2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics
International Conference (PRASA-RobMech)
Stellenbosch, South Africa

Statistical classification of radio frequency interference (RFI) in a radio astronomy environment

Cornelis Johannes Wolfaardt, David Davidson, Thomas Niesler
Department of Electrical and Electronic Engineering, Stellenbosch University
kolijn.wolfaardt@gmail.com
davidson@sun.ac.za
trn@sun.ac.za

***Abstract*—We present the application of statistical classifiers to the problem of automatic identification of radio frequency interference (RFI) in radio astronomy. RFI can corrupt measurements made by radio telescopes and it is therefore very important that such interference can be identified. We compile a dataset of RFI signals gathered at the SKA site near Carnavon, South Africa, and use this data to train and evaluate some statistical classifiers. We find the best performing system to use the k-nearest-neighbour (knn) classifier and achieve an accuracy of 93%. Since our dataset was limited by the capturing equipment in terms of record length, we feel that there is scope to improve on this figure in the future.**

I. Introduction

In radio astronomy radio frequency interference (RFI) can corrupt measurements and thereby render experimental results incorrect. In some cases the RFI saturates the receiver, rendering the measurements worthless. In other cases it may be difficult or even impossible to identify and remove the interference from data gathered by a radio telescope. The RFI signals may originate from external sources, or from the electronics in the radio telescope itself. This paper focuses on the automatic identification of signals generated by external sources.

II. Existing Mitigation Methods

A common method of preventing RFI is to simply keep the area surrounding a radio telescope free of possible sources. This is not always possible, for example when the telescope is situated in a built-up area. Furthermore, new RFI sources may unwittingly be introduced, for example by visitors bearing equipment whose RFI emissions are not well characterised. To ensure the integrity of data gathered from the telescopes, continuous active monitoring of RFI emissions is important.

A. RFI Mitigation Using Additional Antennas

Various approaches to the removal of RFI by means of secondary antennas have been considered. The secondary antennas are low-gain antennas, and are sensitive only to the interference signal. Correlated components between the secondary signal and the primary signal can be removed from the latter. However, because the primary antenna can usually rotate, it is susceptible to differing amounts of RFI[1]. The secondary antenna also illuminates a much larger part of the sky and in order to intercept the interference signal it may include the surrounding horizons. This can introduce environmental thermal noise into the signal.

In [2] a method of RFI mitigation is investigated using a digital adaptive filter. An algorithm continually adjusts the filter in such a way that the output interference power is minimized. In [3] a phased array is used to detect and record interfering signals. A phased array is used for better control of the receiving antenna pattern. The antenna used in these experiments is a six element hexagonal array

B. Thresholding Based Methods

Another common method of mitigating RFI is to flag data as containing RFI when the power of the received signal exceeds a certain threshold. The threshold can also be set globally or varied according to signal properties. In the cumulative sum (CUSUM)[4] method, small frames of samples are summed and an average calculated. If this average exceeds the threshold all the samples fully within the considered frames are flagged.

Combinatory thresholding extends the CUSUM method. Here the frame lengths and the threshold for each frame are varied. The average for small frames must exceed a large threshold, while the average for a larger frame has a lower threshold.

C. Statistical Methods

In [5] a method of removing RFI using surface fitting and smoothing is proposed. A function is fitted to the correlated visibilities. The assumption is made that the combination of the astronomical signal to the image is smooth, while the RFI introduces more rapid changes. This method is not suitable for the detection of pulsars or other narrowband sources.

D. Post-Flagging Techniques

Once data has been flagged in the the frequency domain further processing can be performed to improve the accuracy of the flagging. Analysis of the RFI signal properties can also be performed.

In [6] the statistics of RFI events are investigated. Data from the Parkes Multi beam Pulsar Survey is applied to a thresholding algorithm to flag RFI events. The frequency band, angle of arrival as well as the time of day is used to analyse the statistical distribution of RFI.

978-1-5090-3335-5/16/$31.00 ©2016 IEEE

E. Morphological Algorithms

An algorithm based on the mathematical principle known as dilation was proposed in [7]. The antenna data is first processed by one of the other mitigation techniques, such as thresholding. This will produce an array of flags for the data, which is then processed by the morphological algorithm. The morphological algorithm flags additional samples around already flagged data, based on various criteria. For example, the morphological algorithm assumes that the samples surrounding flagged samples are likely to also contain RFI, but at lower power. These samples are not detected by previous algorithms, but can still interfere with the astronomical observations. The algorithm therefore flags such additional samples based on the number of samples originally flagged. The algorithm processes one dimension at a time, but can be applied to any number of dimensions successively. The order in which the dimensions are processed is important.

As the above brief literature review shows, the application of statistical classifiers to the identification of RFI has not received much attention. We attempt to address this with a first set of experiments.

III. Data Collection

RFI data from various sources is required for analysis and classification, in the form of time-domain signals, containing components from the offending source. Many different captures are required for statistical classification in order to build a statistical model of the signal.

Ideally the data should be captured in an RFI silent environment, to ensure that no other signals are present and to minimize the environment noise present. This type of RFI isolation can be provided by an anechoic chamber. However, we were uncertain whether the signal captured in an anechoic chamber would properly resemble the real world signal. Furthermore some sources were too big or even immobile and could therefore not be characterized in an anechoic chamber.

Therefore data was captured from various sources using the Real Time Analyser and an log periodic dipole antenna (LPDA) during a visit to the SKA site. These were provided by the SKA office in Cape Town. Data was captured in the time domain. For some of the captures, the on-site RFI trailer was used [8]. The trailer uses a Rhode-Schwarz HL033 LPDA antenna attached to a mast.

a) Real Time Analyser: The Real Time Analyser (RTA) is capable of high-speed data capturing in both the time and frequency domain. The RTA samples at 1.8GSa/s, and therefore a wide band. However at the time of writing the RTA could only capture 8 microseconds of the signal continuously. The RTA can capture from any one of four different frequency band as shown in Table I. We attempted data capture from all four bands, but ultimately focused mainly on the lowest band since little signal activity was seen in the higher bands. For our analysis the bands are treated separately because they were not sampled simultaneously.

TABLE I
TABLE OF RTA FREQUENCY BANDS

Band	Frequency
1	50 - 850 MHz
2	800 - 1050 MHz
3	1050 - 1670 MHz
4	1950 - 2550 MHz

The RTA has configurable gain and attenuator sections in the signal chain. These are adjusted on a source by source basis, and for every band used.

In frequency domain capture mode, the RTA accumulates the spectrum of the signal over a configurable duration, usually between 1 and 10 seconds. In this mode the frequency band is divided into 32678 channels by an internal polyphase filter bank. This means that the frequency capture mode is not very suitable for capturing transients. It can however be used to detect low-powered stationary RFI signals [9].

b) LPDA Antenna: The LPDA antenna was chosen because it operates over a wide frequency band and is directional.

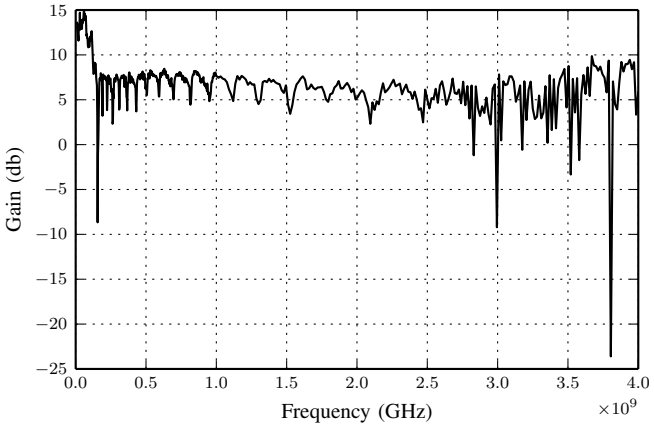

Fig. 1. LPDA gain over frequency

Figure 1 shows the gain of the LPDA antenna over a frequency range. The graph shows a relatively consistent gain over the lower gigahertz range, from 200MHz to 4 GHz.

c) Sources: Many different RFI sources were considered. They were selected based on the availability of the source. The same sources were captured multiple times, in multiple different frequency bands. Some of the sources proved very difficult to capture due to their transient nature and the short ($8\mu S$) capture window of the RTA.

Table II shows the raw number of samples available, before any processing was applied. Some of the sources have captures in all 4 bands, but most only have in the first band. During data capturing an approaching lightning storm was noticed. Captures made during that time were labelled as such and were not used.

TABLE II
NUMBER OF FRAMES OBTAINED FOR EACH RFI SOURCE.

Source Name	Band 1	Band 2	Band 3	Band 4
Bakkie lights	30	0	0	0
Bakkie radio	7	0	0	0
Bakkie start	31	7	0	0
Big crane	15	0	44	31
Big radio	26	0	0	0
Cellphone	7	59	0	0
Cherry picker	10	0	0	0
Compressor	9	0	0	0
Diesel filter	114	16	14	13
Kat7 meysdam	41	0	0	0
Lightning discard	28	0	0	0
Meerkat compressor	209	0	0	0
Meysdam gap	252	0	0	0
Possible lightning	9	0	0	0
Radio	22	0	0	0
Refrigerator unit	13	5	0	0
VW ignition	24	0	0	0
VW indicators	28	0	0	0
Welder	17	0	0	0
Welder_spark	12	0	0	0
Total	904	87	58	4

IV. DATA PROCESSING

The data provided by the RTA consist of sets of 32768 consecutive time samples, which in the following we will refer to as *captures*. After ensuring zero mean, each capture was divided into frames of 1024 samples, overlapping by 512 samples. Each frame was then further divided into four segments of 128 samples. The magnitude of the Fast Fourier Transform of each segment was calculated after applying a Hamming window. The four magnitude spectra were averaged to represent the frequency content of the frame. This process produced a spectrogram of the data.

a) Outlier Removal: The spectrograms were used to visually inspect the captured data. Any obviously corrupt (outlier) captures were removed. These included captures containing any other interference signals such as radio signals. Such interference was identified by comparing all the captures for a specific source, and isolating obvious deviations. There are automated methods of detecting outliers as well.

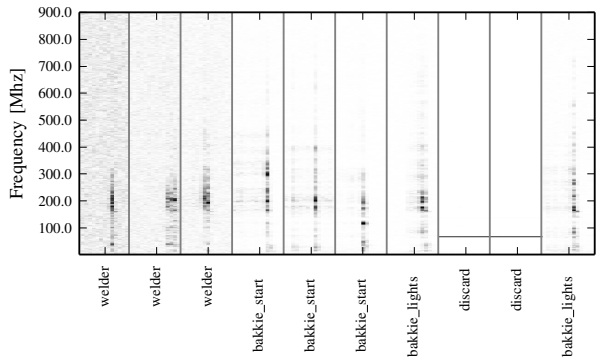

Fig. 2. Example spectrogram of several different sources.

Figure 2 shows several examples of the spectrograms from different sources. Differences in frequency content between the various classes can be seen.

b) Labelling: The data was assigned labels based on notes taken while capturing. Two different methods of labelling were used: per capture labelling, and per frame labelling. Per capture labelling assigned a label to the entire capture, even though some sections might be silent. The frame of the spectrogram containing the greatest power was used as the feature vector.

Per frame labelling assigned a label to each frame, using the same frames mentioned above. If the power in a frame exceeded a threshold, it was labelled as the RFI signal. Otherwise it was labelled as silence. The threshold was manually adjusted on a file by file basis, but was always kept between 5% and 15% of the total power of the capture. This method made much more data available, as each frame is introduced to the classifier. The silence classes were averaged together, under the assumption that they are similar.

Per-frame labelling was motivated by an inspection of the data which revealed that, due to the impulsive and non-stationary nature of many of the interference sources, most captures included a substantial amount of silence, during which no interference was present.

V. EXPERIMENTAL RESULTS

10-Fold cross validation was used in all experiments to make good use of the limited dataset. The data was divided into 10 equal subsets, called folds. Six of the folds were used for training, two for tuning and the remaining two for testing. An experimental result was obtained for this fold. The different subsets of data used for each set were then varied for each of the permutations.

For each of these classifiers, parameter optimization was performed on the tuning folds. These parameters are calculated for each data fold, and the optimal value was then selected.

Two different classifiers were investigated; a KNN classifier and a GMM classifier. A KNN classifier predicts the class of a new data point by selecting the most prevalent class among the k closest neighbours in the training data set. The only parameter to be tuned is the number of nearest neighbours (k). This value is varied over a small range of values, and the optimal value based on the tuning set is selected.

A GMM classifier creates a generative model for the data by fitting one or more Gaussian distributions over the data of a single class. This can be used to calculate the probability that a new data point belongs to a certain class. When repeated for all the classes, the class with the highest probability is selected. The Gaussian distributions are fitted using the expectation maximization algorithm.

The possible parameters for the GMM classifier are the number of distributions, as well as the constraint on the covariance matrix of each distribution. The covariance matrix can be constrained to be a single value, a diagonal matrix or a full matrix.

A. Classification using Per-capture Labels

Classification using the per-capture labelled data was performed first. For the KNN classifier a value of $K = 1$ was found to be optimal. The average classification accuracy was found to be 70.80%, with a standard deviation of 30.72%.

For the GMM, a full covariance matrix using 3 Gaussians was found to be optimal. The GMM achieves an average classification accuracy of 65.56%, with a standard deviation of 31.01%. Some classes are almost unused by the classifier. These results are reflected in the first line of III.

B. Classification using Per-frame Labels

Next, classification was performed using the per-frame labelled data. Again, the KNN and GMM parameters were optimized by using a tuning data set within a 10-fold cross validation framework. Figure 3 shows the classification accuracies for different values of k on the tuning set. Figure 4 shows corresponding results for different GMM parameters.

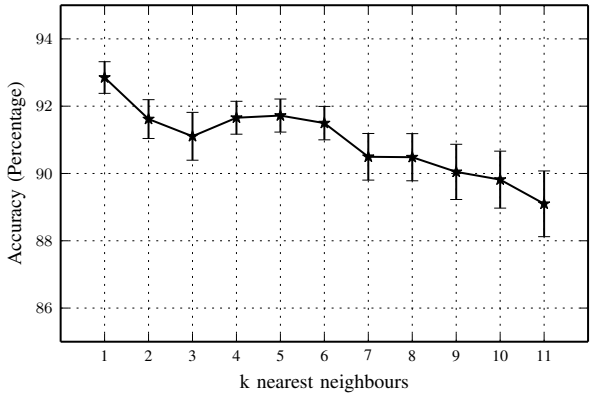

Fig. 3. Comparison of KNN classifier accuracies for various values of k when using frame-based labels.

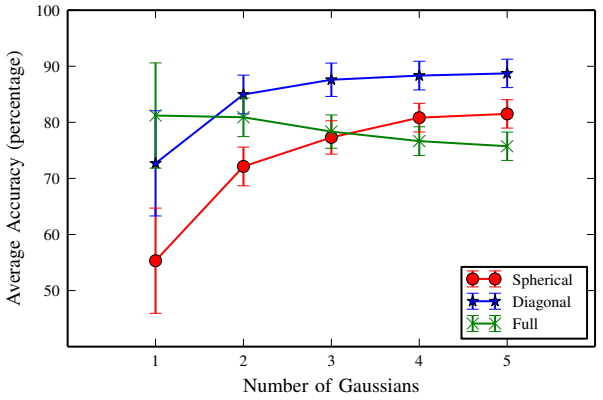

Fig. 4. Comparison of GMM classifier accuracies for different number of Gaussians and covariance matrix, when using per-frame labelling.

For the KNN classifier, 1 nearest neighbour was again found to be optimal, but in this case an accuracy of 93.20% and standard deviation of 6.46% was achieved. The optimal GMM classifier used 3 Gaussians per mixture and a diagonal covariance matrix to achieve an average accuracy of 87.34% and standard deviation of 10.76%. These results are reflected in the second line of Table III.

The per-frame labels achieved a substantially better accuracy for the KNN and GMM classifiers. The KNN achieved a better accuracy than the GMM overall.

C. Classification of Data from Higher Frequency Bands

Similar experiments were performed using the data from the higher frequency bands. However, much less data was available in these bands, because many sources proved difficult to capture or did not emit any detectable signals at all in these bands. Both the per-capture and per-frame labels were used to train both the KNN and GMM classifiers. Again the tuning set was used to optimize the parameters. The optimal KNN classifier again used $k = 1$ while the GMM classifier used 2 Gaussians and a diagonal covariance matrix. The KNN classifier achieved an accuracy of 89.54%, with a standard deviation of 6.58%. The GMM had an accuracy of 91.38% and standard deviation of 4.84%.

VI. Further Feature Extraction Experiments

Two variations of the feature extraction methods were also investigated.

A. Classification using a Reduced Feature Vector

Rather than dividing each frame into segments of 128 samples, they were divided into segments containing just 32 samples. The average magnitude spectrogram was then calculated using these shorter segments, and the DC component removed. This reduces the feature vector dimension to 15 and was expected to make classification much quicker, but also to reduce the accuracy of classification. The experiments of the previous section were repeated for this configuration. Other data reduction methods such as Principle Component Analysis are also available, but were not used.

B. Augmentation with Delta Frames

The spectral feature vector used originally can be extended by appending the finite difference to the next frame (delta). This results in a doubling of the feature vector dimensionality.

Neither reduced feature dimensionality nor the augmentation with delta frame led to improved performance. The results for both are shown alongside the other results in Table III.

VII. Discussion and Conclusion

Overall the work showed that it is possible to classify RFI using machine learning techniques. Both of the classifiers that were investigated were able to classify the data with high accuracies.

Overall the per-frame based labels performed better than the per-capture labels. The per-frame labels divided the signal into finer segments and used all of it to classify, rather than choosing the segment with the most power. The increased detail helped to achieve a better classification result.

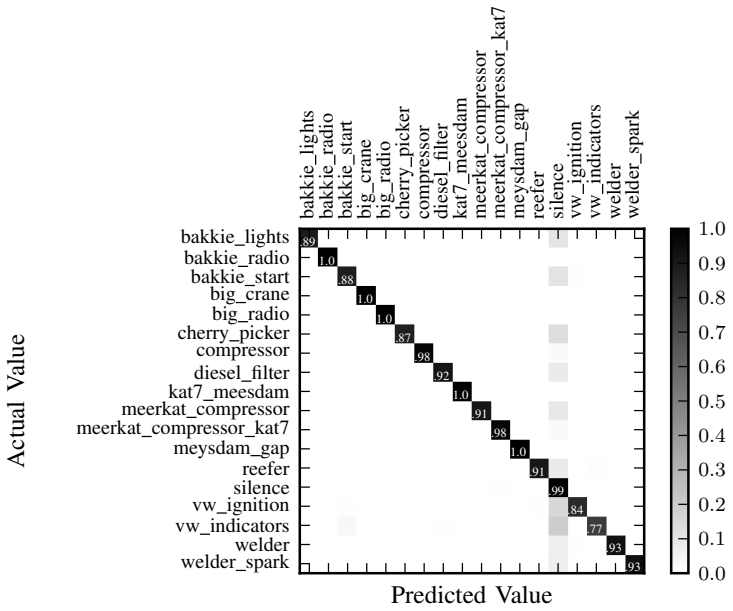

Fig. 5. Confusion matrix of the KNN classifier using frame based labels.

Table III summarizes all the results. Overall the KNN classifier using frame based labels scored the best. In future work other classifiers might be considered, especially if more data is available. This will however require more and specifically longer captured data than is currently available.

Figure 5 shows a confusion matrix for the KNN classifier using frame based labels. The visible diagonal indicates that most classifications are performed correctly. Almost all classes had a few frames misclassified as silence. This seems to suggest that the silence frames identified by the thresholding method still contain some features that can be identified. There is room for a better labelling technique, which extracts more labelled frames from the total dataset.

This also introduces the possibility of a classifier that attempts to distinguish between RFI and silence signals. Another classifier can then distinguish the RFI signals from each other.

TABLE III
CLASSIFICATION ACCURACIES FOR ALL CLASSIFIERS

Feature Type	**KNN Classifier**		**GMM Classifier**	
	Accuracy	Std. Dev.	Accuracy	Std. Dev.
Capture based labels	70.80%	30.72	65.56%	31.01
Frame based labels	93.20%	6.46	87.34%	10.76
Reduced feature vector	81.70%	16.36	78.35%	16.69
Delta frames	86.05%	13.23	78.34%	18.28

For future work longer captures and more captures of a source in different configurations is desirable. Longer captures will enable a better model to be created of the transient properties of the RFI. The currently presented analysis assumes the RFI characteristics of a given source remain constant in time. Longer captures might enable more insight into this aspect of the RFI emissions, and allow for time-varying models.

Data from the source in different configurations will allow more extensive testing of the classifier. Capturing the source from further away, while difficult, will provide an excellent test for the classifier.

VIII. ACKNOWLEDGEMENT

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

REFERENCES

[1] W. A. Baan, "Rfi mitigation in radio astronomy," in *General Assembly and Scientific Symposium, 2011 XXXth URSI*. IEEE, 2011, pp. 1–2.

[2] C. Barnbaum and R. F. Bradley, "A new approach to interference excision in radio astronomy: Real-time adaptive cancellation," *The Astronomical Journal*, vol. 116, no. 5, p. 2598, 1998.

[3] C. K. Hansen, K. F. Warnick, B. D. Jeffs, J. R. Fisher, and R. Bradley, "Interference mitigation using a focal plane array," *Radio Science*, vol. 40, no. 5, 2005.

[4] P. Friedman, "A change point detection method for elimination of industrial interference in radio astronomy receivers," in *Statistical Signal and Array Processing, 1996. Proceedings., 8th IEEE Signal Processing Workshop on (Cat. No. 96TB10004*. IEEE, 1996, pp. 264–266.

[5] 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.

[6] G. Doran, "Characterizing interference in radio astronomy observations through active and unsupervised learning," *JPL Publication 13-12*, 2012.

[7] A. Offringa, J. van de Gronde, and J. Roerdink, "A morphological algorithm for improving radio-frequency interference detection," *arXiv preprint arXiv:1201.3364*, 2012.

[8] Unknown. (2015, May) Rhode and Schwarz HL033 antenna. http://www.rohde-schwarz.com.

[9] A. Botha, "Development of a real-time transient analyser for the SKA," Master's thesis, Stellenbosch: Stellenbosch University, 2014.