related to this topic includes giving an overview of the data set, in a limited display area, without losing necessary details.

4 Procedures and Methods

4.1 Detection

- Choose an algorithm previously designed to be a base for the project.
- Adapt the chosen algorithm to work with the data supplied by Christopher Schollar and the SKA team.
- Choose a platform on which to implement the algorithm, bearing in mind the particular file type used to store the data and the implementation of previous work.
- Implement the adapted algorithm in the chosen language.
- Create test data to validate the implementation, and perform tests using both this and natural data.
- Identify characteristics that may be visible in the supplied data.
- Choose one or possibly two characteristics to try and automatically identify.
- Adapt algorithm to include the characterization.
- Implement the extended algorithm.
- Perform further testing.
- Ensure that data is formatted appropriately for the visualization.

4.2 Visualization

Quantifying the effectiveness of visualizations is a difficult problem [10]. Kosara et al. [13] suggest the appliance of user studies, and Johnson [10] adds to it the scientific method of observation, hypothesis formulation and evaluation to determine efficacy. Quantitative methodologies will be used, as empirical measurements are needed to determine the accuracy of user estimations and performance on visual tasks.

Visualizations will be developed in two distinct phases. The initial design phase will be broken into two iterations, involving rapid prototyping methods such as throw away and paper prototyping. This phase will be explorative and aim to find a visualization among the prototypes most suitable for RFI detection. For each prototype, a list of hypotheses will be formulated beforehand. These hypotheses describe what we expect the visualization to accomplish, for example improving comparisons over multiple data sets. A hypothesis will be confirmed or rejected on the basis of user studies and visualization testing (see Evaluation). The results of the experiments will decide whether the prototype should be discarded and redesigned, or be kept and improved.

After the design phase, feedback gathered from the experiments will be analysed to determine which visual encodings and visualization techniques were the most effective. Thereafter, the implementation phase will begin. The results of the analysis will be used to implement a final visualization.

5 Evaluation

5.1 Detection

The implementation of the detection algorithm can be tested in two ways. The first is to use ‘salted’ data, which is test data with a known RFI signal artificially added. If the system can detect the artificial RFI then it has passed the first test. The second test is to use data where the RFI environment is partially known. If the algorithm detects the RFI corresponding to the known environment, then it is successful at the most basic level. If it is able to detect more RFI than was previously known, which can be proven to be RFI then the algorithm is fully successful.

5.2 Visualization

For user testing, the method of Cleveland and McGill [6] will be adapted, and 15 to 20 human subjects will be used to conduct perceptual experiments. Ideally, people who have some familiarity with design principles will be recruited. Participants might prefer one visualization over the other because it is more aesthetically pleasing (Ellis and Dix [8] suggest similar guidelines). This may reduce the number of potential test candidates, but will increase the validity of the feedback.

For each user experiment, a list of perceptual tasks that require the extraction of quantitative information from a visualization will be constructed. Two visualizations will be used for comparison, where visualizations differ from another by a controlled parameter. The performance of users on tasks for different visualizations are measured during the experiment, and tasks are ordered according to user performance for each visualization. This information can be used to design new visualizations, or improve existing ones by altering visual encodings.

Perceptual tasks include estimation and discrimination tasks, or more abstract tasks such as spotting trends. The time users take and the accuracy of estimations will be used as performance metrics. The estimation error can be accurately calculated by comparing a subject's answer with a computed answer. In order to prevent the effects of outliers, the results can be trimmed using 80% means, similar to Heer et al. [9], where data is trimmed per subject. The learning effect could be reduced by allowing subjects an initial practice session.

6 Ethical, Professional and Legal Issues

The intellectual property rights to this project are held by the University of Cape Town. We add no further restrictions to this work and prefer to release it under a permissive free software license, such as the MIT License. The visualization experiments require the use of human subjects. Ethical clearance will need to be obtained before conducting user testing. The RFI detection algorithms can be evaluated entirely through the use of performance metrics and test data. As such, there is no need for ethical clearance for this. The data used for the visualizations and RFI detection algorithms is supplied by Christopher Schollar. Restrictions to its usage might apply and confirming the permitted operations on the data will be done before prototyping and beginning the implementation.

7 Anticipated Outcomes

Overall the project should produce a system which integrates detection and visualisation of RFI in the environment surrounding the reference antenna. This would mean that the detection algorithm feeds into the visualisation.

7.1 Detection

An implementation of at least one form of detection algorithm will have been produced and tested, which can then be used to show detected RFI on the visualisation. The algorithm should be computationally efficient. It is expected that RFI will be detected more effectively, with fewer false positives and false negatives. It is hoped that it will be possible to add characterization to the algorithm, however, it may not be possible within the time constraints.

7.2 Visualisation

The final implementation is expected to have the following properties in comparison to the prototypes:

- An increase in the proportion of the total size of the graph that is dedicated to displaying data. That is, screen real estate will be utilised more effectively.
- An improved representation of the raw data. Data reduction techniques should preserve interesting features such as trends and outliers.
- A better use of visual encodings and design principles, such as small multiples, to ease the comprehension of data sets. Participants should take less time and make better estimations when performing visual tasks.

8 Project Plan

8.1 Risks

8.1.1 A group member leaves

Risk: Low
Impact: Low
The two parts of the project can be completed entirely independently, with the integration between the two considered a “nice to have”.

8.1.2 Data becomes unavailable

Risk: Low
Impact: High
If the SKA decide that we cannot use the real data we can try to create data, but that would be difficult to do on the scale that the real data is sampled, also interesting features are unlikely to be generated. No real testing can be done without the data.

8.1.3 Output from detection cannot be integrated into visualisation

Risk: Medium
Impact: Medium
The integration of the detection into the visualisation is a non-essential part of the project. Both parts can have successful results without combining.

8.1.4 Loss of work

Risk: Medium
Impact: High
By ensuring that work is backed up at regular intervals to the cloud and to various devices it should be possible to minimise this risk, however, if work is lost it can cause serious delays in the project as it will have to be redone. There are a number of things that could cause loss of work, ranging from physical loss of equipment (e.g. memory sticks) to accidental deletion.

8.1.5 A lack of participants for the visualization experiments

Risk: High
Impact: High
The experiments require a fair number of human subjects for sufficient feedback. Cleveland

and McGill [6] obtained valid results from 51 users, which will be a difficult amount to recruit. They also had problems with 7 subjects not following instructions. Enticements such as rewards could be used to attract subjects; overbooking could be used to compensate for potentially invalid results.

8.1.6 Leaking personal information

Risk: Low
Impact: High
Data from user experiments will have to be anonymised or encrypted before uploading to third party servers. There is a risk of leaking personal information, which could have consequences under the Protection of Personal Information Act.

8.2 Resources Required

The primary resource for both the visualizations and the RFI detection algorithms is the data sets provided by Christopher Schollar. Commodity hardware is sufficient for completing this project. The visualization will make use of third party libraries. Examples include Data Driven Documents (the successor of Protovis) and the JavaScript implementation of Processing, which are used by Heer et al. [9]. The data is stored in HD5 format, and so third party libraries are required to read the files. One such is h5py, used through numpy in Python which is freely available. These will be used for the detection algorithms.

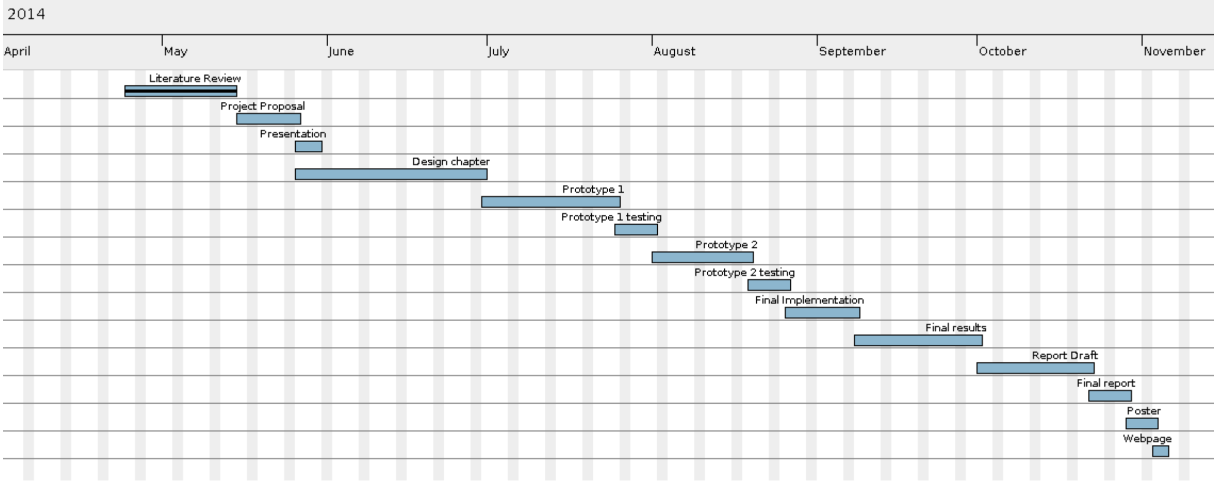
8.3 Deliverables

Section	Due date
Project Proposal	26 - 05 - 2014
Proposal Presentation	30 - 05 - 2014
Project Website	17 - 06 - 2014
Prototype	17 - 08 - 2014
Final implementation	17 - 09 - 2014
Results	10 - 10 - 2014
Draft of report	22 - 10 - 2014
Final report	29 - 10 - 2014
Poster due	03 - 11 - 2014
Final webpage	05 - 11 - 2014
Individual reflection	07 - 11 - 2014
Final presentation	14 - 11 - 2014

8.4 Milestones

What	When
Presentation	30 - 05 - 2014
Final Proposal	02 - 06 - 2014
Design chapter complete	30 - 06 -2014
Prototype 1 complete Pippa – Algorithm developed Gerard – Two visualization prototypes developed	25 - 07 -2014
Prototype 1 tested	01 - 08 - 2014
Prototype 2 complete Pippa – Initial implementation Gerard –Redesign or improve visualizations based on test results	19 - 08 - 2014
Prototype 2 tested	26 - 08 - 2014
Final implementation Pippa – optimisation Gerard –Use test results to develop the final visualizations	08 - 09 - 2014
Final results	01 - 10 - 2014
Draft of report	22 - 10 - 2014
Final report	29 - 10 - 2014
Project demonstration	03 - 11 - 2014
Poster due	03 - 11 - 2014
Final webpage	05 - 11 - 2014
Individual reflection	07 - 11 - 2014
Final presentation	14 - 11 - 2014

8.5 Timeline



8.6 Work Division

Detection algorithms will be done by Philippa Hillebrand.
The work on alternative visualizations will be done by Gerard Nothnagel.

References

[1] J. Antoni, "The spectral kurtosis: a useful tool for characterising non-stationary signals," *Mechanical Systems and Signal Processing*, vol. 20, pp. 282–307, 2004.

[2] E. Bertini and G. Santucci, "Give chance a chance: modeling density to enhance scatter plot quality through random data sampling," *Information Visualization*, vol. 5, no. 2, pp. 95–110, 2006.

[3] P. Bolli, F. Gaudiomonte, F. Messina, R. Ambrosini, C. Bortolotti, and M. Roma, "The RFI monitoring systems for the Medicina and the Sardinia radio telescopes," in *PoS RFI2010*, 2010, p. 29.

[4] A. Boonstra and S. Van der Tol, "Spatial filtering of interfering signals at the initial low frequency array (lofar) phased array test station," *Radio science*, vol. 40, no. 5, 2005.

[5] J. Choo and H. Park, "Customizing computational methods for visual analytics with big data," *IEEE Computer Graphics and Applications*, vol. 33, no. 4, pp. 22–28, 2013.

[6] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American Statistical Association*, vol. 79, no. 387, pp. 531–554, 1984.

[7] R. Ekers and J. Bell, "Radio frequency interference," *arXiv preprint astro-ph/0002515*, 2000.

[8] G. Ellis and A. Dix, "An explorative analysis of user evaluation studies in information visualisation." in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, 2006.

[9] J. Heer, M. Bostock, and V. Ogievetsky, "A tour through the visualization zoo," *Commun. ACM*, vol. 53, no. 6, pp. 59–67, 2010.

[10] C. Johnson, "Top scientific visualization research problems," *IEEE Computer graphics and applications*, vol. 24, no. 4, pp. 13–17, 2004.

[11] D. A. Keim, M. C. Hao, U. Dayal, H. Janetzko, and P. Bak, "Generalized scatter plots," *Information Visualization*, vol. 9, no. 4, pp. 301–311, 2010.

[12] J. Kocz, F. Briggs, and J. Reynolds, "Radio frequency interference removal through the application of spatial filtering techniques on the parkes multibeam receiver," *The Astronomical Journal*, vol. 140, no. 6, pp. 2086–2094, 2010.

[13] R. Kosara, C. Healey, V. Interrante, D. Laidlaw, and C. Ware, "Thoughts on user studies: Why, how, and when," *IEEE Computer graphics and applications*, vol. 23, no. 4, pp. 20–25, 2003.

[14] K. J. Lee, K. Stovall *et al.*, "PEACE: Pulsar Evaluation Algorithm for Candidate Extraction – a software package for post-analysis processing of pulsar survey candidates," *Monthly Notices of the Royal Astronomical Society*, vol. 433, no. 1, pp. 688–694, 2013. [Online]. Available: http://mnras.oxfordjournals.org/content/433/1/688.abstract

[15] Z. Liu, B. Jiang, and J. Heer, "imMens: Real time visual querying of big data," *Computer Graphics Forum*, vol. 32, no. 3, pp. 421–430, June 2013.

[16] R. Lord, "Radio frequency interference and radio astronomy : why the fuss?" *Quest*, vol. 8, no. 3, pp. 18–20, 2012.

[17] G. M. Nita and D. E. Gary, "Statistics of the SK estimator," in *PoS RFI2010*, 2010, p. 19.

[18] A. R. Offringa, A. G. de Bruyn, S. Zaroubi, and M. Biehl, "A LOFAR detection pipeline and its first results," in *PoS RFI2010*, 2010, p. 36.

[19] A. Offringa, A. de Bruyn, M. Biehl, S. Zaroubi, G. Bernardi, and V. Pandey, "Post-correlation radio frequency interference classification methods," *Monthly Notices of the Royal Astronomical Society*, vol. 405, no. 1, pp. 155–167, 2010.

[20] A. Offringa, J. Van de Gronde, and J. Roerdink, "A morphological algorithm for improving radio-frequency interference detection." *Astronomy & Astrophysics/Astronomie et Astrophysique*, vol. 539, p. A95, 2012.

[21] R. Oliva, E. Daganzo, Y. H. Kerr, S. Mecklenburg, S. Nieto, P. Richaume, and C. Gruhier, "Smos radio frequency interference scenario: Status and actions taken to improve the rfi environment in the 1400–1427-mhz passive band," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 50, no. 5, pp. 1427–1439, 2012.

[22] M. Trutschl, G. Grinstein, and U. Cvek, "Intelligently resolving point occlusion," in *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, October 2003, pp. 131–136.

[23] P. C. Wong, H. W. Shen, C. R. Johnson, C. Chen, and R. B. Ross, "Customizing computational methods for visual analytics with big data," *IEEE Computer Graphics and Applications*, vol. 32, no. 4, p. 63, 2012.