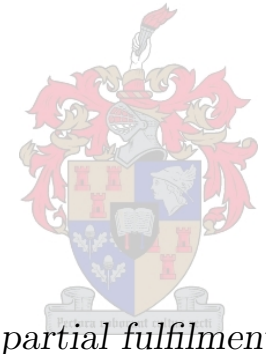
Machine learning approach to radio frequency interference(RFI) classification in radio astronomy

by

Cornelis Johannes Wolfaardt



Thesis presented in partial fulfilment of the requirements for the degree of Master in Electronic Engineering in the Faculty of Engineering at Stellenbosch University

Supervisor: Prof. TR. Niesler

Co-supervisor: Prof. D. B. Davidson

March 2016

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.

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: March 2016 .....

Abstract

Machine learning approach to radio frequency interference(RFI) classification in radio astronomy

CJ. Wolfaardt

Department of Electronic Engineering,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Thesis: MEng (Elec)

November 2015

Radio frequency interference (RFI) presents a large problem for radio telescopes. Interference prevents observations from being made, or extends the duration required for observations. This thesis investigates different methods to automatically classify RFI signals. Data from different sources was captured at the SKA site. Both Gaussian Mixture Model (GMM) and K-nearest neighbors (KNN) classifiers were used to analyse the data. Both performed adequately, with the KNN slightly outperforming the GMM. Different feature extraction methods were also investigated.

Uittreksel

Masjienleer klassifikasie van steurseine in radio astronomie

("Machine learning approach to radio frequency interference(RFI) classification in radio astronomy")

CJ. Wolfaardt

Departement Elektroniese Ingenieurswese,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Tesis: MIng (Elek)

November 2015

Radio frekwensie steurseine verteenwoordig 'n groot probleem vir radio teleskope. Steurseine verhoed teleskope om waarnemings te maak. Hierdie tesis ondersoek verskeie metodes om steurseine automaties te identifiseer en klasifiseer. Data van bekende steurseine op die SKA terrein is versamel. Verkeie voorverwerkingtegnieke word ondersoek en dan geannaliseer met bekende statistiese modelle soos 'n GMM en KNN. Beide lewer aanvaarbare resultate. Verskeie metodes om kenmerke te onttrek word ook ondersoek.

Acknowledgements

I would like to extend my sincere gratitude and thanks to the following persons and organizations:

- My supervisor Prof. Thomas Niesler for always providing guidance and constructive feedback.
- Simon Norval, Siyabulela Tshongweni and Christopher Schollar, as well as the other staff at the SKA office Cape Town for helping to obtain the data used.
- My labmates and friends, for stimulating discussions and positive motivation.
- My family, for your support during the project.
- Thanks to the NRF for providing funding for the project.

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

ADC Analog to digital convertor

ISM Interstellar medium

LNA Low-noise amplifier

RAM Random access memory

RF Radio frequency

RFI Radio frequency interference

ROACH Reconfigurable open architecture computing hardware

RTA Real time analyser

Radio Telescopes

ALMA Atacama Large Millimeter Array

KAT Karoo Array Telescope

LOFAR Low-frequency Array

SKA Square Kilometer Array

VLBA Very Long Baseline Array

Machine Learning

EM Expectation maximization

DET Detection error tradeoff

GMM Gaussian mixture model

HMM Hidden Markov model

KNN K-nearest neighbours

LDA Linear discriminant analysis

PCA Principle component analysis

PDF Probability density function

ROC Receiver operating characteristic

SVD Singular value decomposition

Chapter 1

Background to Radio Astronomy

Radio astronomy involves the study of radio waves emitted by celestial objects. The field of radio astronomy emerged in the 1930s when Karl Jansky noticed a periodic interference signal in his radio communications measurements. The signal had a period of 23 hours and 56 minutes, one sidereal day. The peak of the signal coincided with when the Milky Way was overhead. This led Jansky to conclude that the interfering signal was originating from space. [2]

The first radio telescope was built by Grote Reber, who was inspired by Jansky's work. Reber experimented with several frequencies, finally settling on 160MHz [3]. Reber also completed the first sky survey, which was published in 1941. A sky survey is a map of the sky or part of the sky showing the intensity for a specific frequency.

Reber's work caused a large increase in interest in radio astronomy. In order to increase the resolution of the observations, the size of radio telescope dishes kept increasing. The increased size of the dishes also meant that the complexity of the supporting structure increased.

In the 1940s the technique of radio interferometry was developed. This technique uses observations from multiple antennas to improve the resolution with which observations can be made. This led to the development of radio interferometer arrays such as the Very Large Array and the Atacama Large Millimetre Array. The Square Kilometre Array is also an interferometric array.

There are many types of observations that are well suited to radio astronomy. There are many different sources that emit in the radio frequency

spectrum, which can be received with little interference. These signals are however subject to redshift.

1.1 Redshift

Due to the expansion of the universe all of the astronomical signal sources are moving away from the earth. The Doppler effect causes any signal transmitted between objects moving at different speeds to undergo a change in frequency. This effect lowers the frequency of signals received from astronomical sources. It is named after the similar effect which occurs in optical observations, where it moves signals towards the red (lower frequency) part of the spectrum.

Redshift is expressed as the fractional change of the wavelength and is represented as the dimensionless quantity z. λ_e is the wavelength of the emitted signal, and λ_o is the observed wavelength.

$$z = \frac{\lambda_o - \lambda_e}{\lambda_e} \tag{1.1}$$

A redshift of 0 represents no change in the wavelength. A redshift with z>0 implies an increase in wavelength, while a redshift with z<0 implies a decrease in wavelength (blueshift). Astronomical objects with a high redshift (z>0) are more distant than lower redshift objects. This is due to the accelerating expansion of the universe. Older objects are further away, and are accelerating away from us. Redshift is an important factor to take into account, because it can drastically affect the frequency of a received signal.

By measuring the magnitude of the redshift, the speed of the source can be determined. This measurement can be used to determine the distance to the source [4].

1.2 Atmospheric Absorption

Radio waves between 100MHz and 10GHz are not absorbed appreciably by our atmosphere, so we can easily study them from the ground. The atmosphere does however influence the signals, which must be corrected for. This is in contrast to optical methods, where signals are severely distorted and absorbed by the atmosphere, caused by variations in temperature and pressure.

Radio frequency signals are absorbed by water vapour in the atmosphere. The signals are also affected by refraction in the atmosphere, due to differing temperature layers.

1.3 Origin of signals

There are many different sources of electromagnetic radiation in the universe. The type of signal and the source has an influence on the technique used to study them.

1.4 Basic Black-body Radiation

All objects with a temperature above 0°K emit electromagnetic radiation. The intensity of the emitted waves is determined by the temperature of the object. The intensity at a specific frequency and a specific temperature can be calculated using Planck's radiation law. This law determines the brightness B of a black-body radiator, given its temperature T, and the frequency of interest, ν.

$$B_\nu(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{(kT)}} - 1} \tag{1.2}$$

Where

B	Brightness in $W \cdot sr^{-1} \cdot m^{-2} \cdot Hz^{-}1$
h	Planck's Constant (6.63×10^{-34}joule per second)
ν	Frequency in Hertz
c	Speed of light $3 \times 10^8 \frac{m}{s}$
k	Boltzmann's Constant
T	Temperature in Kelvin

This equation can be used to plot frequency vs brightness at a certain temperature, to show which frequencies an object at that temperature radiates most. The frequency of the radiated waves are heavily dependent on the temperature of the object.

Figure 1.1 shows the radiation intensity at different wavelengths for three different temperatures. The peak of the intensity lies mostly in or close to the visible area. Radio waves have a much longer wavelength, lying between

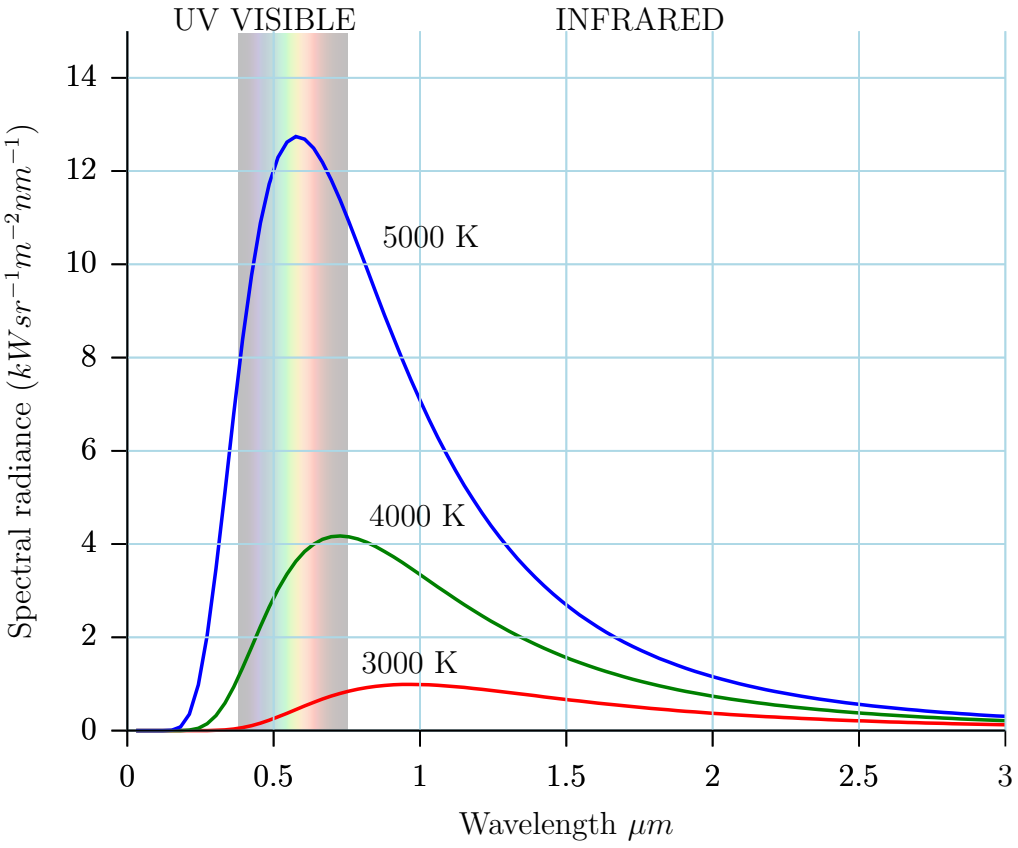


Figure 1.1: Black Body radiation graph, showing spectral radiance of various temperatures for different temperatures. Image by Darth Kule, in public domain

1cm and 3m. Therefore radio astronomy only analyses the upper end of the radiation wavelengths shown in the figure.

1.4.1 Cosmic background radiation

Cosmic background radiation is a remainder of the Big Bang. It is the result of thermal radiation during the early stages of the universe. The signal is characterised as a black-body emission at 2.73K, with a very high red-shift.

1.4.2 Hydrogen Line

Hydrogen is the most common element in the universe. It is prevalent in stars and planets, but is also found in large gas clouds and in the interstellar medium (ISM).

A hydrogen atom consists of a single proton and a single electron. Both the proton and the electron have a spin property with direction. When their spins are in the same direction it is know as a parallel configuration, while spin in opposite directions is know as anti-parallel. In a parallel configuration the atom has a higher energy level than in the anti-parallel configuration. The spin configuration can change from parallel to anti-parallel, but this transformation has a very low probability of occurring, 2.9x10-15/second. However, hydrogen is very abundant in the universe. When the electron changes its spin direction it emits a wave with a frequency of 1420MHz, ($\lambda = 21.106$cm). These signals can be detected on earth when using long integration periods. The redshift of this signal indicates how fast the source is moving away from us. This phenomenon is know as the hydrogen line, or the 21-cm line [5, 6].

The signals useful for hydrogen line observations are very susceptible to interference, because they share a radio band with ground-based transmitters.

1.4.3 Pulsars

Pulsars were first discovered in 1967 by Dame Jocelyn Bell Burnell and Antony Hewish. They are formed when a star becomes a supernova, and subsequently collapses in onto itself. A pulsar is a rotating neutron star, which emits electromagnetic radiation at very precise intervals [7]. Pulsars emit a wideband signal, which we can detect on every rotation. The rotation speeds vary between pulsars, with the slowest known being 8.5 seconds [8].

1.5 Conclusion

This chapter has provided a short introduction to the field of radio astronomy. This thesis will focus on the automatic detection of man-made EM signals that can interfere with the signals captured by a radio telescope. These radio telescopes survey signals with frequency ranges from the low MHz up to a few GHz, which includes cosmic background radiation, the hydrogen line and pulsars. Some radio telescopes observe at much higher frequencies, but these are generally less disturbed by interference.

Chapter 2

Receiving electromagnetic signals

Suppose a plane with area A is receiving electromagnetic signals from the space around it. We will represent the emitting space by a sphere, which is radiating inwards towards the plane. The total power received by the plane depends on the power of the transmitted signals, the frequency of the signals, and the area of the plane. For an infinitesimal point on this plane the power received can be expressed as [9]:

$$\mathrm{d}W = B\cos(\theta)\mathrm{d}\Omega\mathrm{d}\nu\mathrm{d}A \tag{2.1}$$

where

$\mathrm{d}W$	Power in watts
θ	The angle to the transmitting section of the sphere, measured from the vertical reference.
B	Brightness function of the transmitting space, in watts per Hertz per Solid Angle
$\mathrm{d}\Omega$	Solid angle of the transmitting area, which can be expressed in terms of θ and σ
$\mathrm{d}\nu$	Bandwidth of the received signal, in Hertz
$\mathrm{d}A$	Surface area of the plane, in m^2

B is the brightness function of the sky, and is a function of θ,σ and ν. For radio astronomy this is the signal we are interested in. The power received by the entire plane from one solid angle of sky can be obtained by integrating over the bandwidth and the area of the receiver.

$$W = A \int_{\nu}^{\nu+\Delta\nu} \int_{\Omega} B \cos(\theta)\, \mathrm{d}\Omega\, \mathrm{d}\nu \tag{2.2}$$

B is the only term dependant on the bandwidth ν, so by integrating over the bandwidth, $B^{'}$ can be obtained. However, we are usually more interested in the power per unit bandwidth expressed as w, and measured in Watts/Hertz.

$$B^{'} = \int_{\nu}^{\nu+\delta\nu} B \mathrm{d}\nu \tag{2.3}$$

Antennas have a radiation pattern, which is the gain of the antenna in a certain direction. The radiation pattern is usually expressed in spherical coordinates, and replaces the spatial component in the equation. The area component A is also replaced by the effective area, A_e:

$$w = \frac{1}{2} A_e \int_{\sigma} \int_{\theta} B^{'}(\sigma, \theta) P_n(\sigma, \theta) \mathrm{d}\theta \mathrm{d}\sigma \tag{2.4}$$

Antennas are polarized and radio astronomy signals are usually unpolarized, resulting in only half of the signal being received. This leads to the $\frac{1}{2}$ factor in the equation.

This equation shows us that there are two methods of increasing our sensitivity to signals. A larger antenna area can be used (thus increasing A_e), or the radiation pattern can be improved.

2.1 Single Antenna Reception

A radio antenna usually consists of two distinct parts: The feed is the basic EM transmitter or receiver. This can be a small transmitter or a waveguide which provides a signal from a distant transmitter. The signal radiated from the feed (for the transmitting case) usually strikes a reflector. The reflector amplifies the signal by concentrating it into a certain direction. The direction in which radio signals are transmitted, and the power with which they are transmitted are determined by the antenna radiation pattern.

An important concept when dealing with antennas is that of an isotropic radiator. An isotropic radiator is an ideal antenna radiating energy uniformly in all directions. When the antenna radiation pattern is expressed, it is expressed on a log scale relative to an ideal isotropic radiator (dBi).

Another important concept for antennas is the reciprocity principle: All properties of an antenna for the transmitting case will also be valid when receiving signals [9].

2.1.1 Antenna Radiation Pattern

The antenna radiation pattern is a radial graph showing the gain an antenna gives to a signal coming from a certain direction. The radiation pattern for an isotropic radiator is a sphere, since the antenna gives equal gain to signals from all directions. A parabolic reflector has a large gain in the main direction, and several side lobes in other directions. Signals coming in via the side lobes are attenuated, but are still present. In radio astronomy these signals have a significant influence, because they can have a large amplitude even after the attenuation [10]. Signals coming in from the side lobes are also often RFI, which is why they can have relatively high amplitude.

2.1.2 Types of Antennas

Antennas are usually designed to be either omnidirectional or directional. Omnidirectional antennas attempt to radiate or receive in all directions, similar to an isotropic radiator. This is useful when a large area must be covered, as is the case with cellphone or wifi antennas. Directional antennas are used when a point to point connection is required. Directional antennas offer a large gain in a certain direction, allowing the signal to be transmitted and received over large distances. An omnidirectional antenna can however not have high gain as well. In order to be omnidirectional, the antenna generally needs to be small. If the antenna is small it directly means that its gain is low. Only highly directional antennas can have high gain, and consequently are physically large with respect to wavelength.

There are different types of antennas used for radio telescopes. The most common type is a parabolic reflector. This type of antenna employs a large, parabolic shaped dish to reflect the incoming signal to a central receiving feed. The parabolic dish is very desirable. It can be manufactured with a high tolerance and very low sidelobe levels. Unfortunately the feed has to be directly in the field of view of the antenna, and supported very rigidly to ensure low sidelobes. The major disadvantage with the parabolic dish is the required

support structure that sit in the field of view of the antenna. This structure can interfere with the radiation pattern, causing additional sidelobes.

To overcome this problem some antennas only use a part of the parabolic shape, called an offset feed. These antennas offer less gain but a cleaner antenna pattern, because the supporting structure of the feed does not interfere.

The feed does not need to be in front of the antenna, as is the case with Gregorian and Cassegrain antennas. These antennas use an additional reflector, which is located at the focus point of the main reflector, to reflect the signal towards the feed. The additional reflector lowers the signal intensity, but allows the complex receiving hardware to be housed inside the main body of the antenna, rather than being exposed at the feed in front of the dish.

The antennas used by the KAT7 are prime focus antennas, which have the feed at the focus point of the main reflector [11]. The MeerKAT antennas are Gregorian-offset antennas, which use a secondary reflector. The antennas also only employ part of a parabolic shape [12].

2.1.3 Resolution

A single antenna can provide only a certain resolution. The resolution of the antenna determines the minimum separation two sources can have and remain distinguishable. The resolution is determined by the wavelength and the size of the dish. For a parabolic receiver:

$$\theta \approx \frac{\lambda}{D} \tag{2.5}$$

Here D is the effective diameter of the dish, θ is the angular resolution and λ is the wavelength under observation. The only method to increase accuracy is to increase the dish size. This quickly becomes a structural problem, as moving a large dish accurately requires high power, high-precision control. Instead an array of receivers can be used. For an array of parabolic antennas, the angular resolution is determined by

$$\theta \approx \frac{\lambda}{B} \tag{2.6}$$

where B is the longest baseline in the array. The baseline is the distance between an antenna pair, and is explained in Section 2.2. It is easy to increase the baseline by building the antennas further apart from each other. However,

this only results in a single observation, but multiple observations are results are required to fill out the same section of sky as a single antenna.

2.2 Multiple Antenna Reception

Most radio telescopes use an array of antennas to improve the resolution of observations that can be made. This section explains how and why this is done [1, 13, 14].

2.2.1 Multiple Antennas

Two similar antennas P1 and P2 are observing the same part of the sky, shown in Figure 2.1. Two different coordinate systems are used. The (l, m) coordinate system is a direction cosine coordinate system used to reference positions in the sky. For this the sky is represented as a sphere surrounding the observation position. A direction vector ($\mathbf{S_0}$) gives the general direction of the coordinate system. The direction vector is determined by the central direction of the antennas. The direction cosines are calculated as the cosine of the angle between $\mathbf{S_0}$ and the point in the sky under observation, in the main lobe of the antennas.

The (u, v) coordinate system is a right-handed Cartesian system based on a plane located at the observation position. The u, v plane is always normal to the observation direction $\mathbf{S_0}$, and therefore does not in general rotate with the earth. The unit vectors, u and v are defined so that v points towards the celestial north pole, and u is normal to v and lies in the plane. The coordinate system has another dimension, w. The w unit vector is aligned with the observation direction $\mathbf{S_0}$.

Both antennas in Figure 2.1 are pointed at the same section of the sky. They receive the same signal from the sources, albeit with a small delay due to their geometric displacement. The main beams of the individual antennas are not narrow enough to identify individual components in the sky. Instead, they receive a sum representing all the signal sources in their main beam, as well as any signals originating from sources in the side lobes. The Van Cittert-Zernike theorem allows us to further process the signal to form an image of the observed section of the sky inside the main beam.

The Van Cittert-Zernike theorem originated in the field of optics, but is also relevant to interferometry. The theorem states that, given certain conditions, the mutual coherence of an incoherent source is equal to the complex visibility. The source in question must be far away, so that the wavefront received from it appears coherent.

We are interested in the complex visibility, as it is the intensity of radiation from the sky. It cannot be sampled directly as the resolution of the antennas is not fine enough. Instead the mutual coherence function is measured.

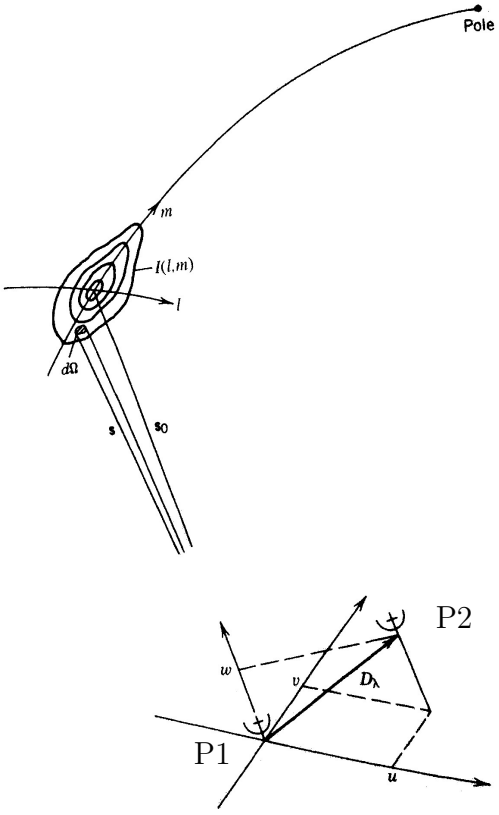


Figure 2.1: Coordinate system for an interferometry setup. Reproduced from [1].

If the mutual coherence is given by Γ, and the complex visibility by $I(l, m)$, the Van Cittert-Zernike theorem can be expressed as:

$$\Gamma_{12}(u, v, 0) = \iint I(l, m) e^{-2\pi i(ul+vm)}\, dl\, dm \tag{2.7}$$

The signals that the telescopes receive are expressed as E_1 and E_2. This allows us to express the mutual coherence is defined as the cross-correlation between the signals:

$$\Gamma_{12}(u,v,w) = \lim_{t\to\infty} \frac{1}{2T} \int_{-T}^{T} E_1(t) E_2^*(t+\tau) d\tau \tag{2.8}$$

The complex- visibility is an image of the sky, and is the goal of the observation. To obtain the image, the (u,v) plane must be filled with observations. Subsequently the complex visibility can be calculated from the (u,v) plane using the Van Cittert-Zernike theorem.

To fill the (u,v) plane the signals from many pairs of two antennas are considered. The signals are filtered to have a very narrow bandwidth, because the Van Cittert-Zernike theorm requires the signal to have a small bandwidth. The filtering can be done using a Fourier transform or a filter bank. The filtered signals are correlated with each other. The value returned by the cross-correlation function can be interpreted as a sample of the mutual coherence function Γ in the (u,v) plane.

Gridding The value returned from the cross-correlation represents a sample from the continuous mutual coherence function. In order to use the FFT to compute the Van Cittert Zernike equation these values should fill in a grid in the (u,v) plane. This process is called gridding.

However, the u and v coordinates are determined by the baseline of the antenna pair and do therefore not necessarily fit on a grid. One option is to place a sample at the closest position in the grid. To express the error made by placing the sample at the wrong position the true value $V(u,v)$ and the sampled value $V(u',v')$ is defined. The sampled value is expressed as convolution of the true value with a gridding kernel G, which is then sampled by a Dirac delta function.

$$V(u',v') = [V(u,v) * G(u,v)]\delta_{(u',v')} \tag{2.9}$$

In the (l,m) domain the equivalent of convolution with the gridding kernal is multiplying the sky with the Fourier transform of the gridding kernel. If the gridding kernel chosen is a rectangular function, the Fourier transform will be a sinc function. This sinc function will be visible in the complex visibility image. Other gridding functions such as a Kaiser window can also be used.

After placing all the samples from observations the (u, v) plane is resampled in order to achieve a uniform filling of the plane.

Baseline The position in the (u, v) plane where the sample is placed is determined by the baseline between the two antennas. The baseline is a vector from the reference antenna to the other antenna. The reference antenna is used as the centre for the (u, v) plane. To determine the position in the (u, v) plane the baseline is expressed as a function of wavelength.

Due to the rotation of the earth, additional baselines are available. When the (u, v) plane is flat on the earth, the baselines are at their longest. As the earth rotates, the length of the baselines change, as well as the relative position of the antennas. This is known as earth rotation synthesis. Figure 2.2 shows a basic example of a two-antenna setup at three different rotational positions. The image shows the positions of the telescopes on the earth, as well as their respective baselines in the UV plane. The antennas are pointed out of the page. Each antenna pair contributes two baseline positions, because each antenna can be used as the reference antenna.

All the possible positions over a certain duration of time can be plotted on a (u, v) plot, giving us the sampling function, $S(u, v)$.

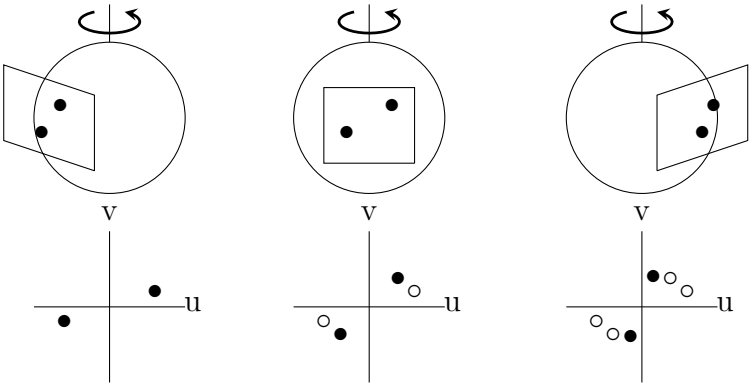


Figure 2.2: Baseline generated by two antennas.

By taking the Fourier transform of the sampling function $S(u, v)$ we obtain the dirty beam $B_d(l, m)$. The dirty beam is also called a point spread function (PSF).

Figure 2.3 shows examples of the different images. The map image is the goal of the observation, while the sampled visibility is the observations. The

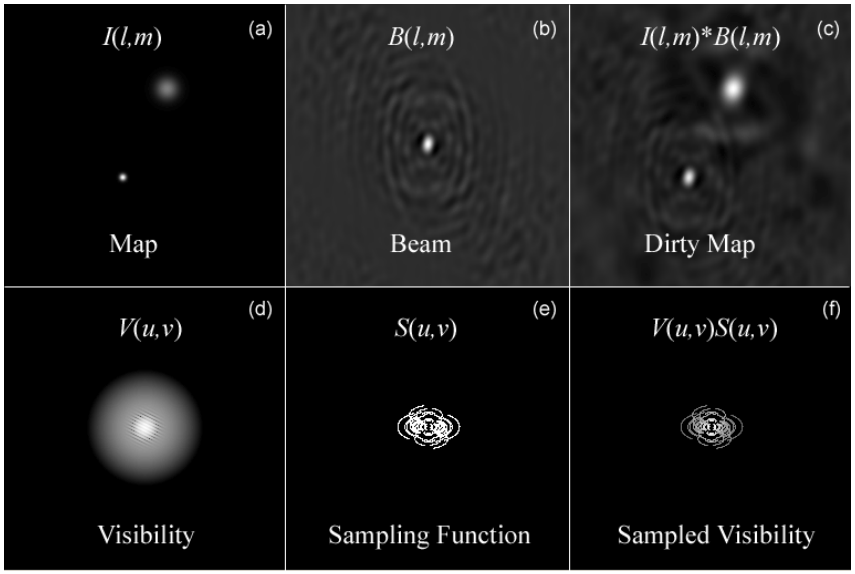


Figure 2.3: Image showing relationship between coordinate systems. Image source https://web.njit.edu/~gary/728/Lecture6.html

sampled visibility is first Fourier transformed to the (l, m) domain to obtain the Dirty Map, and then de-convoluted with the Beam image.

2.2.2 Additional effects

There are additional effects that complicate computing the coverage of the average of the (u, v) plane. Most of these effects are due to geometric assumptions. The earth does not form a perfect sphere, so the UV coordinates of baselines need to be adjusted, taking this into account.

Furthermore, the rotation of the earth causes a Doppler effect, which in turn causes a fringing pattern in the data. However this effect is small compared to other fringing effects. Two antennas receiving the same signal will cause a fringing pattern in the data.

2.3 Conclusion

This section discussed the theory behind electromagnetic signal detection. Some of the basic equations behind EM reception were introduced. After that, the influence of antennas on the signal was discussed. After that the reasons for multiple antennas in radio astronomy was presented, as well as some of the challenges that go with it.

Chapter 3

Literature Review

There are various sources of RFI that pose a problem to radio telescopes [15]. The sources can be broadly divided into accidental radiators, such as construction equipment, and deliberate transmitters, such as radios. Deliberate transmitters generally have a narrow bandwidth, while accidental radiators may have a wide bandwidth. However, the power present in deliberate transmitters can saturate the front-end of the radio telescope receiver. This renders the digitized signal useless, even if the RFI is only present in a single band. An additional category of RFI which is not considered here is that of self-interference. This occurs when a signal internal to the receiver (such as a clock signal or a data signal) leaks out and is transmitted.

The most prominent method of fighting RFI is to simply keep the area surrounding a radio telescope free from possible transmitters. This is possible when constructing new telescopes such as MeerKat. With existing telescopes such as LOFAR this is nearly impossible. In this case, active methods are often required. In other cases, a simple filter would suffice.

When RFI is detected though some method, the data is flagged as containing RFI. The resolution of the data that is marked as RFI depends on the type of observation. When data is flagged as containing RFI, it is removed from the observation.

3.1 RFI mitigation using additional antennas

Various solutions for removing RFI by using secondary antennas have been considered. The secondary antennas are low-gain antennas, and are only sen-

sitive to the interference signal. Correlated components between the secondary signal and the primary signal can be removed from the primary. However, because the primary antenna can usually rotate, it is susceptible to differing amounts of RFI. Hence the secondary antenna may detect RFI when not detected by the primary, or vice versa.

The secondary antenna illuminates a much larger part of the sky and surrounding horizons, in order to intercept the interference signal. Consequently it observes much more noise than the primary antenna. It is mostly undesirable to add any portion of the secondary antenna to the primary antenna as it will increase the noise level.

In [16] 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 minimised.

In [17] a phased array is used to detect and record interfering signals. A phased array is used to better control the antenna pattern of the receiving antenna. The antenna used in these experiments is a six element hexagonal array

Another RFI mitigation method using multiple antennas is discussed in [18]. Here the received signal is divided into different frequency bins by a filter. The cross correlation between frequency bins from different antennas is computed. This results in a correlation matrix. By estimating the rank of the matrix using the eigenvalues, the number of RFI signals can be determined [19].

3.2 Thresholding based methods

A common method of flagging RFI is to use a threshold. Thresholding will flag RFI when the power exceeds a certain level. There are various different methods to calculating the threshold level and for determining which samples should be flagged. The threshold can also be set globally or varied according to signal properties. After samples in the signal are identified as RFI they are usually removed from the signal.

In the cumulative sum (CUSUM) method, small frames of samples are summed together, and an average calculated. If this average exceeds the threshold all the samples fully within the considered frames are flagged. This

method is not as effective for determining precisely which samples contain RFI, but can react quickly to new RFI events.

Combinatory thresholding extends the CUSUM method [20]. Using this method, the frame lengths and the threshold for each frame are varied. The average for small frames needs to exceed a large threshold, while the average for a larger frame has a lower threshold. To find the threshold for each window the VarThreshold or SumThreshold methods can be are proposed.

VarThreshold The threshold is calculated using the formula

$$\chi_i = \frac{\chi_1}{\rho^{log_2 i}} \tag{3.1}$$

Where i is the number of samples in the frame, and χ_1 the threshold for a single sample. A value of $\rho = 1.5$ is suggested based on empirical optimization.

SumThreshold The SumThreshold method is a extension of the VarThreshold method. A large sample will be flagged in a short frame, but might also be detected in a longer frame. If it is detected in a longer frame other samples around it will also be flagged as RFI, even though they contributed little. To avoid this, the SumThreshold starts with the smallest frame length, and replaces any flagged samples with the threshold value for that window [20].

3.3 Statistical methods

3.3.1 Surface Fitting and smoothing

A function $V(v, t)$ can be 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. Such sources are not smooth and will be filtered out by this method. After a function has been fitted over the data the remainder should contain RFI and other spurious noise signals.

Several fitting functions have been suggested. In [20] a two-dimensional, low-order, dimension-independent polynomial is suggested. The time-frequency

data is divided into tiles, and a least squares fit is performed on each tile. Values from previous iterations can be excluded by including a weight function. The act of dividing the data into tiles causes the fringes of the tiles to have an influence. To prevent this overlapping tiles can be used. However, the overlap does not present the astronomical signal very well.

3.3.2 Singular Value decomposition

Data from an antenna is Fourier transformed, and placed through a Singular Value Decomposition. It is assumed the highest singular values correspond to the RFI, and they are set to 0. The values representing RFI are strong outliers, while the Gaussian nature of the source forms a smooth curve. This method does not work when the frequency content of the RFI is not stationary. This method can be applied to the baseline data between each combination of telescopes, or it can be applied to each antenna individually to flag the autocorrelations [20].

3.4 Post-Flagging Techniques

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

In [21] the statistics of RFI events are investigated. Data from the Parkes Multi beam Pulsar Survey is used with thresholding, 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.

3.4.1 Morphological Algorithm

An algorithm based on the mathematical principle known as dilation was proposed in [22]. The antenna data is first processed by some 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 then flags additional samples around already flagged data, based on various criteria. The algorithm will produce an array of additional flags for the data .

The morphological algorithm assumes that the samples surrounding flagged samples are likely to also contain RFI, but with lower power. These samples are not detected by previous algorithms, but can still interfere with processing. The algorithm flags additional samples around flagged samples, based on how many samples were originally flagged. The algorithm only processes one dimension at a time, but can be applied to any number of dimensions. The order in which the dimensions are processed is important.

3.5 Signal classification techniques

In this section we will describe signal and pattern processing techniques that we will later consider for the detection of RFI.

3.5.1 Principle Component Analysis

Principle component Analysis (PCA) is a data reduction method. When given multidimensional data, PCA computes the dimensions along which the data has the most variance. By selecting the dimensions accounting for the greatest variance, the data can be projected down to a lower-dimensional space. This makes it useful for data visualization, since multidimensional data can be projected to two dimensions. However the process is generally lossy and thus non-reversible. PCA can also be used as a general dimension reduction step for higher dimension data which makes the training of models quicker.

PCA requires a data set from which to calculate the new coordinate system. This data set should be representative of the final data, since features that are not present will not be able to contribute to the calculated variances. Consider a data set $\mathbf{D}$ consisting of `n` vectors each of dimension `d`. PCA first computes the covariance matrix for the data $\mathbf{D}$, a `(d x d)` matrix. The eigenvectors $\mathbf{w}_i$ and corresponding eigenvalues λ_i of the covariance matrix are then calculated. The eigenvectors are sorted in order of descending eigenvalue. The eigenvectors corresponding to the largest k eigenvalues are then selected as the new coordinate system, where k is the final (and smaller) number of dimensions required. These are combined into $\mathbf{W}$, a transform matrix.

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots \mathbf{w}_k] \tag{3.2}$$

Here the column vectors $\mathbf{w}_i$ are the sorted eigenvectors. The matrix $\mathbf{W}$ can then be used to transform the data $\mathbf{D}$ from n to k dimensions.

$$\mathbf{G} = \mathbf{W}^T\mathbf{D} \tag{3.3}$$

Here $\mathbf{G}$ is the transformed data matrix with dimensions `(k x n)` and $\mathbf{D}$ is the original data `(n x d)`.

3.5.2 K-nearest neighbour classifier

The K-nearest neighbour (KNN) technique is a non-parametric classification algorithm. KNN uses the training data directly to classify a new data point and does not attempt to represent the data using a model. When a test point to be labelled is introduced, the k closest points to it are determined. The label that occurs most often among these k points is used as the label for the new data point. For non-numeric features the distance to the closest point must be calculated using other functions.

Increasing k increases the computational complexity of the algorithm. By setting $k = 1$ the point closest is chosen as the classification result.

The KNN classifier is conceptually simple, and fairly straightforward to implement. However, because the entire training set is required at run-time, it suffers from a high memory requirement. Finding the k closest vectors can also be computationally demanding, although this can be mitigated by using appropriate data structures and/or search techniques such as dividing the search space into a KD tree [23].

3.5.3 Gaussian Mixture models

A Gaussian mixture model (GMM) is a generative classifier, which fits Gaussian distributions to labelled data. For each of the K classes, N Gaussian distributions are fitted to the data. Each Gaussian mixture is represented by a mean vector μ_i, a covariance matrix Σ_i, and a mixture weight w_i. Due to the high number of products in a full covariance matrix, it is sometimes approximated by a diagonal matrix. The mixture weight w_i represents the prior probability of the mixture within the GMM. The probability that a vector x_i belongs to a class λ is the sum of the probabilities for each of the N Gaussian distributions.

$$P(x|\lambda) = \sum_{i=1}^{N} w_i g(x|\mu_i, \Sigma_i) \tag{3.4}$$

The probability density $g(x|\mu_i, \Sigma_i)$ is defined as

$$g(X|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left\{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)\right\} \tag{3.5}$$

Training The parameters of a Gaussian mixture model must be estimated iteratively using an expectation maximization algorithm. The expectation maximization algorithm works by beginning with an initial estimate of the parameters, and then iteratively improving this estimate. After each iteration the improved estimate replaces the original estimate. This process continues until the successive improvements fall below a predetermined threshold. Initial parameter values can be chosen at random, or estimated be using k-means clustering. The expectation maximization algorithm is vulnerable to local maxima, so it should be run a few times from different initializations and the best model selected.

Optimization To optimise the Gaussian mixture model parameters a tuning data set can be used (See section 3.6.2). The number of Gaussian distributions to fit per class as well as the type of covariance matrix to use can be optimised. In a cross-validation framework, this will generally lead to different optima for each fold. In this case, the median solution can be chosen for the fold of the model.

Classification To classify using this model, a new data point is presented to the system. For every class, the likelihood that the data point was generated by one of the distributions in the class is calculated. The class with the maximum associated probability indicates the classification result. The model can also be used to generate synthetic sample data for a specific class.

3.6 Classifier training and evaluation

3.6.1 Confusion Matrix

A confusion matrix is a simple visualization of the performance of a classification system. It takes the form of a grid, with one axis representing the correct labels, and the other indicating the labels predicted by the classifier. Each element of this grid is an integer value that indicates how many times each true class was classified as each predicted class. This grid can be displayed as an image, giving a quick visual impression of a classifier's accuracy. A perfect classifier will only have values on the diagonal, implying that all points were correctly classified.

3.6.2 Cross-validation

Cross-validation is a method by which small disjoint datasets can be optimally exploited for classifier development. First the entire dataset is divided into N subsets, approximately equal in size. The subsets are also called folds. These N subsets are divided into a training set, a tuning set and a testing set. Often the training set is larger than the tuning and testing sets. The training set along with the labels are used to train the classifier. The tuning set is then classified. This is repeated for all of the N folds, and an average accuracy is determined. The best parameter combination is then selected based on the classification accuracy of the tuning data. The testing data is then classified by the model for that fold, and the results are averaged for a final result. In this way all the data can be used for both training and testing.

3.6.3 F1 score

An F1 score is a measure of classification accuracy in a binary classification problem. The F1 score is calculated as the geometric mean between the precision and the recall. Precision measures how many of the total predictions were accurate and recall measures how many of the positive data points were correctly identified.

$$\text{Precision} = \frac{\text{True Positives}}{\text{Total Classifications}} \tag{3.6}$$

$$\text{Accuracy} = \frac{\text{True Positives}}{\text{Total Positives}} \tag{3.7}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{3.8}$$

The F1 score is always a result between 0 and 1, with 1 corresponding to the perfect classification.

3.6.4 Receiver operator characteristic curve

A receiver operator characteristic (ROC) curve can also be used to describe the accuracy of a binary classification system. The curve plots the true positive rate versus the false positive rate for a varying classifier parameter. This provides a visualization of the influence of the parameter on the classifier. An excellent classifier should have a true positive rate that quickly approaches 1, and then stays there as the parameter is varied.

3.6.5 Detection error tradeoff curve

A Detection error tradeoff(DET) curve is similar to the ROC curve, but rather plots the false negative against the false positive rate. A DET curve is usually plotted with both axis on a logarithmic scale. The DET curve visualises both types of errors, whereas the ROC curve only visualizes the false positive rate. A DET curve for an average classifier would usually be visualized as a line diagonally down. A line that lies closer to the bottom-left corner represents a better classifier [24].

3.7 Conclusion

There are various existing methods used to detect RFI events. This chapter discussed some of these methods. Many of the methods use additional antennas to detect RFI local to the antenna. Other methods such as the SumThreshold method use a threshold to detect and flag RFI events.

Methods to analyse the RFI data are also investigated. A K-nearest neighbour and Gaussian Mixture model classifiers are investigated and explained. Other tools used to determine and visualizing the result and accuracy of the classifiers are also discussed.

Chapter 4

Data Capture and Corpus Compilation

This section discusses the processes and equipment employed to capture data for analysis. Different sources were captured on-site, using a wideband antenna and time-domain capture device. The data capturing setup is described, and some of the sources are discussed.

4.1 Data capturing and processing

To apply machine learning to a problem, data is required. For RFI identification the data will be in the form of a time-domain signal, containing the signal from the offending source. Many different captures are required, 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. An anechoic chamber is a room lined with radio frequency absorbent material.

Capturing signals in this sort of environment presents two problems. First, there is uncertainty if the signal captured has any similarity to the real world signal. Secondly, some some sources of RFI are to big or immobile to transport to an anechoic chamber.

A visit to the SKA site was performed in September 2014, with the goal of capturing data for machine learning analysis. To perform the data capture an RTA was used with an LPDA antenna. These were provided by the SKA

office in Cape Town. When data capturing was performed at the KAT7 site the LPDA antenna on the RFI trailer was used. We captured data from both the time and frequency domain, in various different frequency bands.

Figure 4.1: Karoo site map. Image obtained from Google Earth, 29 September 2015.

Figure 4.1 shows a map of the Karoo site. The site is about 80km from Carnavon in the Northern Cape. It lies in a farming area, and is a radio quiet area.

The KAT7 site hosts all 7 of the KAT7 telescopes. The first Meerkat dish being constructed is M63, which is the closest dish to the KAT7 site. Losberg is a hill sheltering the processing site from the core of the SKA (off the map to the north). The processing site hosts all the processor buildings as well as the assembly shed and accommodation for visitors. The diesel pumps are also located here. Meysdam is an old farm house located to the North-East of the site. The site is now being used to house the workers and equipment used for the construction.

4.2 Equipment

4.2.1 RTA

The Real Time Analyser (RTA) is high-speed data capturing device, which can perform data capturing in both the time and frequency domain [25]. The RTA is based on the ROACH (Reconfigurable Open Architecture Computing

Hardware) platform, and was previously also known as the Ratty2. It can perform 10 bit sampling of an input voltage at 1.8 GSa/s. The RTA has support for 4 different frequency bands.

The RTA can perform a data capture in two different modes: time domain, and frequency domain. In the time domain mode, samples are recorded to the RTA's RAM, and then transferred to a computer via an Ethernet connection. In this mode the length of the capture is limited to 8 microseconds. This is a very short duration capture, but is the best that the available hardware can provide.

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. The filter bank effectively generates a frequency domain representation of the signal. The resulting spectrum of the signal is summed over the specified time. This means that the frequency capture mode is not very suitable for capturing transients, but can instead be used to detect any low-powered stationary RFI signals [25].

The RTA can capture in four different bands. These are shown in Table 4.1. For the data processing the four bands are treated as separate cases. The band must be configured before a capture is started.

Table 4.1: 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. When starting a capture of a new RFI source, the highest attenuation is used. This is done to protect the front end of the RFI. If no signal is detected, the attenuation is lowered until the signal fills the range. The attenuation can subsequently decreased until the signal is at an acceptable level. The attenuation has a maximum setting of +90db, and a minimum of 0db.

The RTA can trigger on data in two different modes when capturing in the time domain. When the RTA triggers it starts storing data in the buffer until the buffer is full, and then transmits data to the computer. It can trigger as soon as the signal exceeds a certain threshold. It can also operate on a free-running trigger. In this mode it will start sampling a new capture as soon as the previous capture has been transmitted to the computer. As far as possible this mode was avoided, since it provides no assurance that any signal will be present in the data.

4.2.2 LPDA antenna

The antennas used for data capturing are Log Periodic Dipole Arrays (LPDA). An LPDA antenna consists of dipoles of increasing size arranged in a straight line. Each second dipole is connected in reversed phase. An LPDA antenna operates over a wide frequency band and is directional, making it ideal for capturing RFI sources.

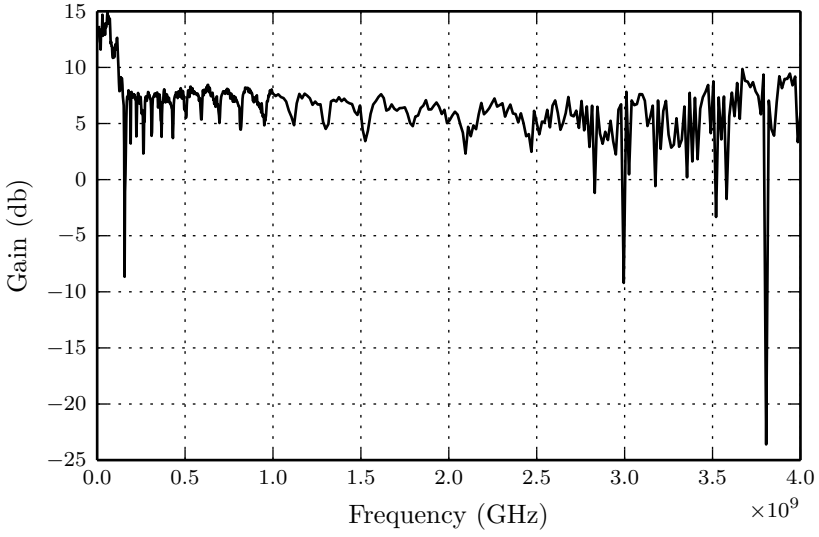


Figure 4.2: LPDA gain over frequency

Figure 4.2 shows the gain the antenna has over a frequency range. It has a very consistent gain over the MHz range, and only breaks up in the lower GHz.

4.2.3 RFI trailer

The RFI trailer is a small unit used on-site to find RFI signals. It contains some detection and processing equipment. We only used the trailer's LPDA antenna, a HL033 LPDA [26]. This antenna has a frequency range from 80 MHz to 2 GHz. The antenna is attached to a mast, so it can be raised, lowered and rotated from ground level. The antenna was used for all captures at the KAT7 site. Figure 4.3 shows the trailer with the mast raised and the antenna attached.

Figure 4.3: The RFI trailer used in some of the captures. Photo credit: Paul Manners (SA-SKA/HartRAO).

4.3 Sources of RFI

Various different sources of RFI available on site were captured. The sources were selected based on availability on site. For each of the sources a baseline capture was also done. This capture is done without the source transmitting, in order to determine the background signal levels in the area where the signal was captured from. Figure 4.4 shows the setup used for capturing. Table 4.2 shows the sources available on site.

Most of the samples were obtained from band 1, which lies between 50MHz and 850MHz. For all of the sources we attempted to capture data in the higher

bands, but for most of them we did not capture any useful or identifiable signals. Most of the further analysis will be on data from band 1 of the RTA.

For most of the sources the antenna was located close to the source, within 3 to 6 meters. The only exceptions to this were the Meysdam refridgeration unit, which was captured from 3m, 15m and 25m, and the Meerkat compressor, which was captured from about 100m away. The antenna was kept in a vertical polarization.

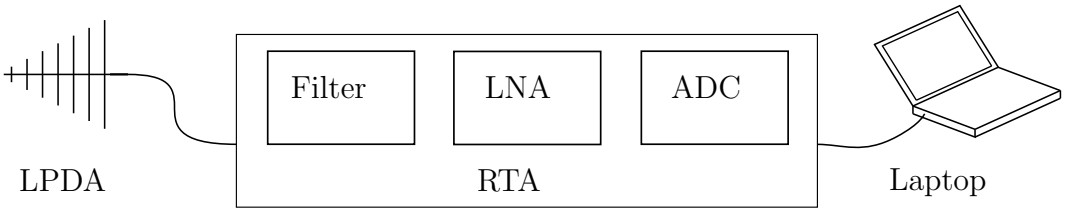


Figure 4.4: The setup used for data capturing

Some of the sources were easier to capture than others. Sources under our direct control such as the welder and the diesel filter pumps could be turned on and off. Other sources such as the Meysdam Refrigeration unit and the Meerkat compressor were not under our control, so we had to wait for them to turn on automatically.

The same source was captured multiple times in order to establish a database of the signal.The gain setting on the RTA was varied per source. The gain setting was recorded, but not taken into account when processing the data. A normalization step replaces the gain setting.

Table 4.2: Table of signal sources

Source
Meysdam Refrigeration unit
Diesel Filter pump
Meerkat Compressor
Crane and Cherry picker
Welder
Vehicle electronics
Radios

Meysdam Refrigeration Unit There is a temporary housing facility at Meysdam for the on-site workers. One of the facilities located there is a refrigerated shipping container, used to store food. The compressor can cause unwanted RF emissions when it turns on and off during the day. The container is referred to as a reefer on site.

Diesel filter pump Diesel is stored on-site and is used both for electricity generators and for on-site vehicles. A filter pump used for the diesel switches on intermittently and can cause unwanted interference. RFI measurements of the pump were taken from up close. The diesel pump was turned on manually.

Meerkat Compressor The compressor used to cool the Meerkat digitizer switches on and off at a regular interval, producing interference. Measurements of the Meerkat compressor were taken from the KAT7 position using the RFI trailer. No closer measurements were possible, because no power was available at the Meerkat site.

The other compressor was a small standalone compressor used with the RFI trailer to raise the mast up.

Crane and Cherry picker There are two cranes that are used on site: one large crane used for construction, and one cherry picker used for lifting people up to the receiver dish for construction. The first can be operated using a remote, which produces an RF signal. This interference was measured at the Meysdam site. RFI generated by the cherry picker was measured inside the assembly shed at the processing site.

Welder A welder was measured inside the assembly shed at the processing site. Two distinct signals were identified, one when the welder was sparking, and another during welding.

Vehicles There are various vehicles used on site. There are a few bakkies, as well as a VW combi used to transport workers. All the vehicles are diesel vehicles. We measured the RFI produced by the alternators of the vehicles, by the lights and by the two-way radios. It proved difficult to capture signals from the vehicle because of the long duration of an alternator cycle the compared to the capture time.

Lightning While capturing data from the KAT7 site, an approaching lightning storm was noticed. It is possible that the lightning interfered with the signals we captured. These samples were marked as lightning and not used in classification, as there is uncertainty if the captured signals contain signals from the lightning.

4.4 Initial Data Labelling

Once the data was captured, it was labelled according to notes taken during the capturing. Any data containing spurious signals are discarded. A assessment of the number of available data points is then made.

Naming conventions In this section, the following names are used to refer to files and captures.

Sample	One instantaneous sample (a scalar value).
Frame	A collection of consecutive samples, usually 1024.
Capture	32768 consecutive time samples, the maximum number of samples the RTA can capture.
File	A file contains multiple captures, all of which are of the same source and use the same attenuation.

A meta-data file is kept for every data file. This file notes the source being captured and the attenuation used. For every capture in the data file, a label is stored in the meta-data file as well. This labelling is used for the first section of Chapter 5.

4.4.1 Visualizing the data

To visualize a single capture, a spectrogram is computed. This is done by dividing a single time capture into overlapping frames. These frames are 1024 samples in length, and overlap by 512 samples. This frame represents a 0.569ns section of the original signal, sampled at 1.8GHz. At this point an FFT can be taken of the data to produce a spectrogram. However, in order to reduce the number of data points, an average spectrogram is calculated instead. The 1024 frame is further divided into 128 point segments. An FFT with a Hamming window is applied to these segments, and the results are averaged. Since this

is a real signal, one half of the FFT result is discarded, along with the DC component. These averaged segments represent the frequency content of the frame, and are later used as feature vectors. Figure 4.5 shows the general process used to extract the feature vectors. Figure 4.6 shows several examples of the feature vectors from different sources.

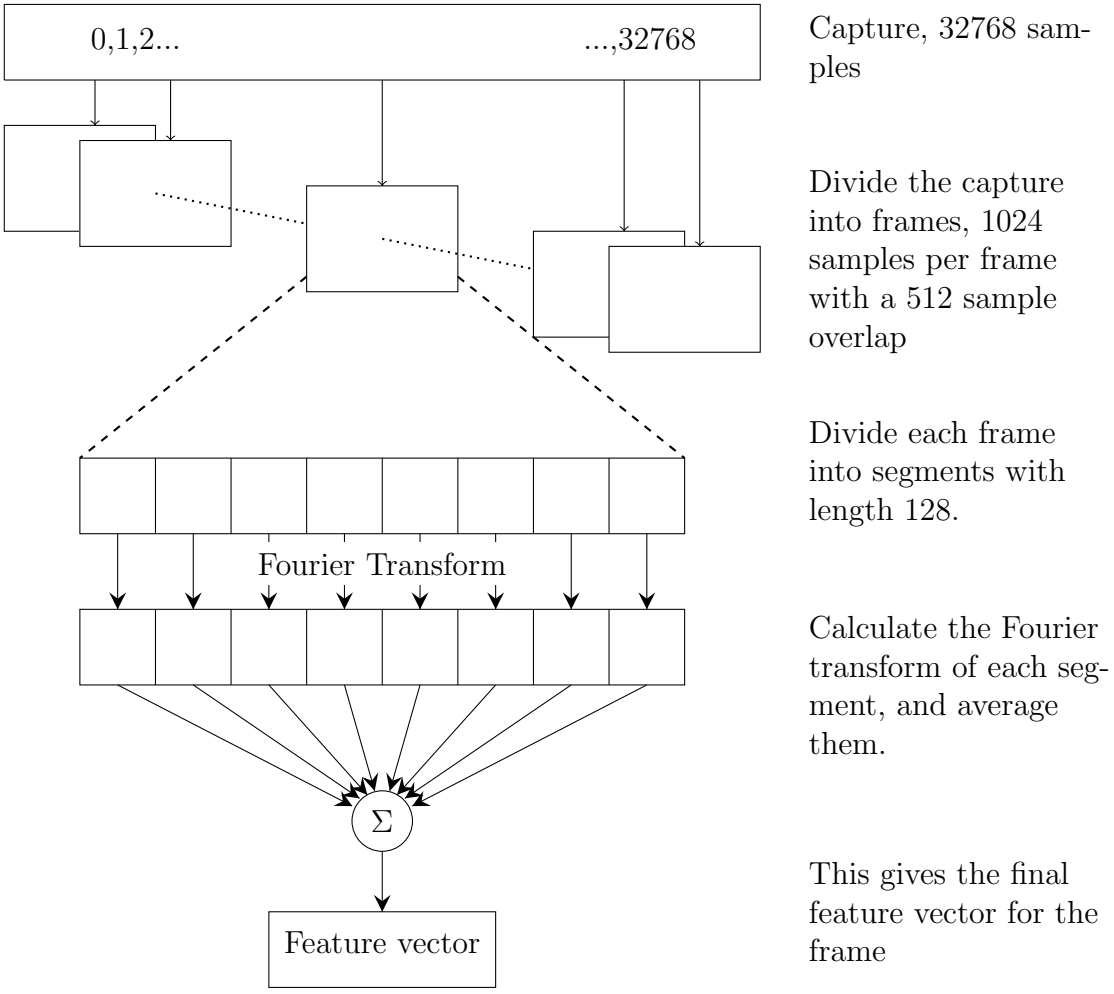


Figure 4.5: Visualization of data division.

Spectrograms of different captures of the same RFI source are then compared to notes made during data capturing. The captures are labelled according to the data source.

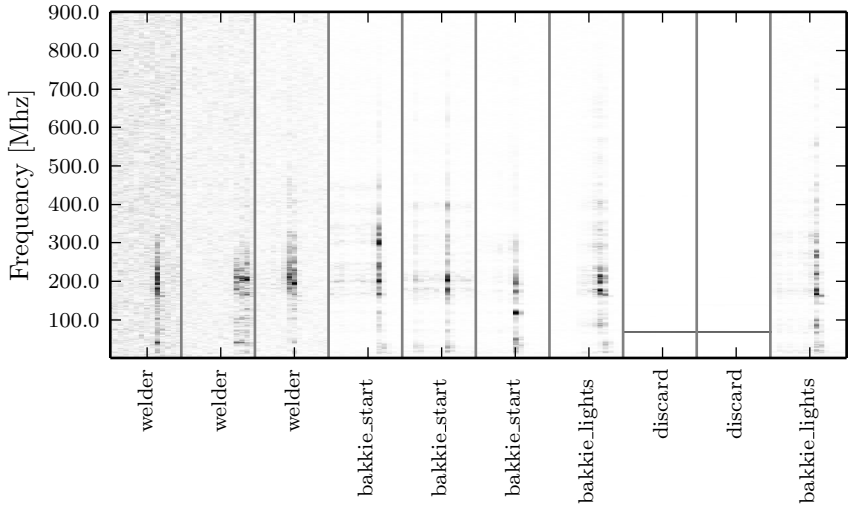


Figure 4.6: Example spectrogram of several different sources.

4.4.2 Removing outliers

Any corrupt captures are removed. These include empty captures containing no data, or captures containing any other interference signals such as radio signals. Interference signals are identified by comparing all the captures for a specific source. Any capture presenting uncharacteristic spectrograms is removed and labelled as (additional) interference. Figure 4.7 shows an example of such a spectrogram. The source is a two-way radio, which emits a single frequency signal. However, additional wideband bursts are visible in the spectra, where the radio signal has not been captured. These captures are labelled with *discard*.

4.4.3 Available Data

The number of available samples are summed and compared. Table 4.3 shows the raw number of samples available, before any processing was applied. Table 4.4 shows the final numbers as well as all the labels used.

4.5 Individual Frame Labelling

The previous section described the labelling of the data on a per-capture basis. This section describes the more detailed labelling of the resultant data set on

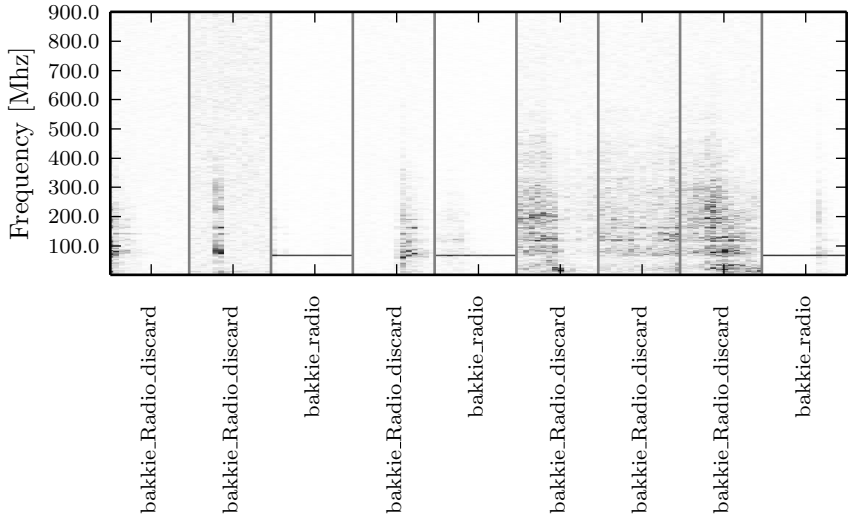


Figure 4.7: Additional Interference in a spectrogram

a per-frame basis. Frames are assigned RFI labels only when their energy exceeds a certain threshold.

The motivation for this step is that inspection of the data 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. The threshold is calculated as a percentage of the total energy in the capture. The threshold was manually adjusted on a file by file basis, but was always kept between 5% and 15%. Any frames that did not exceed this threshold are labelled as silence for that specific class. These silence frames were also considered during classification, to determine whether the models can differentiate between RFI and silence. The per-frame labelling process improves the quality of the labelled data with which classifiers can be developed. It has the negative consequence that some frames containing the signal at a low energy level are labelled as silence. Since it may be expected that the various silence classes are difficult to distinguish between, they were also merged into a single silence class.

Table 4.4 show the number of frames available after labelling the data in this way. Note that for every class an additional class was created to indicate the silence regions taken from captures for this class.

Table 4.3: Number of frames obtained for each RFI source.

Source Name	Band 1	Band 2	Band 3	Band 4
bakkie_radio_discard	25	0	0	0
bakkie_baseline	29	76	64	50
bakkie_lights	30	0	0	0
bakkie_radio	7	0	0	0
bakkie_radio_discard	30	0	0	0
bakkie_start	31	7	0	0
big_crane	15	0	44	31
big_crane_baseline	31	31	29	31
big_radio	26	0	0	0
cellphone	7	59	0	0
cherry_picker	10	0	0	0
cherry_picker_baseline	24	0	0	0
compressor	9	0	0	0
compressor_baseline	75	0	0	0
diesel_filter	114	16	14	13
diesel_filter_baseline	61	46	53	40
discard	548	73	206	52
kat7_meysdam	41	0	0	0
lightning_discard	28	0	0	0
meerkat_compressor	122	0	0	0
meerkat_compressor_kat7	87	0	0	0
meysdam_gap	252	0	0	0
possible_lightning	9	0	0	0
radio	22	0	0	0
reefer	13	5	0	0
vw_baseline	11	0	0	0
vw_discard	8	0	0	0
vw_ignition	24	0	0	0
vw_indicators	28	0	0	0
welder	17	0	0	0
welder_baseline	29	0	0	0
welder_spark	12	0	0	0
total	1775	313	410	217

Table 4.4: Data frames available after per-frame labelling.

Source	Amount
bakkie_lights	180
bakkie_lights_silence	660
bakkie_radio	196
bakkie_start	154
bakkie_start_silence	714
big_crane	420
big_radio	607
big_radio_silence	121
cherry_picker	50
cherry_picker_silence	230
compressor	45
compressor_silence	207
diesel_filter	78
diesel_filter_silence	3142
kat7_meesdam	1148
meerkat_compressor	203
meerkat_compressor_kat7	2293
meerkat_compressor_kat7_silence	143
meerkat_compressor_silence	3213
meysdam_gap	7056
radio_silence	616
reefer	76
reefer_silence	288
vw_ignition	109
vw_ignition_silence	563
vw_indicators	43
vw_indicators_silence	741
welder	110
welder_silence	366
welder_spark	64
welder_spark_silence	272
total silence	11 276
total RFI	12 832

4.6 Conclusion

This chapter explained the method used to record and label data. Data from various RFI sources was captured on-site, using an LPDA antenna and the RTA. Data was captured from multiple frequency bands. The data was labelled, both on a per-capture and per-frame basis, and any outliers were removed. Basic feature extraction in the form of a spectrogram was performed.

Chapter 5

Experimental Results

This chapter describes the application of the classification methods described in Chapter 3 to the data described in Chapter 4, and presents the classification accuracies achieved.

5.1 Classification using capture-based labels

For initial experimentation, a very simple approach to feature extraction was taken. The extracted features are then classified using a selection of classification algorithms.

Feature Extraction First we use only the data described in Table 4.3. The baselines captures representing the background noise levels are not used in this section.

The spectogram for individual frames is calculated, as explained in Section 4.4.1 and shown in Figure 4.5. This results in a feature vector of length 63 for each frame. However, not all frames in a capture represents RFI. The assumption is made that the frame containing the most energy represents the RFI for that capture. The total energy per frame is calculated, and the frame with the most energy is used as feature vector for the corresponding capture.

This feature extraction method is easy to implement, but has some drawbacks. Firstly, it assumes that the part of the signal with the highest energy is representative of the whole signal. This might discard other frames with less energy which can also contribute to the classification. Secondly, this approach

discards any temporal properties of the signal. Thirdly, it greatly reduces the size of the dataset available for classifier training and testing.

Indeed, there are some classes with very few samples and these were therefore removed from the data. The number of frames remaining in the dataset are given in Table 5.1.

Table 5.1: Per-capture dataset available after discarding under-represented classes.

Source	Number of Frames
bakkie_lights	30
bakkie_start	31
big_crane	15
big_radio	26
cherry_picker	10
diesel_filter	114
kat7_meysdam	41
meerkat_compressor	122
meerkat_compressor_kat7	87
meysdam_gap	252
reefer	13
vw_ignition	24
vw_indicators	28
welder	17
welder_spark	12
Total	822

For experimentation, 10-Fold cross validation was used. Six of the folds were used for training, two for tuning and the remaining two for testing.

5.1.1 KNN classification

The KNN classifier was described in Section 3.5.2. Before the KNN classifier can be applied, the value chosen for k has to be decided. This is done by varying k over a range of values, and optimizing the resultant classification accuracy over the tuning set. The average classification results for the various values of k are shown in Figure 5.1. As the value chosen for k decreases, the accuracy of the classifier decreases. The standard deviation is also generally large, indicating that while certain classes may achieve a very accurate result, others might score very poorly.

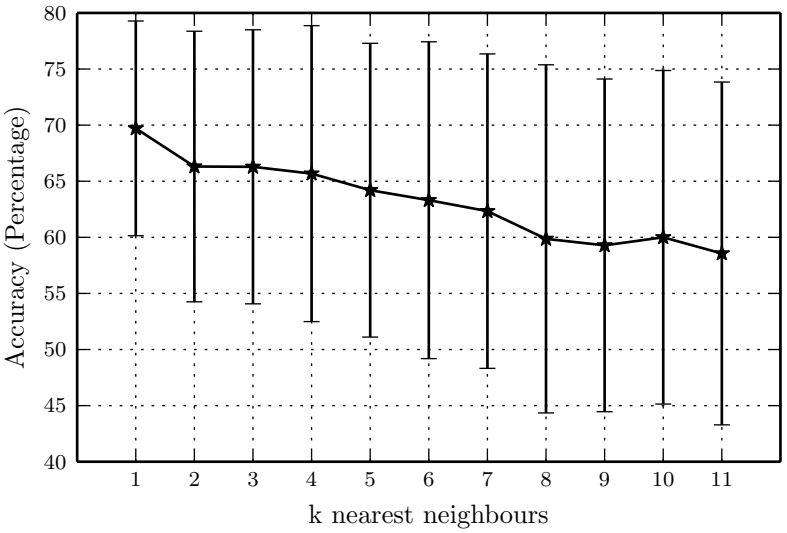


Figure 5.1: Classification accuracy of various KNN models using per-capture labels.

A value of $k = 1$ is selected for the classifier, and the tuning data for each data fold is classified. The results of each fold are averaged. The results are represented in a confusion matrix in Table 5.2. The average accuracy of the classification is 70.80%, with a standard deviation of 30.72.

Table 5.2: Confusion matrix of KNN classification using 1 nearest neighbour and per-capture labelling. The average accuracy is 70.80%with a standard deviation of 30.72.

Actual Value \ Predicted Value	bakkie_lights	bakkie_start	big_crane	big_radio	cherry_picker	diesel_filter	kat7_meesdam	meerkat_compressor	meerkat_compressor_kat7	meysdam_gap	reefer	vw_ignition	vw_indicators	welder	welder_spark
bakkie_lights	**93.3**	3.3										3.3			
bakkie_start	28.6	**37.1**			1.4		2.9	4.3				5.7	11.4	8.6	
big_crane			**100.0**												
big_radio				**100.0**											
cherry_picker	5.0	40.0			**25.0**	10.0					10.0			10.0	
diesel_filter	5.0	20.0			10.0	**15.0**		10.0		10.0		20.0	10.0		
kat7_meesdam							**93.3**		6.7						
meerkat_compressor								**99.2**		0.2		0.6			
meerkat_compressor_kat7							11.4		**68.8**	19.9					
meysdam_gap									6.1	**93.9**					
reefer					3.3	6.7				13.3	**43.3**	3.3	10.0	20.0	
vw_ignition						4.0						**34.0**	38.0	20.0	4.0
vw_indicators							5.0	1.7	5.0	10.0		8.3	**61.7**	8.3	
welder													2.7	**97.3**	
welder_spark															**100.0**

The generally higher values along the diagonal show that many of the samples are correctly classified. However, there is much misclassification in some of the classes, scoring low accuracies. One such misclassification is between the *vw_indicators* and *vw_ignition* classes. Example data points from the two classes are plotted in Figure 5.2. Note that each block has been individually normalized. The spectra of the two classes are visibly similar, with both only

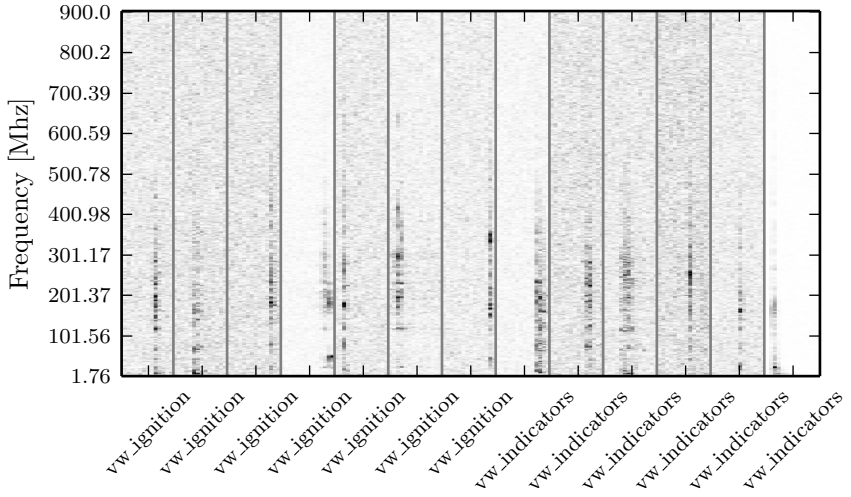


Figure 5.2: Comparison of classes with high confusability in the per-capture KNN experiment.

emitting a short wideband burst. The frequency ranges that they occupy are also similar.

5.1.2 GMM classification

The GMM classifier was described in Section 3.5.3. The most obvious parameter of the GMM under the experimenter's control is the number of Gaussian distributions, n, that are fitted to each of the m classes. Since each distribution will model one of the m classes of RFI, the appropriate number of Gaussians is determined by the shape of the data, and how spread out a class is.

However, the form of the covariance matrix can also be varied by the experimenter. The simplest representation assumes that the data distribution is spherical, so the covariance is represented by a single value. The covariance matrix can also be represented by a diagonal matrix, which assumes statistical

independence between features, or by a full covariance matrix. The number of free parameters in the covariance matrix is 1, d and $\frac{d(d+1)}{2}$ for each approach respectively, where d is the size of the feature vector

GMMs, one for each class of RFI, are trained for each fold of the data. The covariance type is chosen to be either spherical, diagonal or full. The testing data are classified, and the average accuracy over all the folds is used to determine the optimal number of mixtures. The accuracy of a class is determined by the number of correct predictions divided by the total number of data points in that class. The total accuracy is determining by averaging the accuracies of the individual classes. This methods lends equal weight to the accuracies of under and overrepresented classes. The number of mixtures is varied between 1 and 5 to determine the optimal number.

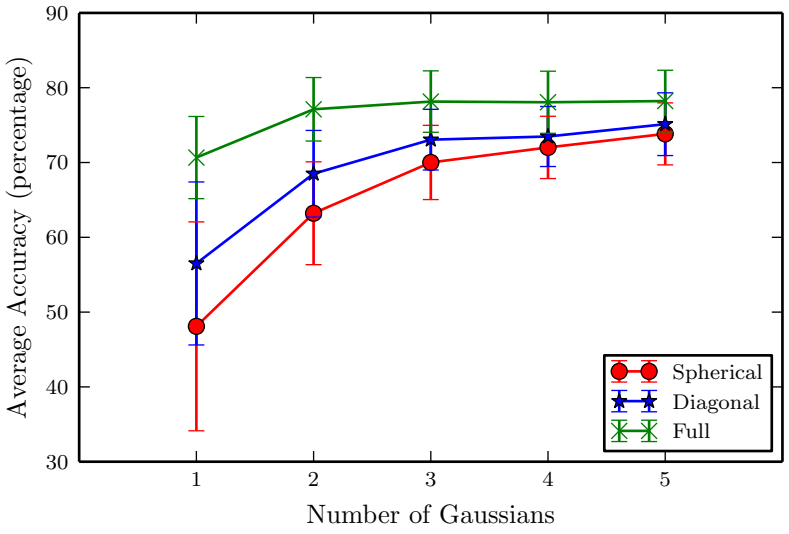


Figure 5.3: GMM classification accuracies for different number of mixtures and different forms of the covariance matrix, using capture-based labels.

Figure 5.3 shows the average cross-validation classification accuracy of the classification of the testing data using different number of Gaussians, and various covariance matrix types. The figure shows that the full covariance matrix scores the best, followed by the diagonal and spherical covariance matrices. The best combination of parameters are a full covariance matrix using 3 Gaussians. Using more Gaussians yields diminishing returns.

Figure 5.3 shows the confusion matrix for a GMM classification using 3 Gaussians per class, and a full covariance matrix. The overall classification

accuracy is acceptable at 65.56%, with a standard deviation of 31.01. Many samples are still labelled incorrectly. Some classes such as the *bakkie_start* and *diesel_filter* are almost unused by the classifier, and have very poor accuracies. Many samples were mislabelled as *kat7_meysdam*, *vw_ignition* and *welder*. Sample data from these classes were inspected visually to assess why the misclassification is so common.

Table 5.3: Confusion matrix of GMM Classification results using 3 Gaussians, full covariance matrix and per-capture labelled data. The average accuracy is 65.56% with a standard deviation of 31.01.

Actual Value \ Predicted Value	bakkie_lights	bakkie_start	big_crane	big_radio	cherry_picker	diesel_filter	kat7_meesdam	meerkat_compressor	meerkat_compressor_kat7	meysdam_gap	reefer	vw_ignition	vw_indicators	welder	welder_spark
bakkie_lights	**90.0**						6.7					3.3			
bakkie_start	41.4	**18.6**			8.6	4.3	5.7					11.4		10.0	
big_crane			**100.0**												
big_radio			13.3	**86.7**											
cherry_picker	15.0	10.0			**50.0**	5.0					10.0			10.0	
diesel_filter		5.0				**5.0**	10.0		10.0			45.0	15.0	10.0	
kat7_meesdam							**100.0**								
meerkat_compressor							1.6	**94.3**		2.2		1.8			
meerkat_compressor_kat7							25.0		**73.3**	1.7					
meysdam_gap							3.1		42.0	**54.9**					
reefer	3.3				16.7		10.0				**23.3**	3.3	23.3	10.0	10.0
vw_ignition							14.0					**46.0**	32.0	4.0	4.0
vw_indicators	1.7						16.7		6.7			11.7	**55.0**	8.3	
welder							13.7							**86.3**	
welder_spark															**100.0**

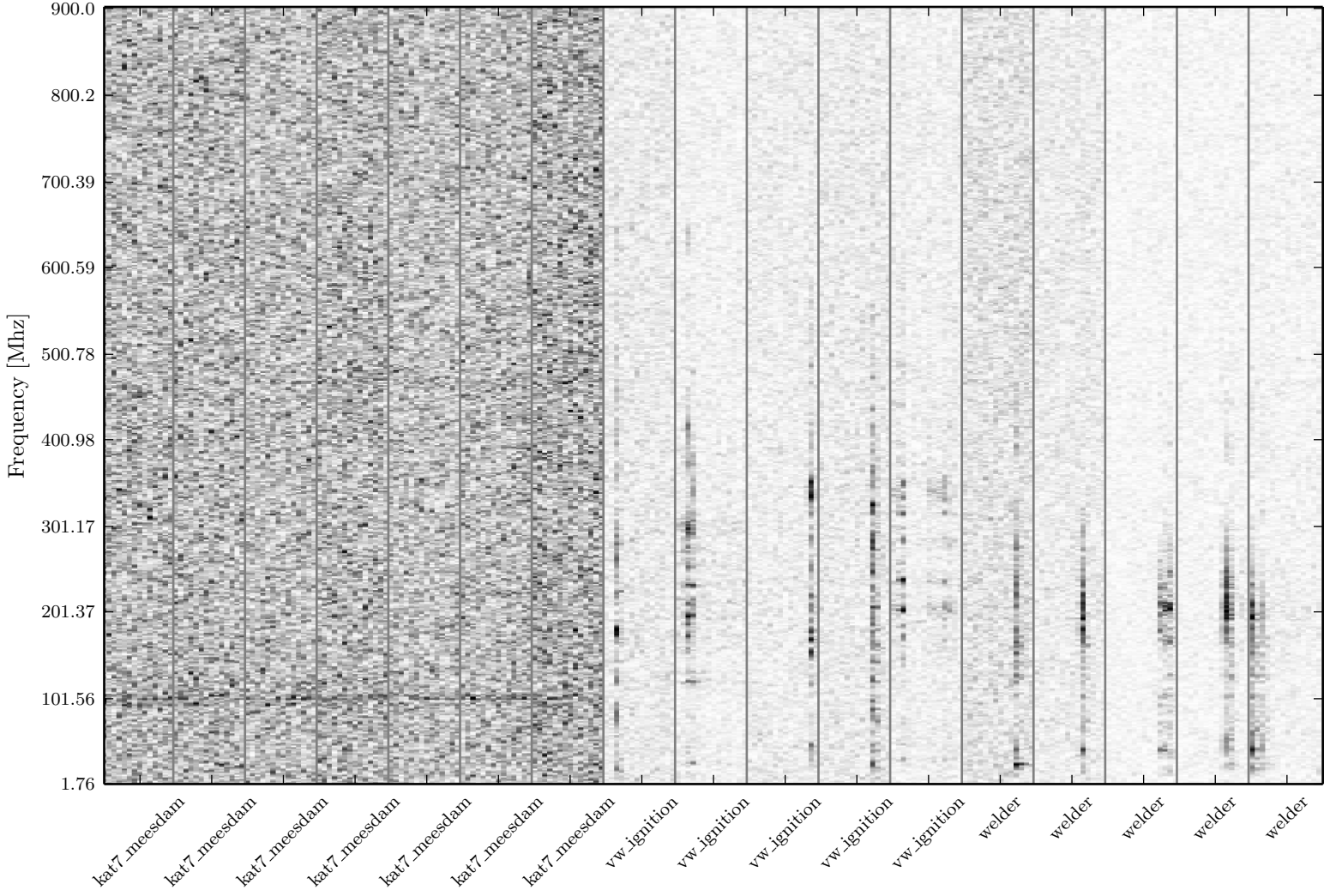


Figure 5.4: Sample data showing classes with high confusability in the per-capture GMM classifier.

Figure 5.4 shows example data from these two RFI classes. Especially the signal in the *kat7_meysdam* samples has very low energies, and the background noise is very clearly present. This class possibly requires better feature extraction to be compared to other classes. The *vw_ignition* and *welder* have more power present. Both present a very common spectra which includes a short, wideband pulse.

For the experiments using capture-based labels, the KNN classifier far outperforms the GMM classifier. The performance of the GMM classifier is surprisingly poor. We suspected that this may in part be due to the approach of using a single high-energy frame from each capture to represent the RFI class. Hence the next section presents results using the data with frame-based labels.

5.2 Classification using frame-based labels

This section describes experiments using the data labelled on a per-frame basis, as described in Section 4.5. This method makes much more data available for classifier training. It also introduces a new *silence* class to the data.

5.2.1 KNN classification

The KNN experiments described in Section 5.1.1 were repeated using the per-frame labelled data. The optimal value for k has to be determined again. The first experiment described in the previous section is repeated. Figure 5.5 shows the accuracy for various values of K. There is almost no difference between any of the values, as all of them have an accuracy in the high 90s. The standard deviation is also very good, having very small values.

For the KNN classifier 1 nearest neighbour is again found to be optimal. The results of the classifier are shown as a confusion matrix in Table 5.4. Combinations where no classifications occur are left empty.

Although it must be kept in mind that the training and testing datasets have changed, the KNN classifier now achieves a much better result than when using per-capture labels. Most remaining misclassifications occur with the *vw_ignition* and *vw_indicators* classes. The classifier also succeeds in classifying the silence class.

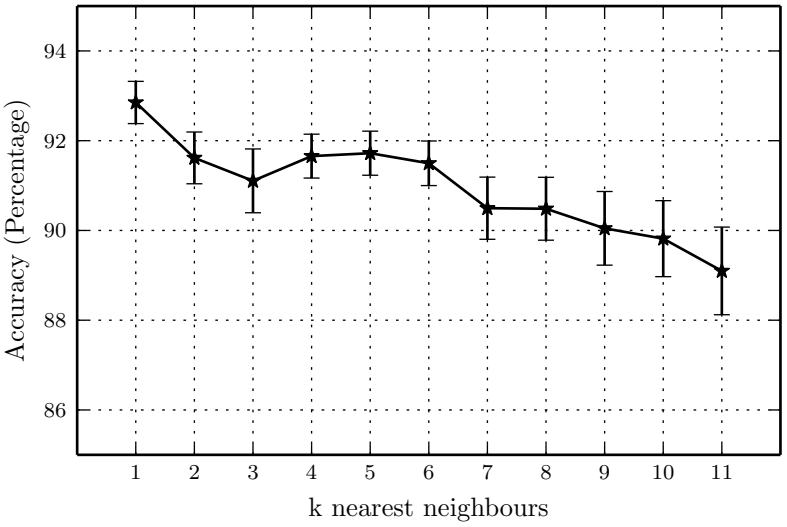


Figure 5.5: Comparison of KNN classifier accuracies for various values of k using frame-based labels.

Table 5.4: Confusion matrix for the KNN classification trained on per-frame labels, using k=1. Average Accuracy of 93.20% and standard deviation of 6.46.

Actual Value \ Predicted Value	bakkie_lights	bakkie_radio	bakkie_start	big_crane	big_radio	cherry_picker	compressor	diesel_filter	kat7_meesdam	meerkat_compressor	meerkat_compressor_kat7	meysdam_gap	reefer	silence	vw_ignition	vw_indicators	welder	welder_spark
bakkie_lights	**89.3**					0.4								10.3				
bakkie_radio		**100.0**																
bakkie_start			**88.0**					0.3						11.3	0.5			
big_crane				**100.0**														
big_radio					**99.9**									0.1				
cherry_picker						**87.1**								12.9				
compressor							**97.8**							2.2				
diesel_filter								**91.7**						8.3				
kat7_meesdam									**100.0**									
meerkat_compressor										**90.7**				9.3				
meerkat_compressor_kat7											**97.6**	0.2		2.1		0.1		
meysdam_gap								0.0			0.0	**99.9**		0.1				
reefer								0.3					**90.8**	8.0		0.9		
silence	0.1		0.2		0.0	0.1		0.0		0.1	0.6		0.1	**98.5**	0.2	0.1	0.0	0.1
vw_ignition			0.6										0.4	15.4	**83.6**			
vw_indicators			3.0					0.5						19.6		**76.9**		
welder			0.1											6.6	0.5		**92.8**	
welder_spark														7.0				**93.0**

5.2.2 GMM classification

This section repeats the GMM experiments of Section 5.1.2 using the per-frame labels. As before, the number of mixtures and the type of covariance matrix is varied and optimised on the tuning set. Figure 5.6 shows the classification accuracies using the various configurations. The diagonal covariance matrix scores the best, outperforming the spherical covariance matrix. The accuracy for the full covariance matrix decreases as the number of Gaussians is increased. This might be due to a lack of data, causing the GMM to be over fitted.

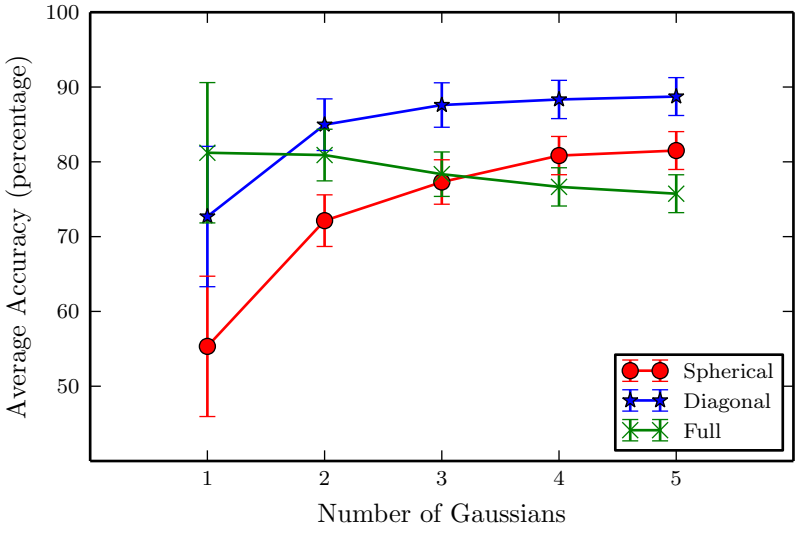


Figure 5.6: GMM classification accuracy for different number of Gaussians and covariance matrices, when using per-frame labelling.

The data is classified using a GMM with 3 Gaussians per mixture and a diagonal covariance matrix. The individual class results are shown in Table 5.5. The classification has an average accuracy of 87.34 and standard deviation of 10.76. The high values on the diagonal clearly show that most of the samples are correctly classified. Many small values are also seen in the silence class, suggesting that some data points are misclassified as silence. The main contributor to these errors is the labelling method used. During labelling the total energy in the signal is compared to a threshold. This means that some frames will be marked as silence even though they contain traces of the interference signal.

The *welder* and *bakkie_start* classes also have many misclassifications.

Table 5.5: Confusion matrix for the GMM classification (3 mixtures, diagonal covariance matrix) when using per-frame labelling. The average accuracy is 87.34% with a standard deviation of 10.76.

Actual Value \ Predicted Value	bakkie_lights	bakkie_radio	bakkie_start	big_crane	big_radio	cherry_picker	compressor	diesel_filter	kat7_meesdam	meerkat_compressor	meerkat_compressor_kat7	meysdam_gap	reefer	silence	vw_ignition	vw_indicators	welder	welder_spark
bakkie_lights	**94.2**									4.7				1.1				
bakkie_radio		**100.0**																
bakkie_start			**80.1**			7.4		1.6					0.6	6.6	2.9		0.8	
big_crane				**100.0**														
big_radio					**99.8**									0.2				
cherry_picker			5.0			**77.5**								17.5				
compressor							**98.3**					0.6		1.1				
diesel_filter			8.7			1.0		**78.6**					5.1	6.6				
kat7_meesdam									**100.0**									
meerkat_compressor			0.2							**85.2**				14.6				
meerkat_compressor_kat7											**77.7**	0.3		22.1				
meysdam_gap											4.0	**95.5**		0.1	0.1		0.3	
reefer			0.3										**83.0**	16.7				
silence	0.6		0.5	0.7	0.2	0.9	0.1	0.2		0.9	0.6	0.1	0.3	**94.0**	0.4	0.2	0.2	0.1
vw_ignition			0.9					0.2				1.8		18.8	**75.3**		3.0	
vw_indicators			2.8					2.3						31.6		**63.3**		
welder			0.1									0.2		20.5	0.9		**78.3**	
welder_spark														8.6				**91.4**

5.3 Higher Frequency Bands

During the data collection process some captures were also made in some of the higher frequency bands. There were however very few usable samples, because few of the sources investigated emit signals in the higher frequency bands.

The features for these experiments were extracted using the same method as previously described, except using data from band 2. No data from band 1 was used.

The data from frequency band 2 (800Mhz-1050Mhz) was classified using both a KNN and a GMM classifier. Various hyper-parameters for the KNN and GMM classifiers are considered. The results are shown in Figure 5.7 and Figure 5.8.

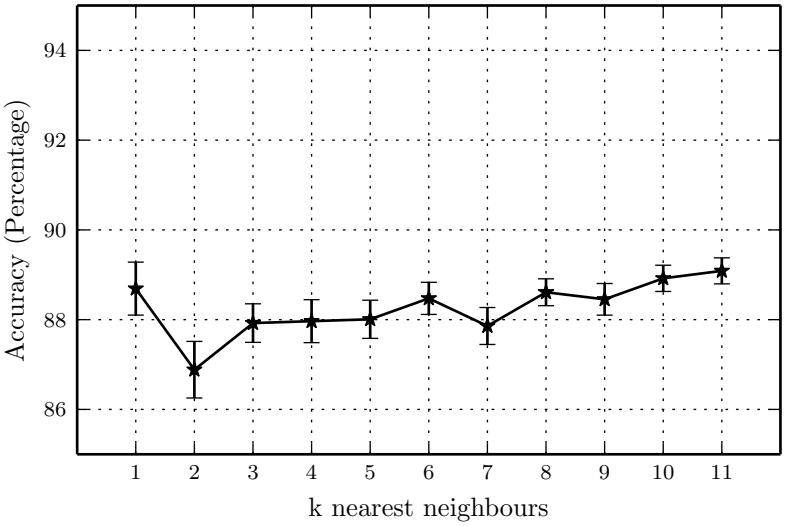


Figure 5.7: GMM classification accuracy for different number of Gaussians and covariance matrices, when using per-frame labelling.

There is no clear optimal value for the KNN classifier. The result for 1, 10 and 11 nearest neighbours are similar. The GMM results show that the diagonal covariance matrix scores the best, followed by the spherical covariance matrix. The full covariance matrix does not achieve an acceptable accuracy.

The KNN classifier uses 1 nearest neighbour; the GMM classifier uses 2 Gaussian mixtures with a diagonal covariance matrix. Tables 5.6 and 5.7 show the covariance matrix results for these experiments. The large values on the

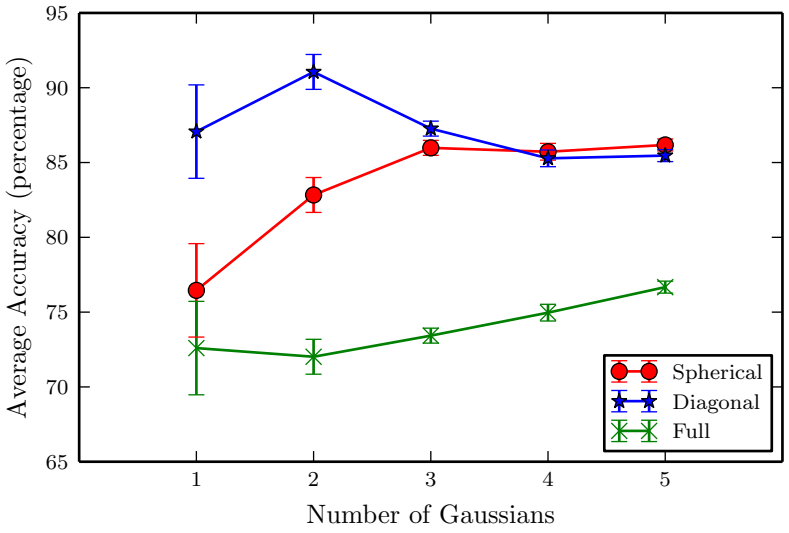


Figure 5.8: GMM classification accuracy for different number of Gaussians and covariance matrices, when using per-frame labelling.

diagonal clearly indicate that most of the classifications are correct. The KNN classifier has an average accuracy of 89.54%, while the GMM classifier achieves an accuracy of 91.38%. Here the GMM outperforms the KNN classifier by a small margin. However the difference between results is less than one standard deviation.

Actual Value \ Predicted Value	bakkie_start	diesel_filter	reefer	silence
bakkie_start	**91.4**			8.6
diesel_filter		**88.5**		11.5
reefer			**80.0**	20.0
silence	0.6	0.6	0.5	**98.3**

Table 5.6: Confusion matrix for KNN classification when trained and evaluated on data from higher frequency bands, with k=1. The per-frame labelled dataset was used. The average accuracy is 89.54%, with a standard deviation of 6.58.

Actual Value \ Predicted Value	bakkie_start	diesel_filter	reefer	silence
bakkie_start	**84.4**			15.6
diesel_filter		**89.3**		10.7
reefer			**96.3**	3.7
silence	1.1	2.6	0.7	**95.5**

Table 5.7: GMM classification confusion matrix using higher frequency bands (2 mixtures, diagonal covariance matrix). The average accuracy is 91.38%, with a standard deviation of 4.84.

5.4 Conclusion

This section presented and discussed the various classification methods investigated. Using capture-based labels proved to be inaccurate, with both the KNN and GMM classifier providing unacceptable results. Table 5.8 shows a comparision between the accuracies of the various classifiers, for data captured from band 1.

Table 5.8: Classification accuracies for GMM and KNN 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

In both conditions the KNN classifier outperforms the GMM classifier.

Classifying the higher frequency data also yields good results, with the GMM achiving an accuracy of 91.38%, outperforming the KNN. However as there are fewer sources emitting in these bands this approach will possibly not be a priority.

Chapter 6

Further feature extraction experiments

In this chapter different feature extraction methods will be investigated. The performance of these methods will be investigated using the same GMM and KNN classifiers as described in previous sections, where applicable. At the end of this section a comparison between all the feature extraction methods is made.

6.1 Classification using reduced feature vectors

This experiment investigates the impact of a smaller feature vector on the classification accuracy. The method modifies the spectrogram calculation in Section 4.4. Rather than dividing each frame into segments of size 128, they are divided into segments containing 32 samples. The average spectrogram of these segments are then calculated. This results in feature vectors with a length of 15, due to it being a real signal and removing the DC component. The frame length is maintained as 1024 samples, in order to preserve the labelling. Using the smaller feature vectors is expected to make classification much quicker, but is also expected to reduce the accuracy of classification.

Experiments similar to those in Section 5 are performed. Both the KNN and GMM classifiers are used. All other configurations, such as data folds, are kept constant.

6.1.1 KNN classifier

The optimal value for k is determined by varying k over a range, classifying the tuning data and selecting the optimal value. Figure 6.1 shows the result of this optimization. All tested values of k show a similar result, and a value of 1 is selected for k. This is different from the previous feature extraction method used in Section 5.2.1, where there was a more gradual difference between values, and a clearer optimal value.

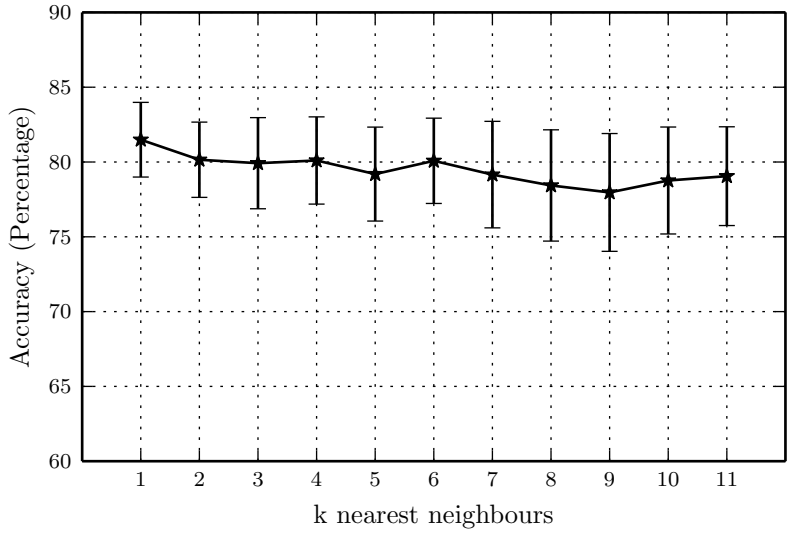


Figure 6.1: Comparison of KNN classifier accuracies for various values of k, using frame based labels and feature vectors reduced to 16 features.

Table 6.1 shows the confusion matrix of a KNN classifier when classifying the testing dataset with $k = 1$. The results are comparable to the result from previous classifiers. The KNN classifier achieves an average accuracy of 81.70% with a standard deviation of 16.36. A comparison between the results is made in Section 6.3.

Table 6.1: Confusion matrix of KNN classification using k=1 and reduced feature vectors. Average accuracy is 81.70% with a standard deviation 16.36.

Actual Value \ Predicted Value	bakkie_lights	bakkie_radio	bakkie_start	big_crane	big_radio	cherry_picker	compressor	diesel_filter	kat7_meesdam	meerkat_compressor	meerkat_compressor_kat7	meysdam_gap	reefer	silence	vw_ignition	vw_indicators	welder	welder_spark
bakkie_lights	**72.1**		4.7			1.6								21.6				
bakkie_radio		**100.0**																
bakkie_start	0.6		**72.9**			1.8		0.6					0.6	21.1	1.2	1.2		
big_crane				**100.0**														
big_radio					**99.2**					0.7				0.2				
cherry_picker	6.5		6.5			**54.8**		6.5						25.8				
compressor							**91.8**					2.0		6.1				
diesel_filter			2.2			2.2		**73.9**		2.2				13.0	2.2	4.3		
kat7_meesdam									**99.9**		0.1							
meerkat_compressor								0.9		**71.6**				27.6				
meerkat_compressor_kat7								0.0	0.2		**93.3**	0.9		5.6				
meysdam_gap		0.0					0.0	0.0		0.0	0.3	**99.3**	0.0	0.3	0.0			
reefer													**82.5**	12.7	1.6	1.6	1.6	
silence	0.5		0.3		0.0	0.1	0.0	0.1		0.5	1.2	0.1	0.2	**96.2**	0.3	0.3	0.2	0.2
vw_ignition			1.6										3.3	27.6	**66.7**		0.8	
vw_indicators			2.3			2.3		4.7					4.7	41.9		**44.2**		
welder			0.8										0.8	20.3	0.8		**77.2**	
welder_spark													2.9	22.1				**75.0**

6.1.2 GMM classifier

The GMM classifier is also used to classify using the reduced feature vectors. GMM models are trained for different covariance matrix types and mixture numbers. The tuning data is classified using these models and compared. The classification results are shown in Figure 6.2.

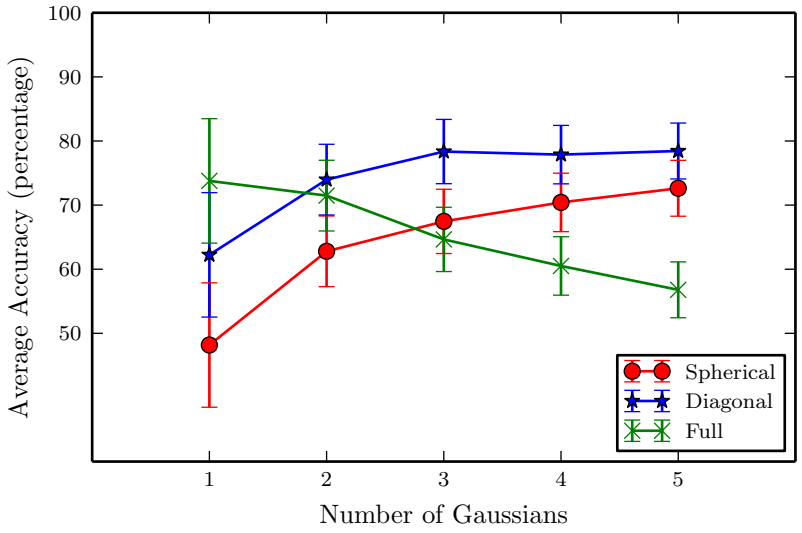


Figure 6.2: Comparison of GMM classifier accuracies for various configurations, using frame based labels and feature vectors reduced to 16 features.

The results are similar to previous experiments with the GMM classifier in Section 5.2.2, with the full covariance matrix performing good at first, and decreasing in accuracy as the number of Gaussians is improved. The diagonal covariance matrix scores the best, followed by the spherical covariance matrix. For classification a diagonal covariance matrix using 3 Gaussians per class is selected.

Table 6.2 shows the result of the classifier. The GMM classifier achieves an average accuracy of 78.35% and a standard accuracy of 16.69. A comparison between the results is made in Section 6.3.

6.1.3 Further possibilities

This feature extraction method can be further refined by optimising the size of the feature vector. This will mean a trade-off between classification speed and accuracy.

Table 6.2: Confusion matrix for a GMM classifier using 4 mixtures and frame-based labels, feature vectors reduced to 16 features. The average accuracy was 78.35% with a standard deviation of 16.69.

Actual Value \ Predicted Value	bakkie_lights	bakkie_radio	bakkie_start	big_crane	big_radio	cherry_picker	compressor	diesel_filter	kat7_meesdam	meerkat_compressor	meerkat_compressor_kat7	meysdam_gap	reefer	silence	vw_ignition	vw_indicators	welder	welder_spark
bakkie_lights	**88.1**		2.5			0.3		0.8		3.3				4.7	0.3			
bakkie_radio		**99.0**	1.0															
bakkie_start	0.3	0.3	**62.3**			24.5		1.9					2.6	4.2	1.3	2.6		
big_crane				**100.0**														
big_radio					**98.8**									1.2				
cherry_picker			26.0			**56.0**								18.0				
compressor							**72.2**				7.8	5.6		14.4				
diesel_filter			19.0			4.0		**48.0**			1.0		4.0	16.0	4.0	4.0		
kat7_meesdam									**99.8**		0.2							
meerkat_compressor	0.7		0.5							**89.3**				9.5				
meerkat_compressor_kat7							0.0		0.4		**64.6**	0.0		34.6	0.0	0.3		
meysdam_gap							0.3				5.7	**88.9**	0.2	4.5	0.5			
reefer			4.0					2.0					**54.0**	25.3	8.0		6.7	
silence	0.7		0.3	0.2	0.0	0.9	0.1	0.1		2.7	0.8	0.1	0.3	**91.7**	0.6	0.8	0.3	0.4
vw_ignition			1.4					3.2					5.5	14.1	**75.5**		0.5	
vw_indicators			11.1											25.6		**63.3**		
welder			0.9										5.0	11.4	1.4		**81.4**	
welder_spark														21.5			0.8	**77.7**

6.2 Classification using delta frames

A feature extraction method using delta frames is investigated. The 63-point feature vector used in Section 5.2 is extended by appending the delta to the next frame to it. This results in a feature vector with 126 dimensions. This method of feature extraction can also be extended by appending the acceleration coefficients to the next frames.

6.2.1 KNN classifier

Optimization of the KNN classifier is performed. Figure 6.3 shows the accuracy for different values of k. There is not much variation between different values, with all values obtaining close to 85% accuracy. A value of 1 is chosen for k.

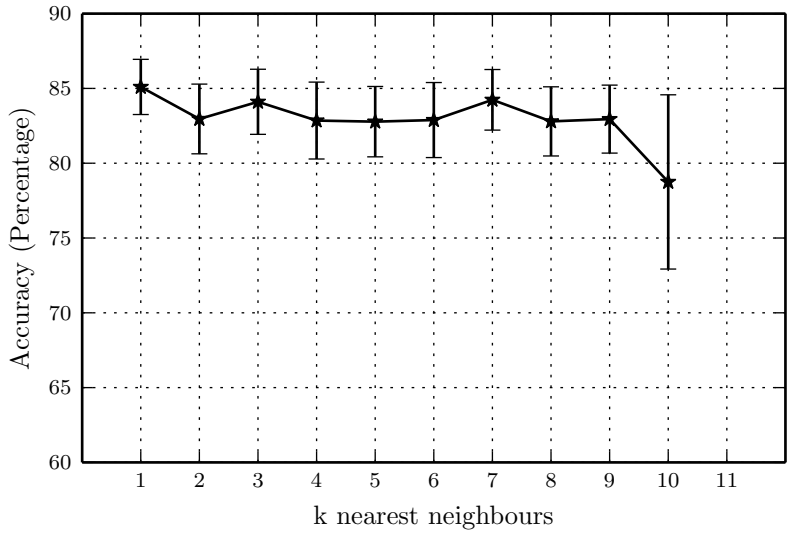


Figure 6.3: Comparison of KNN classifier accuracies for various values of k, using delta frames

Table 6.3 shows the confusion matrix for the KNN classifier. The classifier achieved an average accuracy of 86.05% with a standard deviation of 13.23.

Table 6.3: Confusion matrix for a KNN classifier using k=1 and frame-based labels with delta frames. Average accuracy is 86.05%.

Actual Value \ Predicted Value	bakkie_lights	bakkie_radio	bakkie_start	big_crane	big_radio	cherry_picker	compressor	diesel_filter	kat7_meesdam	meerkat_compressor	meerkat_compressor_kat7	meysdam_gap	reefer	silence	vw_ignition	vw_indicators	welder	welder_spark
bakkie_lights	**81.9**					0.5								17.6				
bakkie_radio		**99.5**								0.5								
bakkie_start			**75.9**			0.6		3.7					0.6	15.4	1.2	1.9	0.6	
big_crane				**100.0**														
big_radio				0.2	**99.5**					0.2				0.2				
cherry_picker						**78.8**							1.9	19.2				
compressor							**91.5**							8.5				
diesel_filter								**86.7**						11.1		2.2		
kat7_meesdam									**100.0**									
meerkat_compressor		0.9								**78.3**				20.7				
meerkat_compressor_kat7											**95.5**	0.2		4.2		0.0		
meysdam_gap							0.0				0.3	**99.7**						
reefer			2.9			1.4		1.4					**81.2**	11.6	1.4			
silence	0.3		0.3		0.0	0.1	0.1	0.0		0.5	1.0		0.2	**96.7**	0.3	0.3	0.2	0.1
vw_ignition			4.3					0.9					2.6	28.7	**63.5**			
vw_indicators			4.7					2.3					2.3	34.9		**55.8**		
welder			2.3										2.3	20.2	4.7		**70.5**	
welder_spark													1.5	4.6				**93.8**

6.2.2 GMM classifier

The GMM classifier is used to classify the data with delta frames. GMM models are trained for different covariance matrix types and mixture numbers. The tuning data is classified using these models and compared. The classification results are shown in Figure 6.4. The behaviour of the diagonal and spherical

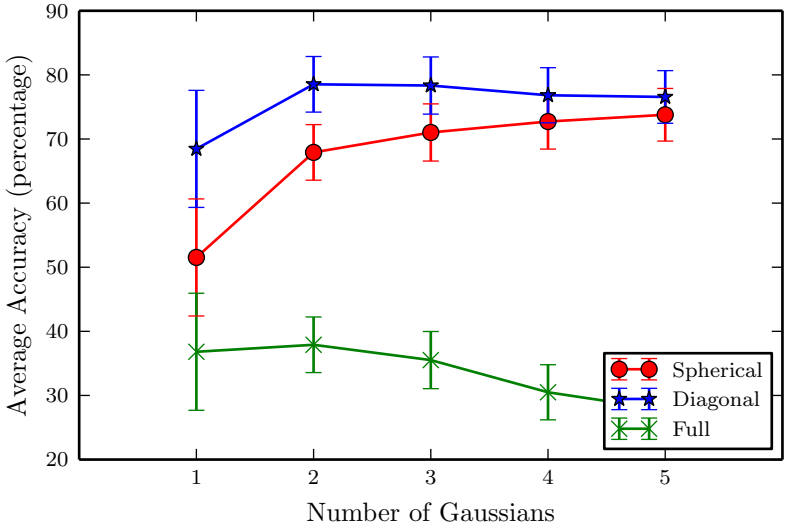


Figure 6.4: Comparison of GMM classifier accuracies for different configurations, using frame based labels with delta frames.

covariance matrices are similar to previous experiments, starting at a lower accuracy and quickly improving as the number of Gaussians is increased. The Full covariance also follows a similar pattern as previous experiments, but has a much lower accuracy overall.

The testing data is classified using a classifier with a diagonal covariance matrix with 3 Gaussians. The results are shown in Table 6.4. The classifier achieved an average accuracy of 78.34%, and a standard deviation of 18.28.

Table 6.4: Confusion matrix for a GMM classifier using frame-based labels with delta frames. The average accuracy is 78.34%.

Actual Value \ Predicted Value	bakkie_lights	bakkie_radio	bakkie_start	big_crane	big_radio	cherry_picker	compressor	diesel_filter	kat7_meesdam	meerkat_compressor	meerkat_compressor_kat7	meysdam_gap	reefer	silence	vw_ignition	vw_indicators	welder	welder_spark
bakkie_lights	**88.6**		0.3			0.8		0.3	0.3	1.9	0.3			7.5				
bakkie_radio		**98.2**	0.8	0.3				0.5						0.3				
bakkie_start			**72.9**			8.7		1.3				0.3	1.9	11.0	2.6		1.3	
big_crane				**99.9**				0.1										
big_radio					**99.4**									0.6				
cherry_picker			20.0			**61.0**								19.0				
compressor							**67.8**				17.8	11.1		3.3				
diesel_filter			28.0					**46.0**	9.0		2.0		7.0	8.0				
kat7_meesdam									**100.0**									
meerkat_compressor			0.5					0.2		**90.5**				8.8				
meerkat_compressor_kat7			0.0								**74.2**	0.5	0.0	25.2		0.0		
meysdam_gap			0.0								5.0	**94.6**	0.1	0.1	0.1		0.1	
reefer			9.3					1.3					**69.3**	19.3	0.7			
silence	0.5		0.6	0.6	0.1	0.7	0.1	0.1	0.0	4.4	0.9	0.1	0.6	**90.4**	0.4	0.3	0.3	0.0
vw_ignition			3.2					0.9				4.5	3.2	20.9	**62.3**		5.0	
vw_indicators			16.7					1.1			2.2			42.2		**37.8**		
welder			0.9								1.4	1.8		21.4	2.7		**71.8**	
welder_spark														14.6				**85.4**

6.3 Comparison between methods

The results from all the used classifiers are combined and compared. Table 6.5 shows the average accuracy when classifying the testing data on all the different classifiers.

Table 6.5: 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

The least accurate feature extraction method was the capture based labels. This method only considered the frame with the highest power, so much less data was available for training purposes.

The most accurate classifier was the KNN classifier using frame based labels, which was discussed in Section 5.2.1. The method also has the smallest standard deviation, indicating that most of the classes had high accuracies.

For each of the feature extraction methods the GMM scores worse than the KNN. The biggest difference in accuracy is between the feature vectors with delta frames.

Changing from reduced feature vectors to delta frames yields an improvement for the KNN classifier, but yields almost no change for the GMM classifier.

6.4 Conclusion

This section investigated two different feature extraction methods. The first method reduced the feature vector down 16 samples. Using this method the accuracy of the KNN and GMM classifiers were lower than the previous experiment.

The second method appended a delta to the next frame to each feature vector. This method also had a lower accuracy than previous methods, although not as substantial as the reduced vector.

Overall the KNN classifier discussed in Section 5.2.1 is still the most accurate.

Chapter 7

Post-processing

In this section some post-processing methods are considered. The two methods presented here attempt to improve the labelling produced by the classifier, and to assess the robustness of a classifier against unknown data.

The classifiers used in the previous section can classify a data point as belonging to a certain class, or as silence. This classification can be reduced to a binary classification by only considering if a data point is RFI or silence. This allows us to use tools such as the F1 score (see Section 3.6.3) to assess the accuracy of the classification.

7.1 Classification of the capture as a time series

One way in which classification may be improved is to consider the temporal relation between the labels assigned to frames. To do this all the frames in a single capture are first classified independently using one of the previous classifiers. The predicted labels are then considered as a temporal sequence. A median filter is applied to the sequence of predicted labels in an attempt to improve the accuracy. The filter will have a smoothing effect on the data, removing isolated frames whose assigned label differs from the surrounding frames.

The filter considers a window of consecutive classifier labels. This window is generally centred on the frame of current interest. The filter selects the most common value from within the window, and assigns this value to the

frame at the centre of the window. Subsequently the window advances by one frame, and the process is repeated. The number of frames in the window must be determined empirically. If the value is too small, the filter is sensitive to high-frequency changes. If the value is too high the filtering effect will be too strong.

The GMM classifier describer in Section 5.2.2 is used. A capture (consisting of multiple frames) is classified using the classifier. The classifications are converted to a binary classification by only considering if a data point is classified as RFI or not. The F1 score is calculated. The classifications are then processed by median filters whose windows extend 1,2,3 and 4 frames either side of the frame of interest, and the new F1 score is calculated. The two scores are compared to asses any improvements as a result of the filter. This process is repeated for all captures, and every data fold.

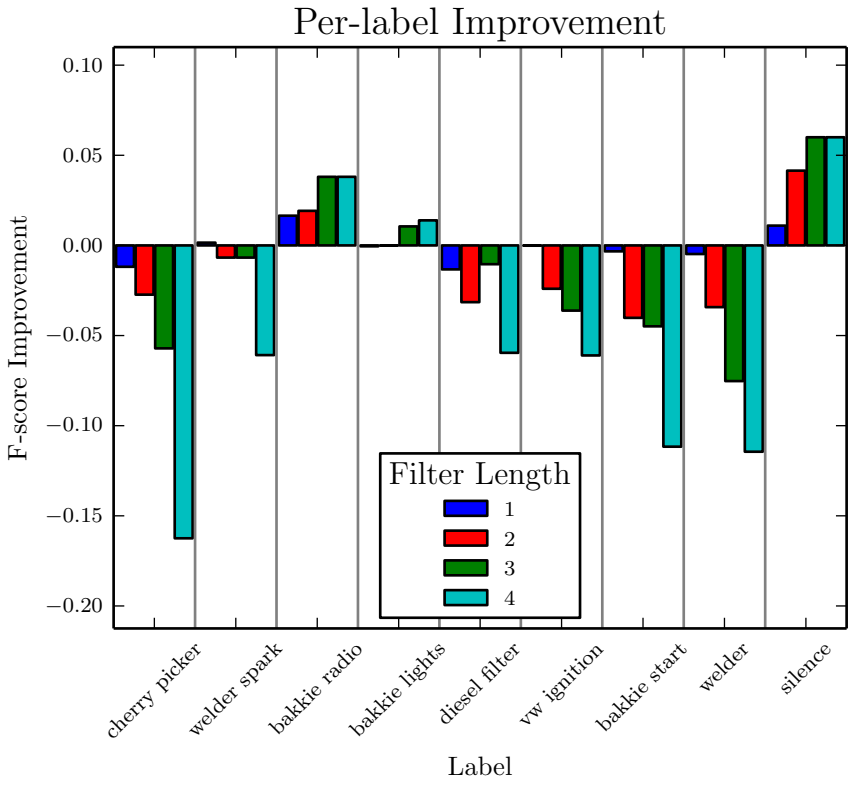


Figure 7.1: Improvements for various lengths of median filter, divided by class.

Figure 7.1 shows the difference in F1 scores for each label. The four columns per label are the different lengths for the filter. The figure shows that the

overall classification accuracy is decreased by applying the filter, and worsens as the length is increased.

7.2 Classification performance for unseen types of RFI

A classifier running in the field will encounter data that does not necessarily belong to any of the RFI classes encountered during training. A good classifier should be able to detect the new data and to flag it accordingly. To evaluate how our classifier would handle this situation, we omit one class of data training. This class is then used during testing. This process is repeated for all the classes and for all folds of the data. This will give an accurate representation of the capabilities of the classifier.

The GMM classifier was modified slightly and used for this experiment. For each input frame the classifier returns the probability that the frame belongs to a class. If the probability for any class is above a threshold the frame is labelled as the relevant class. However, if the probability is below the threshold the frame is labelled as a missing class. This allows the classifier to detect when new data is presented to it. The GMM is used for this, as it returns a probability. A KNN does not return a probability, only a label. The GMM used 3 Gaussians and a diagonal covariance matrix.

To determine the optimum threshold of the classifier the result of the classification is first converted to a binary result. The positive result is defined as unknown RFI data, and the negative as known RFI data. A good classifier should be able to detect the unknown classes (returning many true positives) but should also detect known RFI (returning many true negatives). These values in conjunction with the false positive and false negative rates can be used to asses the ability of the classifier to identify new sources.

The threshold can now be varied to determine the optimum point. The results of different threshold is plotted on an DET curve.

Figure 7.2 shows the DET curve for the GMM.

In order to select the optimal value for the threshold, the frequency of RFI events at the telescope site has to be considered. As this data is not available to us, no attempt is made to estimate an optimal value.

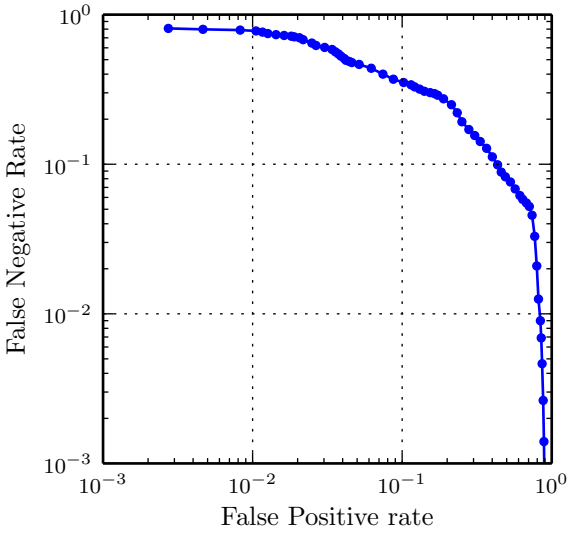


Figure 7.2: DET curve of classification with a unknown source

7.3 Conclusion

Some post-processing techniques were investigated in this chapter. A filter was applied to the post-classification labels. This filter attempted to sacrifice accuracy in order to improve the false-negative rate, but was largely unsuccessful.

Another technique that was investigated was classification using a missing class. A class was omitted when training the classifier models, and the ability of the classifier to detect the missing class was investigated. The threshold for the classifier has to be set depending on the site characteristics.

Chapter 8

Summary, conclusion and further work

The work described in this thesis involved automatic identification and classification of RFI in a radio telescope environment. RFI presents a large challenge to telescope observations, as even low-powered interference can overpower the astronomical signals. It is generally not possible to remove such interference from the astronomical signals. However, by detecting the presence and type of interference, it can be identified and removed from site. The detection of RFI will also allow potentially corrupt observations to be omitted from astronomical experiments.

This thesis has investigated the use of machine learning algorithms to identify RFI signals.

Data from different sources was captured from the SKA site. The data was captured in the time domain, using the real time analyser (RTA). The signals were captured close to the source, using an LPDA antenna. The RTA can only capture time signals of up to $8\mu s$ in length. The data was labelled, and different feature extraction and classification methods were implemented and tested.

The initial feature extraction technique considered only a single feature vector per capture. KNN and GMM classifiers were applied to the data, and were able to distinguish between the 15 classes with accuracies of 70.80% and 65.56%, respectively.

Performance was improved using a thresholding method to select multiple feature vectors per capture. This made more data available to the classifiers

for training and evaluation. It also allowed them to distinguish between silence (absence of RFI) and RFI. Using the improved feature extraction technique both KNN and GMM classifiers exhibited improved accuracies of 93.2% and 87.34%, respectively.

Variations of the feature extraction methods were also considered. One method investigated the effect of reducing the dimensionality of the feature vector from 63 to only 15 features. Another method extended the feature vector by adding delta features. Both of these methods performed adequately, but did not outperform the methods previously used.

An attempt was also made to reduce the false negative rate of the classifier by applying a post-classification filter. Finally the versatility of the GMM classifier was tested by omitting classes, hereby emulating the occurrence of previous unseen RFI classes. It was found that the GMM classifier can detect unknown sources, but must be carefully tuned to the characteristics of the site.

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.

8.1 Recommendations and Future Work

In future work other classifiers might be considered. Using hidden Markov models will allow the temporal properties in the data to be captured, possibly allowing more accurate classification. This will however require more and specifically longer captured data than is currently available.

In line with this, the data sampling capabilities should be improved. The current sampling hardware is able to capture only a very short segments of the signal, albeit a very high sampling rate. A longer capture may reveal temporal relations that are not visible using the current method, and that can be exploited for classification. The longer capture can even be done at the cost of sampling rate. Most of the RFI signals contained energy in the range from 50 to 400MHz. This means the 1.8GHz sampling rate is excessive and possibly simpler recording equipment could be used.

The data used for classification was captured close to the source. In most cases the signal was clearly visible in the spectrogram. Capturing data from further away, or using a less directional antenna will allow more rigorous testing

of classifiers. However, capturing and correctly labelling the data will be more difficult.

List of References

[1] A. R. Thompson, J. M. Moran, and G. W. Swenson Jr, *Interferometry and synthesis in radio astronomy.* John Wiley & Sons, 2008.

[2] K. G. Jansky, "Electrical disturbances apparently of extraterrestrial origin," *Radio Engineers, Proceedings of the Institute of*, vol. 21, no. 10, pp. 1387–1398, 1933.

[3] Unknown. (2014, Feb.) Grote reber. [Online]. Available: http://www.nrao.edu/whatisra/hist_reber.shtml

[4] J. Binney, *Galactic astronomy.* Princeton University Press, 1998.

[5] Unknown. (2014, Feb.) The hydrogen 21-cm line. [Online]. Available: http://hyperphysics.phy-astr.gsu.edu/hbase/quantum/h21.html

[6] H. I. Ewen and E. Purcell, "Observation of a line in the galactic radio spectrum," *Nature*, vol. 168, no. 4270, pp. 356–358, 1951.

[7] N. Kanekar and F. Briggs, "21-cm absorption studies with the Square Kilometer Array," *New Astronomy Reviews*, vol. 48, no. 11, pp. 1259–1270, 2004.

[8] O. Maron, J. Kijak, M. Kramer, and R. Wielebinski, "Pulsar spectra of radio emission," *arXiv preprint astro-ph/0010233*, 2000.

[9] J. D. Kraus, *Radio astronomy.* McGraw-Hill, 1966.

[10] K. Rohlfs and T. Wilson, *Tools of radio astronomy.* Springer Science & Business Media, 2013.

[11] "KAT-7 (seven-dish MeerKAT precursor array)," http://www.ska.ac.za/meerkat/kat7.php, Accessed: June 2014.

[12] "MeerKAT," http://www.ska.ac.za/meerkat/, Accessed: June 2014.

[13] A. S. Ostrovsky, G. Martínez-Niconoff, P. Martinez-Vara, and M. A. Olvera-Santamaría, "The van Cittert-Zernike theorem for electromagnetic fields," *Optics express*, vol. 17, no. 3, pp. 1746–1752, 2009.

[14] G. Swenson Jr and N. Mathur, "The interferometer in radio astronomy," *Proceedings of the IEEE*, vol. 56, no. 12, pp. 2114–2130, 1968.

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

[16] 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.

[17] 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.

[18] A. Leshem, A.-J. van der Veen, and A.-J. Boonstra, "Multichannel interference mitigation techniques in radio astronomy," *The Astrophysical Journal Supplement Series*, vol. 131, no. 1, p. 355, 2000.

[19] F. Briggs, J. Bell, and M. Kesteven, "Removing radio interference from contaminated astronomical spectra using an independent reference signal and closure relations," *The Astronomical Journal*, vol. 120, no. 6, p. 3351, 2000.

[20] 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.

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

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

[23] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[24] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," DTIC Document, Tech. Rep., 1997.

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

[26] Unknown. (2015, May) Rhode and schwarz hl033 antenna. [Online]. Available: http://www.rohde-schwarz.com