Chapter 1

Signals in Radio Astronomy

Rajaram Nityananda

1.1 Introduction

The record of the electric field $E(t)$, received at a point on earth from a source of radio waves can be called a "signal", so long as we do not take this to imply intelligence at the transmitting end. Emanating as it does from a large object with many independently radiating parts, at different distances from our point, and containing many frequencies, this signal is naturally random in character. In fact, this randomness is of an extreme form. All measured statistical properties are consistent with a model in which different frequencies have completely unrelated phases, and each of these phases can vary randomly from 0 to 2π. A sketch of such a signal is given in Fig. 1.1. The strength (squared amplitude or power) of the different frequencies ω has a systematic variation which we call the "power spectrum" $S(\omega)$. This chapter covers the basic properties of such signals, which go by the name of "time-stationary gaussian noise". Both the signal from the source of interest, as well as the noise added to this cosmic signal by the radio telescope recievers can be described as time-stationary gaussian noise. The word noise of course refers to the random character. "Noise" also evokes unwanted disturbance, but this of course does not apply to the signal from the source (but does apply to what our receivers unavoidably add to it). The whole goal of radio astronomy is to receive, process, and interpret these cosmic signals, (which were, ironically enough, first discovered as a "noise" which affected trans-atlantic radio communication). "Time–Stationary" means that the signal in one time interval is statistically indistinguishable from that in another equal duration but time shifted interval. Like all probabilistic statements, this can never be precisely checked but its validity can be made more probable (circularity intended!) by repeated experiments. For example, we could look at the probability distribution of the signal amplitude. An experimenter could take a stretch of the signal say, from times 0 to T, select N equally spaced values $E(t_i), i$ going from 1 to N, and make a histogram of them. The property of time stationarity says that this histogram will turn out to be (statistically) the same — with calculable errors decreasing as N increases! — if one had chosen instead the stretch from t to $t+T$, for any t. The second important characteristic property of our random phase superposition of many frequencies is that this histogram will tend to a gaussian, with zero mean as N tends to infinity.

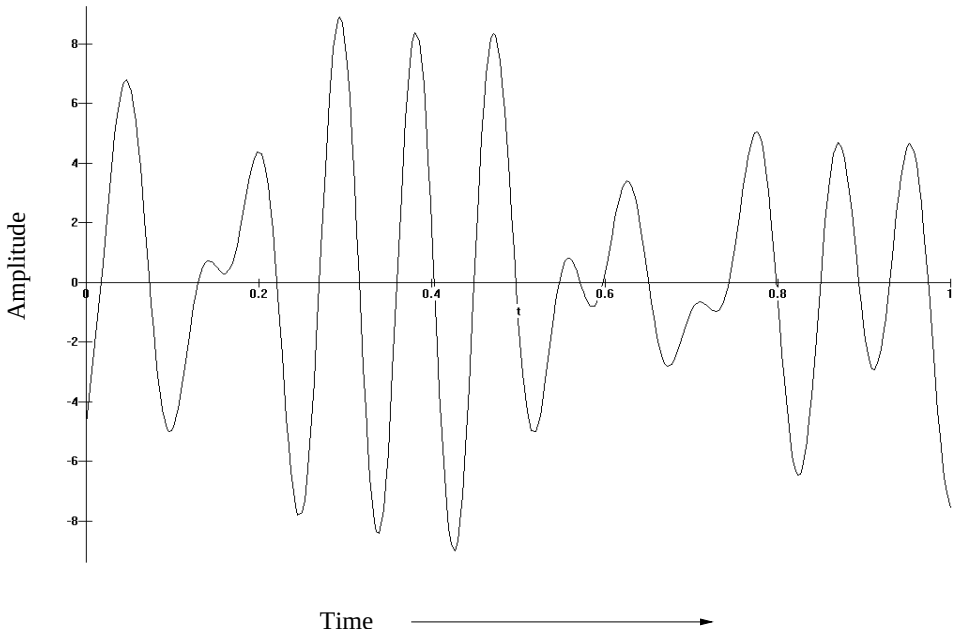


Figure 1.1: A signal made by superposition of many frequencies with random phases

1.2 Properties of the Gaussian

The general statement of gaussianity is that we look at the **joint** distribution of N amplitudes $x_1 = E(t_1), x_2 = E(t_2), \ldots$ etc. This is of the form

$$P(x_1 \ldots x_k) = const \times \exp\left(-Q(x_1, x_2, \ldots x_k)\right)$$

Q is a quadratic expression which clearly has to increase to $+\infty$ in any direction in the k dimensional space of the x's. For just one amplitude,

$$P(x_1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x_1^2/2\sigma^2}$$

does the job and has one parameter, the "Variance"σ, the mean being zero. This variance is a measure of the power in the signal. For two variables, x_1 and x_2, the general mathematical form is the "bivariate gaussian"

$$P(x_1, x_2) = const \times \exp\left(-\frac{1}{2}(a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2)\right)$$

.

Such a distribution can be visualised as a cloud of points in $x_1 - x_2$ space, whose density is constant along ellipses Q =constant (see Fig. 1.2).

The following basic properties are worth noting (and even checking!).

1. We need a_{11}, a_{22}, and $a_{11}a_{22} - a_{12}^2$ all > 0 to have ellipses for the contours of constant P (hyperbolas or parabolas would be a disaster, since P would not fall off at infinity).

2. The constant in front is

$$(1/2\pi) \times \sqrt{det \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix}}$$

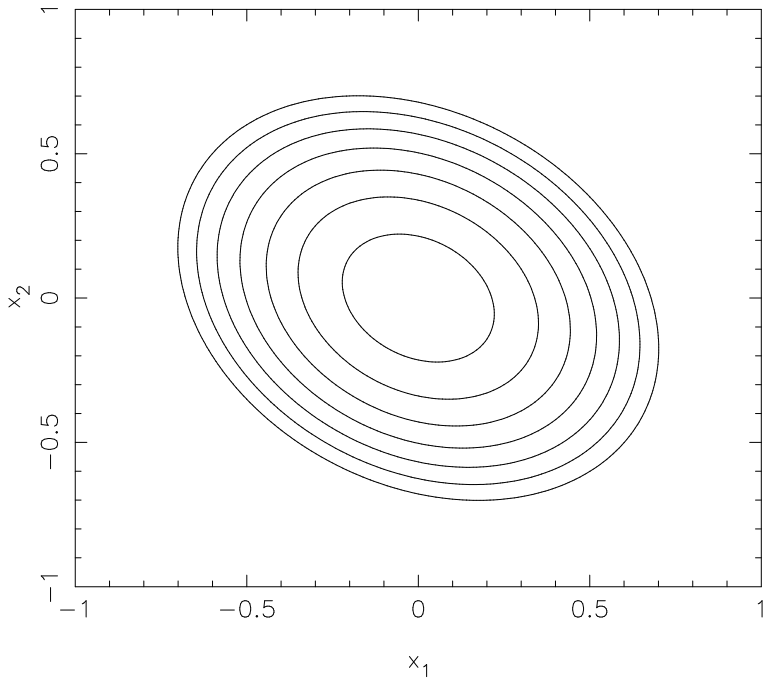


Figure 1.2: Contour lines of a bivariate gaussian distribution

3. The average values of x_1^2, x_2^2 and $x_1 x_2$, when arranged as a matrix (the so called covariance matrix) are the **inverse** of the matrix of a's. For example,

$$\langle x_1^2 \rangle = a_{22}/detA$$

$$\langle x_1 x_2 \rangle = a_{12}/detA$$

etc.

4. By time stationarity,

$$< x_1^2 >=< x_2^2 >= \sigma^2$$

$$< x_1^2 >=< x_2^2 >= \sigma^2$$

The extra information about the correlation between x_1 and x_2 is contained in $< x_1 x_2 >$, i.e. in a_{12} which (again by stationarity) can only be a function of the time separation $\tau = t_1 - t_2$. We can hence write $< E(t)E(t+\tau) >= C(\tau)$ independent of t. $C(\tau)$ is called the autocorrelation function. From (1) above, $C^2(\tau) \leq \sigma^2$. This suggests that the quantity $r(\tau) = C(\tau)/\sigma^2$ is worth defining, as a dimensionless correlation coefficient, normalised so that $r(0) = 1$. The generalisation of all these results for a k variable gaussian is given in the Section 1.8

1.3 The Wiener-Khinchin Theorem

So far, we have only *asserted* that the sum of waves with random phases generates a time-stationary gaussian signal. We now have to check this. It is convenient to start with

a signal going from 0 to T, and only later take the limit $T \to \infty$. The usual theory of Fourier series tells us that we can write

$$\begin{aligned} E(t) &\equiv \sum a_n \cos\omega_n t + b_n \sin\omega_n t \\ &\equiv \sum r_n \cos(\omega_n t + \varphi_n) \end{aligned}$$

where,

$$\omega_n = \frac{2\pi}{T},\ r_n = \sqrt{a_n x^2 + b_n^2}, and\ \tan\varphi_n = -b_n/a_n$$

Notice that the frequencies come in multiples of the "fundamental" $2\pi/T$ which is very small since T is large, and hence they form a closely spaced set. We can now compute the autocorrelation

$$C(\tau) = \langle E(t)E(t+\tau)\rangle = \langle \sum_n r_n \cos(\omega_n t + \varphi_n) \sum_m r_m \cos(\omega_m(t+\tau) + \varphi_m)\rangle$$

The averaging on the right hand side has to be carried out by letting each of the phases φ_k vary independently from 0 to 2π. When we do this, only terms with $m = n$ can survive, and we get

$$C(\tau) = \sum \frac{1}{2} r_n^2 \cos\omega_n\tau$$

.

Putting τ equal to zero, we get the variance

$$C(0) = \langle E(t)^2\rangle = \sum \frac{1}{2} r_n^2$$

We note that the autocorrelation is independent of t and hence we have checked time stationarity, at least for this statistical property. We now have to face the limit $T \to \infty$. The number of frequencies in a given range $\Delta\omega$ blows up as

$$\frac{\Delta\omega}{(2\pi/T)} = \frac{T\Delta\omega}{2\pi}.$$

Clearly, the r_n^2 have to scale inversely with T if statistical qualities like $C(\tau)$ are to have a well defined $T \to \infty$ behaviour. Further, since the number of r_n's even in a small interval $\Delta\omega$ blows up, what is important is their combined effect rather than the behaviour of any individual one. All this motivates the definition.

$$\sum_{\omega < \omega_n < \omega + \Delta\omega} \frac{r_n^2}{2} = 2S(\omega)\Delta\omega$$

as $T \to \infty$. Physically, $2S(\omega)\Delta\omega$ is the contribution to the variance $\langle E^2(t)\rangle$ from the interval ω to $\omega + \Delta\omega$. Hence the term "power spectrum" for $S(\omega)$. Our basic result for the autocorrelation now reads

$$C(\tau) = \int_0^\infty 2S(\omega)\cos\omega\tau d\omega = \int_{-\infty}^{+\infty} S(\omega) e^{-i\omega\tau} d\omega$$

if we define $S(-\omega) = S(\omega)$.

This is the "Wiener–Khinchin theorem" stating that the autocorrelation function is the Fourier transform of the power spectrum. It can also be written with the frequency measured in cycles (rather than radians) per second and denoted by ν.

$$C(\tau) = \int_0^\infty 2P(\nu)\cos(2\pi\nu\tau)d\nu = \int_{-\infty}^{+\infty} P(\nu)e^{-2\pi i\nu\tau}d\nu$$

and as before, $P(-\nu) = P(\nu)$.

In this particular case of the autocorrelation, we did not use independence of the φ 's. Thus the theorem is valid even for a non-gaussian random process. (for which different φ 's are not independent). Notice also that we could have averaged over t instead of over all the φ's and we would have obtained the same result, viz. that contributions are nonzero only when we multiply a given frequency with itself. One could even argue that the operation of integrating over the φ's is summing over a fictitious collection (i.e "ensemble") of signals, while integrating over t and dividing by T is closer to what we do in practice. The idea that the ensemble average can be realised by the more practical time average is called "ergodicity" and like everything else here, needs better proof than we have given it. A rigorous treatment would in fact start by worrying about existence of a well-defined $T \to \infty$ limit for all statistical quantities, not just the autocorrelation. This is called "proving the existence of the random process".

The autocorrelation $C(\tau)$ and the power spectrum $S(\omega)$ could in principle be measured in two different kinds of experiments. In the time domain, one could record samples of the voltage and calculate averages of lagged products to get C. In the frequency domain one would pass the signal through a filter admitting a narrow band of frequencies around ω, and measure the average power that gets through.

A simple but instructive application of the Wiener Khinchin theorem is to a power spectrum which is constant ("flat band") between $\nu_0 - B/2$ and $\nu_0 + B/2$. A simple calculation shows that

$$C(\tau) \;=\; 2KB\left(\cos(2\pi\nu_0\tau)\right)\left(\frac{\sin(\pi B\tau)}{\pi B\tau}\right)$$

The first factor $2KB$ is the value at $\tau = 0$, hence the total power/variance to radio astronomers/statisticians. The second factor is an oscillation at the centre frequency. This is easily understood. If the bandwidth B is very small compared to ν_0, the third factor would be close to unity for values of τ extending over say $1/4B$, which is still many cycles of the centre frequency. This approaches the limiting case of a single sinusoidal wave, whose autocorrelation is sinusoidal. The third sinc function factor describes "bandwidth decorrelation[1]", which occurs when τ becomes comparable to or larger than $1/B$.

Another important case, in some ways opposite to the preceding one, occurs when $\nu_0 = B/2$, so that the band extends from 0 to B. This is a so-called "baseband". In this case, the autocorrelation is proportional to a sinc function of $2\pi B\tau$. Now, the correlation between a pair of voltages measured at an interval of $1/2B$ or any multiple (except zero!) thereof is zero, a special property of our flat band. In this case, we see very clearly that a set of samples measured at this interval of $1/2B$, the so-called "Nyquist sampling interval", would actually be statistically independent since correlations between any pair vanish (this would be clearer after going through Section 1.8). Clearly, this is the minimum number of measurements which would have to be made to reproduce the signal, since if we missed one of them the others would give us no clue about it. As we will now see, it is also the maximum number for this bandwidth!

1.4 The Sampling Theorem

This more general property of a band-limited signal (one with zero power outside a bandwidth B) goes by the name of the "Shannon Sampling Theorem". It states that a set of

[1] also called "fringe washing" in Chapter 4

samples separated by $1/2B$ is sufficient to reconstruct the signal. One can obtain a preliminary feel for the theorem by counting Fourier coefficients. The number of parameters defining our signal is twice the number of frequencies, (since we have an a and a b, or an r and a φ, for each ω_n). Hence the number of real values needed to specify our signal for a time T is

$$2 \times \frac{\Delta\omega T}{2\pi} = 2\left(\frac{\Delta\omega}{2\pi}\right)T = 2BT$$

This rate at which new real numbers need to be measured to keep pace with the signal is $2B$. The so called "Nyquist sampling interval" is therefore $(2B)^{-1}$. A real proof (sketched in Section 1.8) would give a reconstruction of the signal from these samples!

In words, the Shannon criterion is two samples per cycle of the maximum frequency *difference* present. The usual intuition is that the centre frequency ν_0 does not play a role in these considerations. It just acts a kind of rapid modulation which is completely known and one does not have to sample variations at this frequency. This intuition is consistent with radio engineers/astronomers fundamental right to move the centre frequency around by heterodyning[2] with local (or even imported[3]) oscillators, but a more careful examination shows that the centre frequency should satisfy $\nu_0 = (n + \frac{1}{2})B$ for the sampling at a rate $2B$ to work.

1.5 The Central Limit and Pairing Theorems

We now come to the statistics of $E(t)$. For example, we already know that $\langle E^2(t)\rangle = \sum r_n^2/2$. How about $\langle E^3(t)\rangle$? Quite easy to check that it is zero because

$$\langle r_l r_m r_n \cos(\omega_m t + \varphi_m)\cos(\omega_n t + \varphi_n)\cos(\omega_l t + \varphi_l)\rangle = 0$$

when we let the φ's each vary independently over the full circle 0 to 2π. This is true whether l, m, n are distinct or not. But coming to even powers like $\langle E^4(t)\rangle$, something interesting happens. When we integrate a product like $r_l r_m r_n r_p \cos(\omega_m t + \varphi_m)\cos(\omega_n t + \varphi_n)\cos(\omega_\ell t + \varphi_l)\cos(\omega_p t + \varphi_p)$ over all the four φ's we can get non-zero answers, provided the φ's occur in pairs, i.e., if $l = m$ and $n = p$, then we encounter $\cos^2\varphi_l \times \cos^2\varphi_n$ which has a non-zero average. (We saw a particular case of this when we calculated $\langle E(t)E(t+\tau)\rangle$ and only r_m^2 type terms survived).

Because of the random and independent phases of the large number of different frequencies, we can now state the "pairing theorem".

$$\langle E(t_1)E(T_2)\ldots E(t_{2k})\rangle = \sum_{\text{pairs}} \langle E(t_1)E(t_2)\rangle \ldots \langle E(t_{2k-1})E(t_{2k})\rangle$$

As discussed in Section 1.8, this pairing theorem proves that the statistics is gaussian. (A careful treatment shows that only the $r_m^2 r_n^2$ terms are equal on the two sides- we have not quite got the r_m^4 terms right, but there are many more (of the order of N times more) of the former type and they dominate as $T \to \infty$ and the numbers of sines and cosines we are adding is very large). This result — that the sum of a large number of small, finite variance, independent terms has a gaussian distribution — is a particular case of the "central limit theorem". We only need the particular case where these terms are cosines with random phases.

[2] see Chapter 3

[3] aaaaagggh! beware of weak puns. (eds.)

1.6 Quasimonochromatic and Complex Signals

For a strictly monochromatic signal, electrical engineers have known for a long time that it is very convenient to use a *complex* voltage $V(t) = E_0 \exp(i(\omega t + \varphi))$ whose real part gives the actual signal $E_r(t) = E_0 \cos(\omega t + \varphi)$. One need not think of the imaginary part as a pure fiction since it can be obtained from the given signal by a phase shift of $\pi/2$, viz. as $E_i(t) = E_0 \cos(\omega t + \phi - \pi/2)$. In practice, since one invariably deals with signals at an intermediate frequency derived by beating with a local oscillator, both the real and imaginary parts are available by using two such oscillators $\pi/2$ out of phase. Squaring and adding the real and imaginary parts give $E_r^2(t) + E_i^2(t) = V(t)^* V(t) = E_0^2$ which is the power averaged over a cycle. This is actually closer to what is practically measured than the instantaneous power, which fluctuates at a frequency 2ω.

These ideas go through even when we have a range of frequencies present, by simply imagining the complex voltages corresponding to each of the monochromatic components to be added. In mathematical terms, this operation of deriving $E_i(t)$ from $E_r(t)$ goes by the name of the "Hilbert Transform", and the time domain equivalent is described in Section 1.8 But the physical interpretation is easiest when the different components occupy a range $\Delta\omega$ - the so called "bandwidth" - which is small compared to the "centre frequency" ω_0. Such a signal is called "quasimonochromatic", and can be represented as below

$$E_q(t) = Re\ \exp(i\omega_0 t) \sum_{-\Delta\omega/2 < \omega_1 < \Delta\omega/2} E(\omega_1)\ \exp(i\omega_1 t + i\varphi(\omega_1))$$

In this expression, ω_1 is a frequency offset from the chosen centre ω_0, so that $E(\omega_1)$ actually represents the amplitude at a frequency $\omega_0 + \omega_1$, and $\varphi(\omega_1)$ the phase. We can now think of our quasimonochromatic signal as a rapidly varying phasor at the centre frequency ω_0, modulated by a complex voltage

$$V_m(t) = \sum_{-\Delta\omega/2 < \omega_1 < \Delta\omega/2} E(\omega_1)\ \exp(i\omega_1 t + i\varphi\omega_1)$$

This latter phasor varies much more slowly than $\exp(-i\omega_0 t)$. In fact, it takes a time $\Delta\omega^{-1}$ for $V_m(t)$ to vary significantly since the highest frequencies present are of order $\Delta\omega$. This time scale is much longer than the timescale ω^{-1} associated with the centre frequency. Writing $V_m(t)$ in the polar form as $R(t)\exp(i\alpha(t))$, our original real signal reads

$$E_q(t) = R(t)\cos(\omega_0 t + \alpha(t))$$

.

We can think of R and α as time dependent, slowly varying, amplitude and phase modulation of an otherwise (hence "quasi") monochromatic signal.

While the mathematics did not assume smallness of $\Delta\omega$, the physical interpretation does. If $R(t)$ changes significantly during a cycle, some of its values may not be attained as maxima and hence its square cannot be regarded as measuring average power. This is as it should be. No amount of algebra can uniquely extract *two* real functions $R(t)$ and $\alpha(t)$ from a single real signal without further conditions (and the condition imposed is explained in section 1.8).

But returning to the quasimonochromatic case, we can now think of $V_m(t)^* V_m(t)$ as the (slowly) time varying power in the signal. Likewise we can think of $\langle V_m^*(t) V_m(t+\tau)\rangle$ as the autocorrelation. (A little algebra checks that this is the same as the autocorrelation of the original real signal). One advantage in working with the complex signal is that the centre frequency cancels in any such product containing one voltage and one complex

conjugate voltage. We can therefore think of such products as referring to properties of the fluctuations of the signal amplitude and phase, and measure them even after heterodyning has changed the centre frequency.

1.7 Cross Correlations

We have so far thought of the signal as a function of time, after it enters the antenna. Let us now liberate ourselves from one dimension (time) and think of the electric field as existing in space and time, before it is collected by the antenna. In this view, one can obtain a delayed version of the signal by moving along the longitudinal direction (direction of the source). Thus, the frequency content is obtained by Fourier transforming a *longitudinal* spatial correlation. As explained in Chapter 2, the spatial correlations *transverse* to the direction of propagation carry information on the angular power spectrum of the signal, i.e. the energy as a function of direction in the sky. With hindsight, this can be viewed as a generalisation of the Wiener- Khinchin theorem to spatial correlations of a complex electric field which is the sum of waves propagating in many different directions. Historically, it arose quite independently (and about at the same time!) in the context of optical interference. This is the van Cittert-Zernike theorem of Chapter 2. Since one is now multiplying and averaging signals coming from different antennas, this is called a "cross correlation function". To get a non-vanishing average, one needs to multiply $E_1(x,t)$ by $E_2^*(y,t)$. The complex conjugate sign in one of the terms ensures that this kind of product looks at the phase *difference*. Writing out each signal as a sum with random phases, the terms which leave a non-zero average are the ones in which an $e^{i\varphi_n}$ in an E cancels a $e^{-i\varphi_n}$ in an E^*. An (ill-starred?) product of two complex E's with zero (or two!) complex conjugate signs would average to zero.

1.8 Mathematical details

This section gives some more mathematical details of topics mentioned in the main text of the chapter.

We first give the generalisation of the two variable gaussian to the joint distribution of k variables. Defining the covariance matrix $C_{ij} = \langle x_i x_j \rangle$, and $A = C^{-1}$, then we have

$$P(x_1 \dots x_k) = (2\pi)^{-k/2} (det\ A)^{1/2} \exp\left(-\frac{1}{2} x^T A x\right)$$

The quadratic function Q in the exponent has been written in matrix notation with T for transpose. In full, it is $Q = \sum_{ij} x_i a_{ij} x_j$. Notice that the only information we need for the statistics of the amplitudes at k different times is the autocorrelation function $C(\tau)$, evaluated at all time differences $t_i - t_j$. Formally this is stated as "the gaussian process is defined by its second order statistics".

What would be practically useful is an explicit formula for the average value of an arbitrary product $x_i x_j x_l \dots$ in terms of the second order statistics $\langle x_1 x_2 \rangle \langle x_3 x_7 \rangle \dots$ etc. The first step is to see that a product of an **odd** number of x's averages to **zero**. (The contributions from $x_1 \dots x_k$ & $-x_1 \dots - x_k$ cancel).

For the case of an even number of gaussian variables to be multiplied and averaged, there is a standard trick to evaluate an integral like $\int P(x_1 \dots x_k) x_3 x_7 \dots dx_1 \dots$. Define the Fourier transform of P,

$$G(k_1 \dots k_k) = \int \int P(x_1 \dots x_k) e^{-ik_1 x_1 \dots ik_k x_k} dx_1 \dots dx_k$$

It is a standard result, derived by the usual device of completing the square, that this Fourier transform is itself a gaussian function of the k's, given by

$$G(k_1, \ldots, k_k) = \exp\left(-\frac{1}{2}\sum_{ij} C_{ij}k_i k_j\right) \equiv \exp\left(-\frac{1}{2}k^T C k\right).$$

Differentiating with respect to k_1 and then k_2, and putting all k's equal to zero, pulls down a factor $-x_1x_2$ into the integral and gives the desired average of x_1x_2. This trick now gives the average of the product of a string of x's in the form of the "pairing theorem". This is easier to state by an example.

$$\begin{aligned}\langle x_1x_2x_3x_4\rangle &= \langle x_1x_2\rangle\langle x_3x_4\rangle + \langle x_1x_3\rangle\langle x_2x_4\rangle + \langle x_1x_4\rangle\langle x_2x_3\rangle \\ &\equiv C_{12}C_{34} + C_{13}C_{24} + C_{14}C_{23}\end{aligned}$$

A sincere attempt to differentiate G with respect to $k_1k_2k_3$ and k_4 and then put all k's to zero will show that the C's get pulled down in precisely this combination. Deeper thought shows that the pairing rule works even when the x's are not all identical, i.e.,

$$\langle x^4\rangle = \langle x^2\rangle\langle x^2\rangle + \langle x^2\rangle\langle x^2\rangle + \langle x^2\rangle\langle x^2\rangle = 3\langle x^2\rangle^2 = 3\sigma^4$$

or even $\langle x^{2n}\rangle = 1, 3, 5 \ldots (2n-1)\sigma^{2n}$.

The last property is easily checked from the single variable gaussian

$$(2\pi\sigma^2)^{-1/2}\exp(-x^2/2\sigma^2)$$

Since the pairing theorem allows one to calculate **all** averages, it could even be taken to define a gaussian signal, and that is what we do in the main text.

We now sketch a proof of the sampling theorem. Start with a band limited (i.e containing only frequencies less than B) signal sampled at the Nyquist rate, $E_r(n/2B)$. The following expression gives a way of constructing a continuous signal $E_c(t)$ from our samples.

$$E_c(t) = \sum_n E_r(n/2B)\ \mathrm{sinc}(2\pi B(t - \frac{n}{2B}))$$

It is also known as Whitaker's interpolation formula. Each sinc function is diabolically chosen to give unity at one sample point and zero at all the others, so $E_c(t)$ is guaranteed to agree with our samples of $E_r(t)$. It is also band limited (Fourier transform of a flat function extending from $-B$ to $+B$). All that is left to check is that it has the same Fourier coefficients as $E_r(t)$ (it does). And hence, we have reconstructed a band limited function from its Nyquist samples, as promised.

We add a few comments on the notion of Hilbert transform mentioned in the context of associating a complex signal with a real one. It looks rather innocent in the frequency domain, just subtract $\pi/2$ from the phase of each cosine in the Fourier series of $E_r(t)$ and reassemble to get $E_i(t)$. In terms of complex Fourier coefficients, it is a multiplication of the positive frequency component by $-i$ and of the corresponding negative frequency component by $+i$, Apart from the i, this is just multiplication by a step function of the symmetric type, jumping from minus 1 to plus 1 at zero frequency. Hence, in the time domain, it is a convolution of $E_r(t)$ by a kernel which is the Fourier transform of this step function, viz $1/t$ (the value t=0 being excluded by the usual principal value rule). Explicitly, we have

$$E_i(t) = \int E_r(s) P\left[1/(t-s)\right]\ ds/\pi$$

There is a similar formula relating E_r to E_i which only differs by a minus sign. This is sufficient to show that one needs values from the infinite past, and more disturbingly, future, of t to compute $E_i(t)$. This is beyond the reach of ordinary mortals, even those equipped with the best filters and phase shifters. Practical schemes to derive the complex signal in real time thus have to make approximations as a concession to causality.

As remarked in the main text, there are many complex signals whose real parts would give our measured $E_r(t)$. The choice made above seemed natural because it was motivated by the quasimonochromatic case. It also has the mathematical property of creating a function which is very well behaved in the upper half plane of t regarded as a complex variable, (should one ever want to go there). The reason is that $V(t)$ is constructed to have terms like $e^{i\omega t}$ with only positive values of ω. Hence the pedantic name of "analytic signal" for this descendant of the humble phasor. It was the more general problem of continuing something given on the real axis to be well behaved in the upper half plane which attracted someone of Hilbert's IQ to this transform.

Chapter 2

Interferometry and Aperture Synthesis

A. P. Rao

2.1 Introduction

Radio astronomy is the study of the sky at radio wavelengths. While optical astronomy has been a field of study from time immemorial, the "new" astronomies viz. radioastronomy, X-ray, IR and UV astronomy are only about 50 years old. At many of these wavelengths it is essential to put the telescopes outside the confines of the Earth's atmosphere and so most of these "new" astronomies have become possible only with the advent of space technology. However, since the atmosphere is transparent in the radio band (which covers a frequency range of 10 MHz to 300 GHz or a wavelength range of approximately 1mm to 30m) radio astronomy can be done by ground based telescopes (see also Chapter 3).

The field of radioastronomy was started in 1923 when Karl Jansky, (working at the Bell Labs on trying to reduce the noise in radio receivers), discovered that his antenna was receiving radiation from outside the Earth's atmosphere. He noticed that this radiation appeared at the same sidereal (as opposed to solar) time on different days and that its source must hence lie far outside the solar system. Further observations enabled him to identify this radio source as the centre of the Galaxy. To honour this discovery, the unit of flux density in radioastronomy is named after Jansky where

$$1 \text{ Jansky} = 10^{-26} W m^{-2} H z^{-1} \tag{2.1.1}$$

Radio astronomy matured during the second world war when many scientists worked on projects related to radar technology. One of the major discoveries of that period (made while trying to identify the locations of jamming radar signals), was that the sun is a strong emitter of radio waves and its emission is time variable. After the war, the scientists involved in these projects returned to academic pursuits and used surplus equipment from the war to rapidly develop this new field of radioastronomy. In the early phases, radioastronomy was dominated by radio and electronic engineers and the astronomy community, (dominated by optical astronomers), needed considerable persuasion to be convinced that these new radio astronomical discoveries were of relevance to astronomy in general. While the situation has changed considerably since then much

of the jargon of radio astronomy (which is largely borrowed from electrical engineering) remains unfamiliar to a person with a pure physics background. The coherent detection techniques pioneered by radio astronomers also remains by and large not well understood by astronomers working at other wavelength bands. This set of lecture notes aims to familiarize students of physics (or students of astronomy at other wavelengths) with the techniques of radio astronomy.

2.2 The Radio Sky

The sky looks dramatically different at different wave bands and this is the primary reason multi-wavelength astronomy is interesting. In the optical band, the dominant emitters are stars, luminous clouds of gas, and galaxies all of which are thermal sources with temperatures in the range $10^3 - 10^4$ K. At these temperatures the emitted spectrum peaks in the optical band. Sources with temperatures outside this range and emitters of non thermal radiation are relatively weak emitters in the optical band but can be strong emitters in other bands. For example, cold (~ 100 K) objects emit strongly in the infra red and very hot objects ($> 10^5$ K) emit strongly in X-rays. Since the universe contains all of these objects one needs to make multiband studies in order to fully understand it.

For a thermal source with temperature greater than 100 K, the flux density in the radio band can be well approximated by the Rayleigh-Jeans Law[1], viz.

$$S = (2kT/\lambda^2)d\Omega \tag{2.2.2}$$

The predicted flux densities at radio wavelengths are miniscule and one might hence imagine that the radio sky should be dark and empty. However, radio observations reveal a variety of radio sources all of which have flux densities much greater than given by the Rayleigh-Jeans Law, i.e. the radio emission that they emit is not thermal in nature. Today it is known that the bulk of radio emission is produced via the synchrotron mechanism. Energetic electrons spiraling in magnetic fields emit synchrotron radiation. Unlike thermal emission where the flux density increases with frequency, for synchrotron emitters, the flux density increases with wavelength (see Figure 2.1). Synchrotron emitting sources are hence best studied at low radio frequencies.

The dominant sources seen in the radio sky are the Sun, supernova remnants, radio galaxies, pulsars etc. The Sun has a typical flux density of 10^5 Jy while the next strongest sources are the radio galaxy Cygnus A and the supernova remnant Cassiopeia A, both of which have flux densities of $\sim 10^4$ Jy. Current technology permits the detection of sources as weak as a few μJy. It turns out also that not all thermal sources are too weak to detect, the thermal emission from large and relatively nearby HII regions can also be detected easily in the radio band.

Radio emission from synchrotron and thermal emitters is "broad band", i.e. the emission varies smoothly (often by a power law) over the whole radio band. Since the spectrum is relatively smooth, one can determine it by measurements of flux density at a finite number of frequencies. This is a major advantage since radio telescopes tend to be narrow band devices with small frequency spreads ($\Delta\nu/\nu \sim 0.1$). This is partly because it is not practical to build a single radio telescope that can cover the whole radio-band (see eg. Chapter 3) but mainly because radio astronomers share the radio band with a variety of other users (eg. radar, cellular phones, pagers, TV etc.) all of who radiate at power levels high enough to completely swamp the typical radio telescope. By international agreement, the radio spectrum is allocated to different users. Radio astronomy has

[1]The Rayleigh-Jeans Law, as can be easily verified, is the limit of the Plank law when $h\nu << kT$. This inequality is easily satisfied in the radio regime for generally encountered astrophysical temperatures.

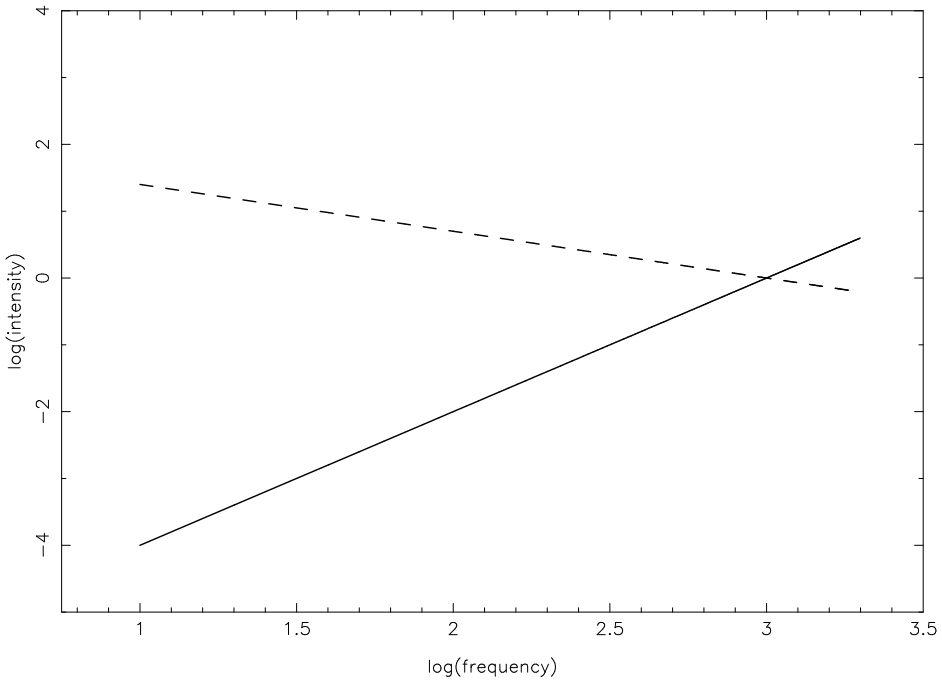


Figure 2.1: Intensity as a function of frequency ("power spectra") for synchrotron (dashed) and thermal (solid) radio sources.

a limited number of protected bands where no one else is permitted to radiate and most radio telescopes work only at these protected frequencies.

Several atoms and molecules have spectral lines in the radio band. For example, the hyperfine transition of the Hydrogen atom corresponds to a line with a wavelength of ~ 21cm. Since atomic hydrogen (HI) is an extremely abundant species in the universe this line is one of the brightest naturally occurring radio lines. The HI 21cm line has been extensively used to study the kinematics of nearby galaxies. High quantum number recombination lines emitted by hydrogen and carbon also fall in the radio band and can be used to study the physical conditions in the ionized interstellar medium. Further the radio line emission from molecules like OH, SiO, H_2O etc. tend to be maser amplified in the interstellar medium and can often be detected to very large distances. Of course, these lines can be studied only if they fall within the protected radio bands. In fact, the presence of radio lines is one of the justifications for asking for protection in a specific part of the radio spectrum. While many of the important radio lines have been protected there are many outside the protected bands that cannot be studied, which is a source of concern. Further, with radio telescopes becoming more and more sensitive, it is possible to study lines like the 21cm line to greater and greater distances. Since in the expanding universe, distance translates to a redshift, this often means that these lines emitted by distant objects move out of the protected radio band and can become unobservable because of interference.

2.3 Signals in Radio Astronomy

A fundamental property of the radio waves emitted by cosmic sources is that they are stochastic in nature, i.e. the electric field at Earth due to a distant cosmic source can

be treated as a random process[2]. Random processes can be simply understood as a generalization of random variables. Recall that a random variable x can be defined as follows. For every outcome o of some given experiment (say the tossing of a die) one assigns a given number to x. Given the probabilities of the different outcomes of the experiment one can then compute the mean value of x, the variance of x etc. If for every outcome of the experiment instead of a number one assigns a given function to x, then the associated process $x(t)$ is called a random process. For a fixed value of t, $x(t)$ is simply a random variable and one can compute its mean, variance etc. as before.

A commonly used statistic for random processes is the auto-correlation function. The auto-correlation function is defined as

$$r_{xx}(t,\tau) = \big\langle x(t)x(t+\tau)\big\rangle$$

where the angular brackets indicate taking the mean value. For a particularly important class of random processes, called wide sense stationary (WSS) processes the auto-correlation function is independent of changes of the origin of t and is a function of τ alone, i.e.

$$r_{xx}(\tau) = \big\langle x(t)x(t+\tau)\big\rangle$$

For $\tau = 0$, $r(\tau)$ is simply the variance σ^2 of $x(t)$ (which for a WSS process is independent of t).

The Fourier transform $S(\nu)$ of the auto-correlation function is called the power spectrum, i.e.

$$S(\nu) = \int_{-\infty}^{\infty} r_{xx}(\tau)e^{-i2\pi\tau\nu}d\tau$$

Equivalently, $S(\nu)$ is the inverse Fourier transform of $r(\tau)$ or

$$r_{xx}(\tau) = \int_{-\infty}^{\infty} S(\nu)e^{i2\pi\tau\nu}d\nu$$

Hence

$$r_{xx}(0) = \sigma^2 = \int_{-\infty}^{\infty} S(\nu)d\nu$$

i.e. since σ^2 is the "power" in the signal, $S(\nu)$ is a function describing how that power is distributed in frequency space, i.e. the "power spectrum".

A process whose auto-correlation function is a delta function has a power spectrum that is flat – such a process is called "white noise". As mentioned in Section 2.2, many radio astronomical signals have spectra that are relatively flat; these signals can hence be approximated as white noise. Radio astronomical receivers however have limited bandwidths, that means that even if the signal input to the receiver is white noise, the signal after passing through the receiver has power only in a finite frequency range. Its auto-correlation function is hence no longer a delta function, but is a sinc function (see Section 2.5) with a width $\sim 1/\Delta\nu$, where $\Delta\nu$ is the bandwidth of the receiver. The width of the auto-correlation function is also called the "coherence time" of the signal. The bandwidth $\Delta\nu$ is typically much smaller than the central frequency ν at which the radio receiver operates. Such signals are hence also often called "quasi-monochromatic" signals. Much like a monochromatic signal can be represented by a constant complex phasor, quasi-monochromatic signals can be represented by complex random processes.

Given two random processes $x(t)$ and $y(t)$, one can define a cross-correlation function

$$r_{xy}(\tau) = \big\langle x(t)y(t-\tau)\big\rangle$$

[2]see Chapter 1 for a more detailed discussion of topics discussed in this section.

where one has assumed that the signals are WSS so that the cross-correlation function is a function of τ alone. The cross-correlation function and its Fourier transform, the cross power spectrum, are also widely used in radio astronomy.

We have so far been dealing with random processes that are a function of time alone. The signal received from a distant cosmic source is in general a function both of the receivers location as well as of time. Much as we defined temporal correlation functions above, one can also define spatial correlation functions. If the signal at the observer's plane at any instant is E(**r**), then spatial correlation function is defined as:

$$V(\mathbf{x}) = \left\langle E(\mathbf{r})E^*(\mathbf{r}+\mathbf{x})\right\rangle$$

Note that strictly speaking the angular brackets imply ensemble averaging. In practice one averages over time[3] and assumes that the two averaging procedures are equivalent. The function V is referred to as the "visibility function" (or just the "visibility") and as we shall see below, it is of fundamental interest in interferometry.

2.4 Interferometry

2.4.1 The Need for Interferometry

The idea that the resolution of optical instruments is limited due to the wave nature of light is familiar to students of optics and is embodied in the Rayleigh's criterion which states that the angular resolution of a telescope/microscope is ultimately diffraction limited and is given by

$$\theta \sim \lambda/D \tag{2.4.3}$$

where D is some measure of the aperture size. The need for higher angular resolution has led to the development of instruments with larger size and which operate at smaller wavelengths. In radioastronomy, the wavelengths are so large that even though the sizes of radio telescopes are large, the angular resolution is still poor compared to optical instruments. Thus while the human eye has a diffraction limit of $\sim\ 20^{''}$ and even modest optical telescopes have diffraction limits[4] of $0.1^{''}$, even the largest radio telescopes (300m in diameter) have angular resolutions of only $\sim 10^{'}$ at 1 metre wavelength. To achieve higher resolutions one has to either increase the diameter of the telescope further (which is not practical) or decrease the observing wavelength. The second option has led to a tendency for radio telescopes to operate at centimetre and millimetre wavelengths, which leads to high angular resolutions. These telescopes are however restricted to studying sources that are bright at cm and mm wavelengths. To achieve high angular resolutions at metre wavelengths one need telescopes with apertures that are hundreds of kilometers in size. Single telescopes of this size are clearly impossible to build. Instead radio astronomers achieve such angular resolutions using a technique called aperture synthesis. Aperture synthesis is based on interferometry, the principles of which are familiar to most physics students. There is in fact a deep analogy between the double slit experiment with quasi-monochromatic light and the radio two element interferometer. Instead of setting up this analogy we choose the more common route to radio interferometry via the van Cittert-Zernike theorem.

[3]For typical radio receiver bandwidths of a few MHz, the coherence time is of the order of micro seconds, so in a few seconds time one gets several million independent samples to average over.

[4]The actual resolution achieved by these telescopes is however usually limited by atmospheric seeing.

2.4.2 The Van Cittert Zernike Theorem

The van Cittert-Zernike theorem relates the spatial coherence function $V(\mathbf{r}_1, \mathbf{r}_2) = \left\langle \mathrm{E}(\mathrm{r}_1) E^*(\mathrm{r}_2) \right\rangle$ to the distribution of intensity of the incoming radiation, $\mathcal{I}(\mathbf{s})$. It shows that the spatial correlation function $V(\mathbf{r}_1, \mathbf{r}_2)$ depends only on $\mathbf{r}_1 - \mathbf{r}_2$ and that if all the measurements are in a plane, then

$$V(\mathbf{r}_1, \mathbf{r}_2) = \mathcal{F}\{I(\mathbf{s})\} \tag{2.4.4}$$

where $\mathcal{F}$ implies taking the Fourier transform. Proof of the van Cittert-Zernike theorem can be found in a number of textbooks, eg. "Optics" by Born and Wolf, "Statistical Optics" by Goodman, "Interferometry and Synthesis in radio astronomy" by Thompson et al. We give here only a rough proof to illustrate the basic ideas.

Let us assume that the source is distant and can be approximated as a brightness distribution on the celestial sphere of radius R (see Figure 2.2). Let the electric field[5] at a point $P_1'(x_1', y_1', z_1')$ at the source be given by $\mathcal{E}(P_1')$. The field $E(P_1)$ at the observation point $P_1(x_1, y_1, z_1)$ is given by[6]

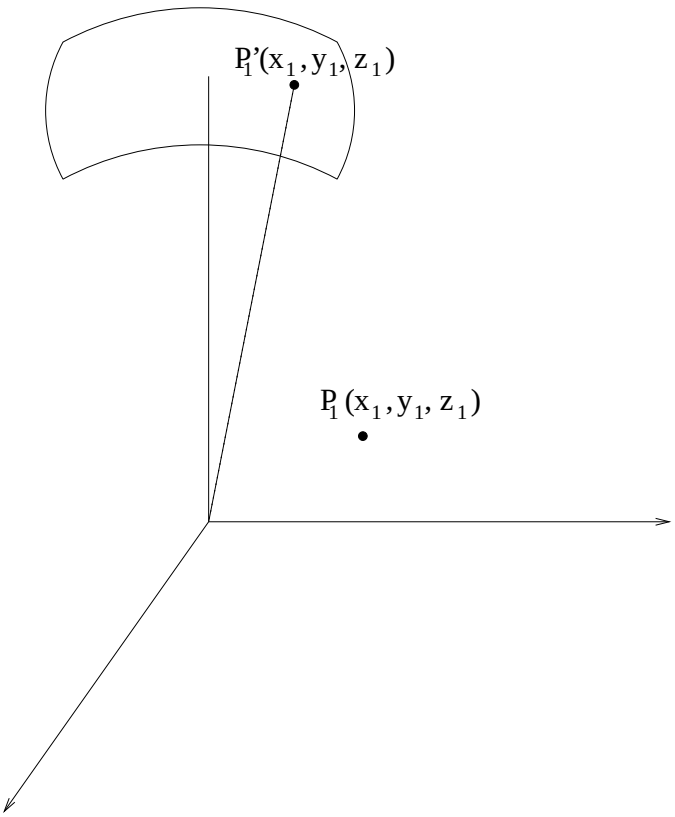


Figure 2.2: Geometry for the van Cittert-Zernike theorem

$$E(P_1) = \int \mathcal{E}(P_1') \frac{e^{-ikD(P_1', P_1)}}{D(P_1', P_1)} d\Omega_1 \tag{2.4.5}$$

[5]We assume here for the moment that the electric field is a scalar quantity. See Chapter 15 for the extension to vector fields.

[6]Where we have invoked Huygens principle. A more rigorous proof would use scalar diffraction theory.

where $D(P_1', P_1)$ is the distance between P_1' and P_1. Similarly if $E(P_2)$ is the field at some other observing point $P_2(x_2, y_2, z_2)$ then the cross-correlation between these two fields is given by

$$\left\langle E(P_1)E^*(P_2)\right\rangle = \int \left\langle \mathcal{E}(P_1')\mathcal{E}^*(P_2')\right\rangle \frac{e^{-ik[D(P_1',P_1)-D(P_2',P_2)]}}{D(P_1',P_1)D(P_2',P_2)} d\Omega_1 d\Omega_2 \tag{2.4.6}$$

If we further assume that the emission from the source is spatially incoherent, i.e. that $\left\langle \mathcal{E}(P_1')\mathcal{E}^*(P_2')\right\rangle = 0$ except when $P_1' = P_2'$, then we have

$$\left\langle E(P_1)E^*(P_2)\right\rangle = \int \mathcal{I}(P_1') \frac{e^{-ik[D(P_1',P_1)-D(P_1',P_2)]}}{D(P_1',P_1)D(P_1',P_2)} d\Omega_1 \tag{2.4.7}$$

where $\mathcal{I}(P_1')$ is the intensity at the point P_1'. Since we have assumed that the source can be approximated as lying on a celestial sphere of radius R we have $x_1' = R\cos(\theta_x) = Rl$, $y_1' = R\cos(\theta_y) = Rm$, and $z_1' = R\cos(\theta_z) = Rn$; (l, m, n) are called "direction cosines". It can be easily shown[7] that $l^2 + m^2 + n^2 = 1$ and that $d\Omega = \frac{dl\ dm}{\sqrt{1-l^2-m^2}}$. We then have:

$$\begin{aligned}
D(P_1', P_1) &= \left[(x_1' - x_1)^2 + (y_1' - y_1)^2 + (z_1' - z_1)^2\right]^{1/2} && (2.4.8)\\
&= \left[(Rl - x_1)^2 + (Rm - y_1)^2 + (Rn - z_1)^2\right]^{1/2} && (2.4.9)\\
&= R\left[(l - x_1/R)^2 + (m - y_1/R)^2 + (n - z_1/R)^2\right]^{1/2} && (2.4.10)\\
&\simeq R\left[(l^2 + m^2 + n^2) - 2/R(lx_1 + my_1 + nz_1)\right]^{1/2} && (2.4.11)\\
&\simeq R - (lx_1 + my_1 + nz_1) && (2.4.12)\\
& && (2.4.13)
\end{aligned}$$

Putting this back into equation 2.4.7 we get

$$\left\langle E(P_1)E^*(P_2)\right\rangle = \frac{1}{R^2}\int \mathcal{I}(l,m) e^{-ik[l(x_2-x_1)+m(y_2-y_1)+n(z_2-z_1)]} \frac{dl\ dm}{\sqrt{1-l^2-m^2}} \tag{2.4.14}$$

Note that since $l^2 + m^2 + n^2 = 1$, the two directions cosines (l, m) are sufficient to uniquely specify any given point on the celestial sphere, which is why the intensity $\mathcal{I}$ has been written out as a function of (l, m) only. It is customary to measure distances in the observing plane in units of the wavelength λ, and to define "baseline co-ordinates" u, v, w such that $u = (x_2 - x_1)/\lambda$, $v = (y_2 - y_1)/\lambda$, and $w = (z_2 - z_1)/\lambda$. The spatial correlation function $\left\langle E(P_1)E^*(P_2)\right\rangle$ is also referred to as the "visibility" $\mathcal{V}(u, v, w)$. Apart from the constant factor $1/R^2$ (which we will ignore hence forth) equation 2.4.14 can then be written as

$$\mathcal{V}(u,v,w) = \int \mathcal{I}(l,m) e^{-i2\pi[lu+mv+nw]} \frac{dl\ dm}{\sqrt{1-l^2-m^2}} \tag{2.4.15}$$

This fundamental relationship between the visibility and the source intensity distribution is the basis of radio interferometry. In the optical literature this relationship is also referred to as the van Cittert-Zernike theorum.

Equation 2.4.15 resembles a Fourier transform. There are two situations in which it does reduce to a Fourier transform. The first is when the observations are confined to a the $U - V$ plane, i.e. when $w = 0$. In this case we have

$$\mathcal{V}(u,v) = \int \frac{\mathcal{I}(l,m)}{\sqrt{1-l^2-m^2}} e^{-i2\pi[lu+mv]} dl\ dm \tag{2.4.16}$$

[7] see for example, Christiansen & Hogbom, "Radio telescopes", Cambridge University Press

i.e. the visibility $\mathcal{V}(u,v)$ is the Fourier transform of the modified brightness distribution $\frac{\mathcal{I}(l,m)}{\sqrt{1-l^2-m^2}}$. The second situation is when the source brightness distribution is limited to a small region of the sky. This is a good approximation for arrays of parabolic antennas because each antenna responds only to sources which lie within its primary beam (see Chapter 3). The primary beam is typically $< 1^o$, which is a very small area of sky. In this case $n = \sqrt{1-l^2-m^2} \simeq 1$. Equation 2.4.15 then becomes

$$\mathcal{V}(u,v,w) = e^{-i2\pi w} \int \mathcal{I}(l,m) e^{-i2\pi[lu+mv]} dl\ dm \tag{2.4.17}$$

or if we define a modified visibility $\widetilde{\mathcal{V}}(u,v) = \mathcal{V}(u,v,w)e^{i2\pi w}$ we have

$$\widetilde{\mathcal{V}}(u,v) = \int \mathcal{I}(l,m) e^{-i2\pi[lu+mv]} dl\ dm \tag{2.4.18}$$

2.4.3 Aperture Synthesis

As we saw in the previous section, the spatial correlation of the electric field in the U-V plane is related to the source brightness distribution. Further, for the typical radio array the relationship between the measured visibility and the source brightness distribution is a simple Fourier transform. Correlation of the voltages from any two radio antennas then allows the measurement of a single Fourier component of the source brightness distribution. Given sufficient number of measurements the source brightness distribution can then be obtained by Fourier inversion. The derived image of the sky is usually called a "map" in radio astronomy, and the process of producing the image from the visibilities is called "mapping".

The radio sky (apart from a few rare sources) does not vary[8]. This means that it is not necessary to measure all the Fourier components simultaneously. Thus for example one can imagine measuring all required Fourier components with just two antennas, (one of which is mobile), by laboriously moving the second antenna from place to place. This method of gradually building up all the required Fourier components and using them to image the source is called "aperture synthesis". If for example one has measured all Fourier components up to a baseline length of say 25 km, then one could obtain an image of the sky with the same resolution as that of a telescope of aperture size 25 km, i.e. one has synthesized a 25 km aperture. In practice one can use the fact that the Earth rotates to sample the U-V plane quite rapidly. As seen from a distant cosmic source, the baseline vector between two antennas on the Earth is continuously changing because of the Earth's rotation (see Figure 2.3). Or equivalently, as the source rises and sets the Fourier components measured by a given pair of antennas is continuously changing. If one has an array of N antennas spread on the Earth's surface, then at any given instant one measures ${}^{N}C_2$ Fourier components (or in radio astronomy jargon one has ${}^{N}C_2$ samples in the U-V plane). As the Earth rotates one samples more and more of the U-V plane. For arrays like the GMRT with 30 antennas, if one tracks a source from rise to set, the sampling of the U-V plane is sufficiently dense to allow extremely high fidelity reconstructions of even complex sources. This technique of using the Earth's rotation to improve "U-V coverage" was traditionally called "Earth rotation aperture synthesis", but in modern usage is usually also simply referred to as "aperture synthesis".

From the inverse relationship of Fourier conjugate variables it follows that short baselines are sensitive to large angular structures in the source and that long baselines are

[8]Or, in the terminology of random processes cosmic radio signals are stationary, i.e. their statistical properties like the mean, auto and cross-correlation functions etc. are independent of the absolute time.

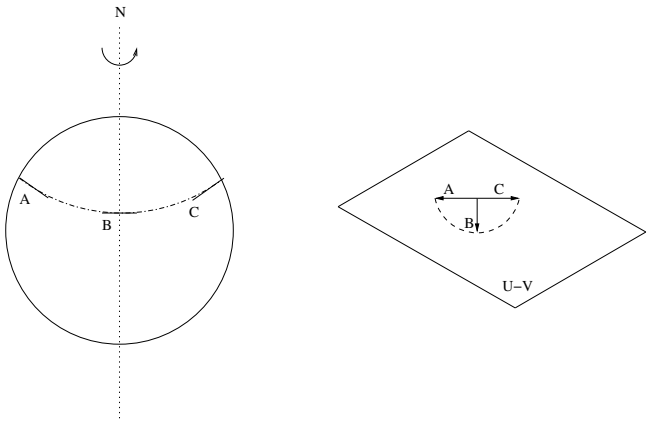


Figure 2.3: The track in the U-V plane traced out by an east-west baseline due to the Earth's rotation.

sensitive to fine scale structure. To image large, smooth sources one would hence like an array with the antennas closely packed together, while for a source with considerable fine scale structure one needs antennas spread out to large distances. The array configuration hence has a major influence on the kind of sources that can be imaged. The GMRT array configuration consists of a combination of a central 1x1 km cluster of densely packed antennas and three 14 km long arms along which the remaining antennas are spread out. This gives a combination of both short and long spacings, and gives considerable flexibility in the kind of sources that can be imaged. Arrays like the VLA on the other hand have all their antennas mounted on rails, allowing even more flexibility in determining how the U-V plane is sampled.

Other chapters in these notes discuss the practical details of aperture synthesis. Chapter 3 discusses how one can use radio antennas and receivers to measure the electric field from cosmic sources. For an N antenna array one needs to measure ${}^{N}C_{2}$ correlations simultaneously, this is done by a (usually digital) machine called the "correlator". The spatial correlation that one needs to measure (see equation 2.4.6) is the correlation between the instantaneous fields at points P_1 and P_2. In an interferometer the signals from antennas at points P_1 and P_2 are transported by cable to some central location where the correlator is – this means that the correlator has also to continuously adjust the delays of the signals from different antennas before correlating them. This and other corrections that need to be made are discussed in Chapter 4, and exactly how these corrections are implemented in the correlator are discussed in Chapters 8 and 9. The astronomical calibration of the measured visibilities is discussed in Chapter 5, while Chapter 16 deals with the various ways in which passage through the Earth's ionosphere corrupts the astronomical signal. Chapters 10, 12 and 14 discuss the nitty gritty of going from the calibrated visibilities to the image of the sky. Chapters 13 and 15 discuss two refinements, viz. measuring the spectra and polarization properties of the sources respectively.

2.5 The Fourier Transform

The Fourier transform $U(\nu)$ of a function $u(t)$ is defined as

$$U(\nu) = \int_{-\infty}^{\infty} u(t) e^{-i2\pi\nu t} dt$$

and can be shown to exist for any function $u(t)$ for which

$$\int_{-\infty}^{\infty} |u(t)| dt < \infty$$

The Fourier transform is invertible, i.e. given $U(\nu)$, $u(t)$ can be obtained using the inverse Fourier transform, viz.

$$u(t) = \int_{-\infty}^{\infty} U(\nu) e^{i2\pi\nu t} d\nu$$

Some important properties of the Fourier transform are listed below (where by convention capitalized functions refer to the Fourier transform)

1. Linearity
$$\mathcal{F}\{au(t) + bv(t)\} = aU(\nu) + bV(\nu)$$
where a, b are arbitrary complex constants.

2. Similarity
$$\mathcal{F}\{u(at)\} = \frac{1}{a} U(\frac{\nu}{a})$$
where a is an arbitrary real constant.

3. Shift
$$\mathcal{F}\{u(t-a)\} = e^{-i2\pi a} U(\nu)$$
where a is an arbitrary real constant.

4. Parseval's Theorem
$$\int_{-\infty}^{\infty} |u(t)|^2 dt = \int_{-\infty}^{\infty} |U(\nu)|^2 d\nu$$

5. Convolution Theorem
$$\mathcal{F} \int_{-\infty}^{\infty} u(t) v(t-\tau) dt = U(\nu) V(\nu)$$

6. Autocorrelation Theorem
$$\mathcal{F} \int_{-\infty}^{\infty} u(t) u(t+\tau) dt = |U(\nu)|^2$$

Some commonly used Fourier transform pairs are:

Table 2.1: Fourier transform pairs

Function	Transform
$e^{\pi t^2}$	$e^{\pi \nu^2}$
1	$\delta(\nu)$
$\cos(\pi t)$	$\frac{1}{2}\delta(\nu - \frac{1}{2}) + \frac{1}{2}\delta(\nu + \frac{1}{2})$
$\sin(\pi t)$	$\frac{i}{2}\delta(\nu - \frac{1}{2}) - \frac{i}{2}\delta(\nu + \frac{1}{2})$
$rect(t)$	$sinc(\nu)$

Chapter 3

Single Dish Radio Telescopes

Jayaram N. Chengalur

3.1 Introduction

As a preliminary to describing radio telescopes, it is useful to have a look at the transparency of the atmosphere to electro-magnetic waves of different frequencies. Figure 3.1 is a plot of the height in the atmosphere at which the radiation is attenuated by a factor of 2 as a function of frequency. There are only two bands at which radiation from outerspace can reach the surface of the Earth, one from $3000\ \mathring{A}$ to $10000\ \mathring{A}$ – the optical/near-infrared window, and one from a few mm to tens of meters – the radio window. Radio waves longer than a few tens of meters get absorbed in the ionosphere, and those shorter than a few mm are strongly absorbed by water vapor. Since mm wave absorption is such a strong function of the amount of water vapour in the air, mm wave telescopes are usually located on high mountains in desert regions.

The optical window extends about a factor of ~ 3 in wavelength, whereas the radio window extends almost a factor of $\sim 10^4$ in wavelength. Hence while all optical telescopes 'look similar', radio telescopes at long wavelengths have little resemblance to radio telescopes at short wavelengths. At long wavelengths, radio telescopes usually consist of arrays of resonant structures, like dipole or helical antennas (Figure 3.2). At short wavelengths reflecting telescopes (usually parabolic antennas, which focus incoming energy on to the focus, where it is absorbed by a small feed antenna) are used (Figure 3.3).

Apart from this difference in morphology of antennas, the principal difference between radio and optical telescopes is the use of coherent (i.e. with the preservation of phase information) amplifiers in radio astronomy. The block diagram for a typical single dish radio astronomy telescope is shown in Figure 3.4. Radio waves from the distant cosmic source impinge on the antenna and create a fluctuating voltage at the antenna terminals. This voltage varies at the same frequency as the cosmic electro-magnetic wave, referred to as the **Radio Frequency** (RF). The voltage is first amplified by the front-end (or Radio Frequency) amplifier. The signal is weakest here, and hence it is very important that the amplifier introduce as little noise as possible. Front end amplifiers hence usually use low noise solid state devices, High Electron Mobility Transistors (HEMTs), often cooled to liquid helium temperatures.

After amplification, the signal is passed into a mixer. A mixer is a device that changes the frequency of the input signal. Mixers have two inputs, one for the signal whose frequency is to be changed (the RF signal in this case), the other input is usually a pure sine

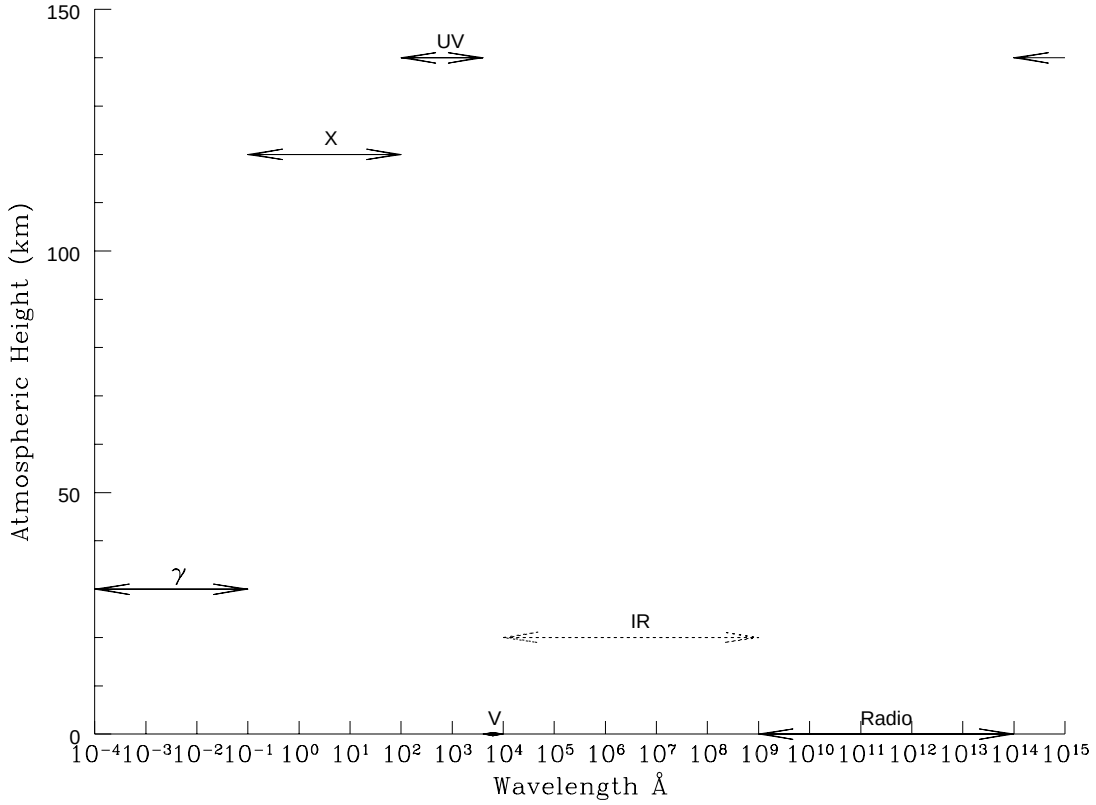


Figure 3.1: The height above the Earth's surface where cosmic electro-magnetic radiation is attenuated by a factor of two. There are two clear windows the optical (V) ($\sim 4000 - 10000$ Å) and the radio $\sim 1\mathrm{mm} - 10\mathrm{m}$. In addition there are a few narrow windows in the infra-red (IR) wavelength range. At all other wavelengths astronomy is possible only through satellites.

wave generated by a tunable signal generator, the **Local Oscillator** (LO). The output of the mixer is at the beat frequency of the radio frequency and the local oscillator frequency. So after mixing, the signal is now at a different (and usually lower) frequency than the RF, this frequency is called the **Intermediate Frequency** (IF). The main reason for mixing (also called heterodyning) is that though most radio telescopes operate at a wide range of radio frequencies, the processing required at each of these frequencies is identical. The economical solution is to convert each of these incoming radio frequencies to a standard IF and then to use the exact same back-end equipment for all possible RF frequencies that the telescope accepts. In telescopes that use co-axial cables to transport the signal across long distances, the IF frequency is also usually chosen so as to minimize transmission loss in the cable. Sometimes there is more than one mixer along the signal path, creating a series of IF frequencies, one of which is optimum for signal transport, another which is optimum for amplification etc. This is called a 'super-heterodyne' system. For example, the GMRT (see Chapter 21) accepts radio waves in six bands from 50 MHz to 1.4 GHz and has IFs at 130 MHz, 175 MHz and 70 MHz[1].

[1]There are IFs at 130 MHz and 175 MHz to allow the signals from the two different polarizations received

Figure 3.2: The Mauritius Radio Telescope. This is a low frequency (150 MHz) array of which the individual elements are helical antennas.

After conversion to IF, the signal is once again amplified (by the IF amplifier), and then mixed to a frequency range near 0 Hz (the **Base Band** (BB) and then fed into the backend for further specialized processing. What backend is used depends on the nature of the observations. If what you want to measure is simply the total power that the telescope receives then the backend could be a simple square law detector followed by an integrator. (Remember the signal is a *voltage* that is proportional to amplitude of the electric field of the incoming wave, and since the power in the wave goes like the square of its amplitude, the square of the voltage is a measure of the strength of the cosmic source). The integrator merely averages the signal to improve the signal to noise ratio. For spectral line observations the signal is passed into a spectrometer instead of a broad band detector. For pulsar observations the signal is passed into special purpose 'pulsar machines'. Spectrometers (usually implemented as "correlators") and pulsar machines are fairly complex and will not be discussed further here (see instead Chapters 8 and 17 more more details). The rest of this chapter discusses only the first part of this block diagram, viz. the antenna itself.

3.2 EM Wave Basics

A cosmic source typically emits radio waves over a wide range of frequencies, but the radio telescope is sensitive to only a narrow band of emission centered on the RF. We can hence, to zeroth order, approximate this narrow band emission as a monochromatic wave. (More realistic approximations are discussed in Chapter 15). The waves leaving the cosmic source have spherical wavefronts which propagate away from the source at the speed of light. Since most sources of interest are very far away from the Earth, the radio telescope only sees a very small part of this spherical wave front, which can be well approximated by a plane wave front. Electro-magnetic waves are vector waves, i.e. the

by the antenna to be frequency division multiplexed onto the same optical fiber for transport to the central electronics building.

Figure 3.3: The Caltech Sub-millimeter Observatory (CSO) at Mauna Kea in Hawaii. The telescope operates in the the sub-mm wavlength range.

electric field has a direction as well as an amplitude. In free space, the electric field of a plane wave is constrained to be perpendicular to its direction of propagation and the power carried by the wave is proportional to the square of the amplitude of the electric field.

Consider a plane EM wave of frequency ν propagating along the Z axis (Figure 3.6). The electric field then can have only two components, one along the X axis, and one along the Y axis. Since the wave is varying with a frequency ν, each of these components also varies with a frequency ν, and at any one point in space the electric field vector will also vary with a frequency ν. The **polarization** of the wave characterizes how the direction of the electric field vector varies at a given point in space as a function of time.

The most general expression for each of the components of the electric field of a plane monochromatic wave[2] is:

$$E_X = A_X \cos(2\pi\nu t + \delta_X)$$
$$E_Y = A_Y \cos(2\pi\nu t + \delta_Y)$$

where $A_X,\ A_Y,\ \delta_X,\ \delta_Y$ are constants. If $A_Y = 0$, then the field only has one component along the X axis, which increases in amplitude from $-A_X$ to $+A_X$ and back to $-A_X$ over one period. Such a wave is said to be **linearly polarized** along the X axis. Similarly if A_X is zero then the wave is linearly polarized along the Y axis. Waves which are generated by **dipole** antennas are linearly polarized along the length of the dipole. Now consider a wave for which $A_X = A_Y, \delta_X = 0$, and $\delta_Y = -\pi/2$. If we start at a time at which the X component is a maximum, then the Y component is zero and the total field points along the +X axis. A quarter period later, the X component will be zero and the Y component will be at maximum, the total field points along the +Y direction. Another quarter of a period later, the Y component is again zero, and the X component is at minimum, the total field points

[2]Monochromatic waves are necessarily 100% polarized. As discussed in Chapter 15 quasi-monochromatic waves can be partially polarized.

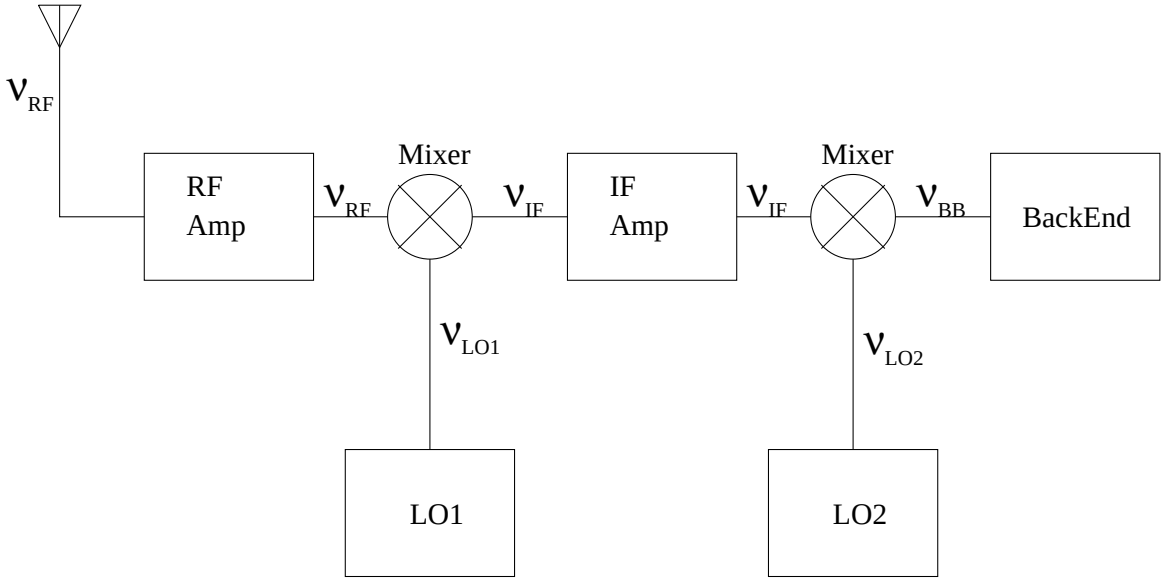


Figure 3.4: Block diagram of a single dish radio telescope.

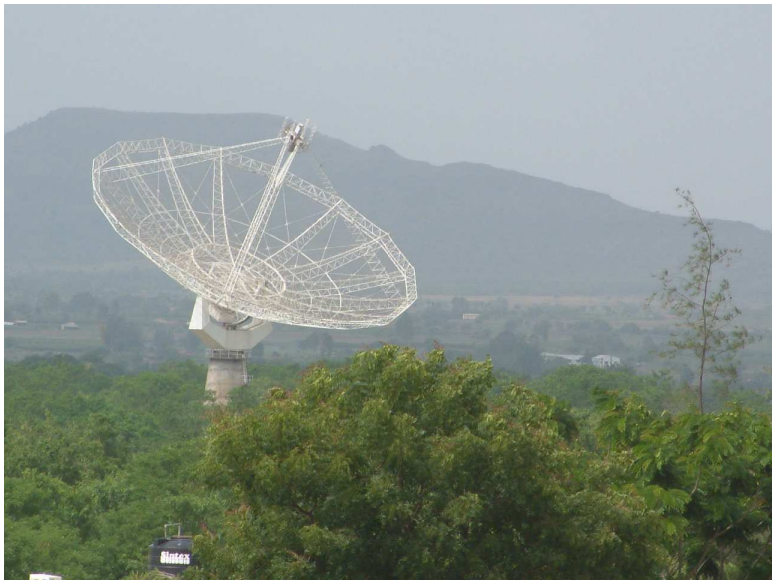


Figure 3.5: One of the 30 GMRT antennas

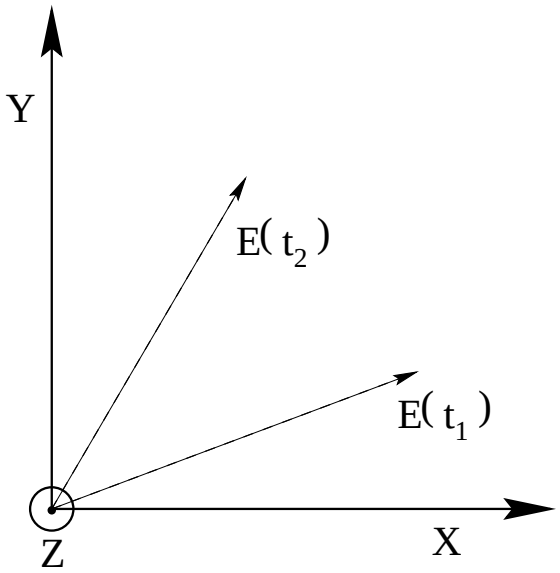


Figure 3.6: Electric field of a plane wave

along the -X direction. Thus over one period, the tip of the electric field vector describes a circle in the XY plane. Such a wave is called **circularly polarized**. If δ_Y were $= \pi/2$ then the electric field vector would still describe a circle in the XY plane, but it would rotate in the opposite direction. The former is called **Right Circular Polarization** (RCP) and the latter **Left Circular Polarization** (LCP).[3] Waves generated by **Helical** antennas are circularly polarized. In the general case when all the constants have arbitrary values, the tip of the electric wave describes an ellipse in the XY plane, and the wave is said to be **elliptically polarized**.

Any monochromatic wave can be decomposed into the sum of two orthogonal polarizations. What we did above was to decompose a circularly polarized wave into the sum of two linearly polarized components. One could also decompose a linearly polarized wave into the sum of LCP and RCP waves, with the same amplitude and π radians out of phase. Any antenna is sensitive to only *one* polarization (for example a dipole antenna only absorbs waves with electric field along the axis of the dipole, while a helical antenna will accept only one sense of circular polarization). Note that the reflecting surface of a telescope could well [4] work for both polarizations, but the feed antenna will respond to only one polarization. To detect both polarizations one need to put two feeds (which could possibly be combined into one *mechanical* structure) at the focus. Each feed will require its own set of electronics like amplifiers and mixers etc.

EM waves are usually described by writing explicitly how the electric field strength varies in space and time. For example, a plane wave of frequency ν and wave number k ($k = 2\pi/\lambda,\ \lambda = c/\nu$) propagating along the Z axis and linearly polarized along the X axis could be described as

$$E(z,t) = A\cos(2\pi\nu t - kz)$$

[3]This RCP-LCP convention is unfortunately not fixed, and the reverse convention is also occasionally used, leading to endless confusion. It turns out however, that most cosmic sources have very little circular polarization.

[4]Not all reflecting radio telescopes have surfaces that reflect both polarizations. For example, the Ooty radio telescope's (Figure 3.16) reflecting surface consists of a parallel set of thin stainless steel wires, which only reflect the polarization with the electric field parallel to the wires.

This could also be written as

$$E(z,t) = \text{Real}(Ae^{j(2\pi\nu t - kz)})$$

where Real() implies taking the real part of () and j is the imaginary square root of -1. Since all the time variation is at the same frequency ν, one could suppress writing it out explicitly and introduce it only when one needs to deal with physical quantities. So, one could equally well describe the wave by the complex quantity **A**, where $\mathbf{A} = A\, e^{-jkz}$, and understand that the physical field is obtained by multiplying **A** by $e^{j2\pi\nu t}$ and taking the real part of the product. The field **A** is called the **phasor** field[5]. So for example the phasor field of the wave

$$E = A\cos(2\pi\nu t - kz + \delta)$$

is simply $\mathbf{A}e^{j\delta}$.

3.3 Signals and Noise in Radio Astronomy

3.3.1 Signals

At radio frequencies, cosmic source strengths are usually measured in **Janskys**[6] **(Jy)**. Consider a plane wave from a distant point source falling on the Earth. If the energy per unit frequency passing through an area of 1 square meter held perpendicular to the line of sight to the source is 10^{-26} watts then the source is said to have a brightness of 1 Jy, i.e.

$$1\ Jy = 10^{-26}\ W/m^2/Hz,$$

For an extended source, there is no longer a unique direction to hold the square meter, such sources are hence characterized by a sky brightness B, the energy flow at Earth per unit area, per unit time, per unit solid angle, per unit Frequency, i.e. the units of brightness are $W/m^2/Hz/sr$.

Very often the sky brightness is also measured in temperature units. To motivate these units, consider a black body at temperature T. The radiation from the black body is described by the Planck spectrum

$$B(\nu) = \frac{2h\nu^3}{c^2}\frac{1}{e^{h\nu/kT}-1} \qquad W/m^2/Hz/sr$$

i.e. the same units as the brightness. For a typical radio frequency of 1000 MHz, $h\nu/k = 0.048$, hence

$$e^{h\nu/kT} \sim 1 + h\nu/kT$$

and

$$B(\nu) \simeq \frac{2\nu^2}{c^2}kT = 2kT/\lambda^2$$

This approximation to the Planck spectrum is called the **Rayleigh-Jeans** approximation, and is valid through most of the radio regime. From the R-J approximation,

$$T = \frac{\lambda^2}{2k}B(\nu)$$

[5]For qasi monochromatic waves, (see Chapter 1), one has the related concept of the complex analytical signal

[6]As befitting its relative youth, this is a linear, MKS based scale. At most other wavelengths, the brightness is traditionally measured in units far too idiosyncratic to be described in this footnote.

In analogy, the brightness temperature T_B of an extended source is *defined* as

$$T_B = \frac{\lambda^2}{2k} B(\nu).$$

where $B(\nu)$ is the sky brightness of the source. Note that in general the brightness temperature T_B has no relation to the physical temperature of the source.

For certain sources, like the quiet sun and HII regions, the emission mechanism is **thermal bremstrahlung**, and for these sources, provided the optical depth is large enough, the observed spectrum will be the Rayleigh-Jeans tail of the black body spectrum. In this case, the brightness temperature is a directly related to the physical temperature of the electrons in the source. Sources for which the **synchrotron** emission mechanism dominates, the spectrum is not black-body, but is usually what is called *steep spectrum*[7], i.e. the flux increases sharply with increasing wavelength. At low frequencies, the most prominent such source is the Galactic non-thermal continuum, for which the flux $S \propto \nu^{-\alpha},\ \alpha \sim 1$. At low frequencies hence, the sky brightness temperature dominates the system temperature[8]. Pulsars and extended extra-galactic radio sources too in general have steep spectra and are brightest at low frequencies. At the extreme end of the brightness temperature are masers where a lot of energy is pumped out in a narrow collimated molecular line, the brightness temperatures could reach $\sim 10^{12}$ K. This could certainly not be the physical temperature of the source since the molecules disintegrate at temperatures well below 10^{12} K.

3.3.2 Noise

An antenna absorbs power from the radio waves that fall on it. This power is also usually specified in temperature units, i.e. degrees Kelvin. To motivate these units, consider a resistor placed in a thermal bath at a temperature T. The electrons in the resistor undergo random thermal motion, and this random motion causes a current to flow in the resistor. On the average there are as many electrons moving in one direction as in the opposite direction, and the average current is zero. The power in the resistor however depends on the *square* of the current and is not zero. From the equipartition principle one could compute this power as a function of the temperature, and in the radio regime the power per unit frequency is well approximated by the **Nyquist formula**:

$$P = kT,$$

where k is the same Boltzmann constant as in the Planck law. In analogy with this, if a power P (per unit frequency) is available at an antenna's terminals the antenna is *defined* to have an antenna temperature of

$$T_A = \frac{P}{k}$$

Note again that the antenna temperature does not correspond to the physical temperature of the antenna. Similarly the total power available at a radio telescope terminals, referred to the receiver (i.e. the RF amplifier) inputs[9] is defined as the system temperature T_{sys}, i.e.

$$T_{sys} = \frac{\text{Total Power referred to receiver inputs}}{k}$$

[7] provided that the source is optically thin

[8] See the discussion on system temperature later in this section

[9] By 'referred to the reciever inputs' we mean the following. Suppose you have a noise power P at the output of the radio telescope. If there is only one stage of amplification with gain G, then the power referred to the inputs is P/G. If there is more than one stage of amplification, one has to rescale each noise source along the signal path by the gain of all the preceeding amplifiers.

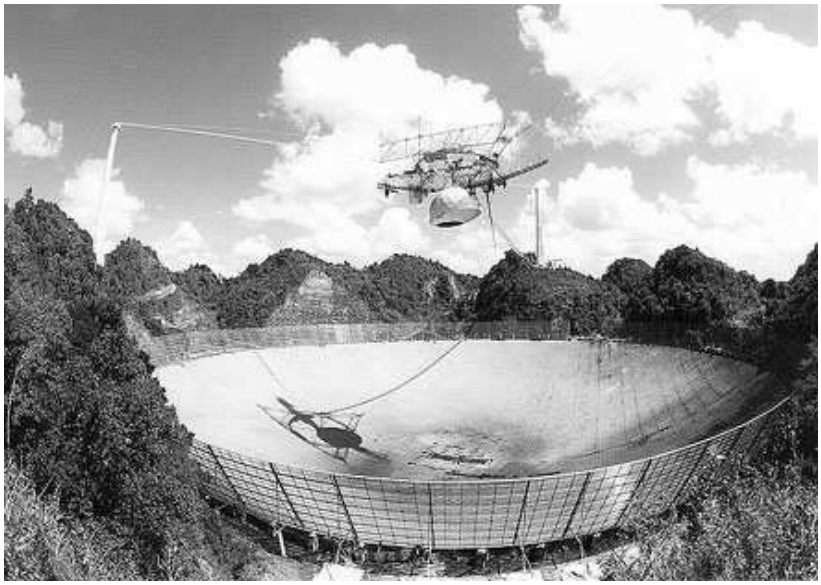


Figure 3.7: The Arecibo telescope consists of a large (300 m) spherical reflector fitted into a naturally occuring valley. The telescope has feeds which are suspended from cables which originate from towers on the surrounding hills. Photo courtesy of NAIC, Arecibo observatory.

The system temperature when looking at blank sky is a measure of the total random noise in the system and hence it is desirable to make the system temperature as low as possible. Noise from the various sub systems that make up the radio telescope are uncorrelated and hence add up linearly. The system temperature can be very generally written as

$$T_{sys} = T_{sky} + T_{spill} + T_{loss} + T_{rec}$$

T_{sky} is the contribution of the background sky brightness. For example the galaxy is a strong emitter of non thermal[10] continum radiation, which at low frequencies usually dominates the system temperature. At all frequencies the sky contributes at least 3K from the cosmic background radiation.[11]

The feed antenna is supposed to collect the radiation focused by the reflector. Often the feed antenna also picks up stray radiation from the ground (which radiates approximately like a black body at 300 K) around the edge of the reflector. This added noise is called spillover noise, and is a very important contribution to the system temperature at a telescope like Arecibo. In Figure 3.8 is shown (schematically) the system temperature for the (pre-upgrade) Arecibo telescope at 12cm as a function of the zenith angle at which the telescope is pointed. At high zenith angles the feed radiation spills over the edge of the dish and picks up a lot of radiation from the surrounding hills and the

[10] By non thermal radiation one means simply that the source function is not the Planck spectrum.

[11] Historicaly, this fact was discovered by Penzias and Wilson when they set out to perform the relatively mundane task of calibrating the system temperature of their radio telescope. This excess 3K discovered to come from the sky was identified with the radiation from the Big Bang, and was one of the powerful pieces of evidence in favour of the Big Bang model. The field of Radio Astronomy itself was started by Karl Jansky, who too was engaged in the task of calibrating the system temperature of his antenna (he had been set the task of characterizing the various kinds of noise which radio receivers picked up, this noise was harmful to trans-atlantic communication, and was hence essential to understand). Jansky discovered that one component of the 'radio noise' was associated with the Galactic center, the first detection of extra-terrestrial radio waves.

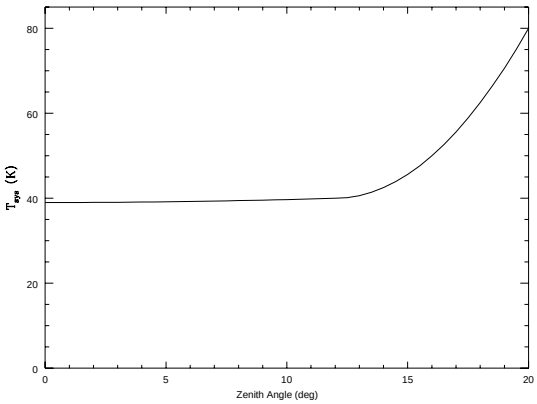


Figure 3.8: Schematic of the variation of T_{sys} with zenith angle for the pre-upgrade Arecibo.

system temperature changes from under 40 K to over 80 K. If a reflecting screen were to be placed around the telescope edges, then, the spill over radiation will be sky radiation reflected by the screen, and not thermal radiation from the ground. At cm wavelengths, $T_{sky} << T_{ground}$, so such a ground screen would significantly reduce the system temperature at high zenith angles[12].

Any lossy element in the feed path will also contribute noise (T_{loss}) to the system. This follows from Kirchoff's law which states that good absorbers are also good emitters, and that the ratio of emission to absorption in thermodynamic equilibrium is given by the Planck spectrum at the absorber's physical temperature. This is the reason why there are rarely any uncooled elements between the feed and the first amplifier. Finally, the receiver also adds noise to the system, which is characterized by T_{rec}. The noise added after the first few stages of amplification is usually an insignificant fraction of the signal strength and can often be ignored.

The final, increasingly important contributor to the system temperature is terrestrial interference. If the bandwidth of the interference is large compared to the spectral resolution, the interference is called broad band. Steady, broad band interference increases the system temperature, and provided this increase is small its effects are relatively benign. However, typically interference varies on a very rapid time scale, causing a rapid fluctuation in the system temperature. This is considerably more harmful, since such fluctuations could have harmonics which are mistaken for pulsars etc. In aperture synthesis telescopes such time varying effects will also produce artifacts in the resulting image[13]. Interference whose bandwidth is small compared to the spectral resolution is called narrow band interference. Such interference, provided it is weak enough will corrupt only one spectral channel in the receiver. Provided this spectral channel is not important (i.e. does not coincide with for eg. a spectral line from the source) it can be flagged with little

[12]As can be seen from Figure 3.7, such a screen has indeed been built, and it has dramatically reduced the system temperature at high zenith angles. The wire mesh for this screen was produced, with the co-ordination of NCRA by the same contractor who fabricated the mesh for the GMRT antennas, and was exported to the USA.

[13]It is often claimed that interferometers are immune from interference because different antennas "see" different interfering sources and these do not correlate with one another. However since the interference is typically varying on timescales faster than the system temperature is calibrated, the resulting variations in the system temperatures of the different antennas cause variations in the observed correlation coefficent (for telescopes which do a continuous normalization by the auto-correlation of each antenna's signal) and hence artifacts in the image plane.

loss of information. However, if the interference is strong enough, the receiver saturates, which has several deleterious effects. Firstly since the receiver is no longer in its linear range, the increase in antenna temperature on looking at a cosmic source is no longer simply related to the source brightness, making it difficult, and usually impossible to derive the actual source brightness. This is called **compression**. Further if some other spectral feature is present, perhaps even a spectral line from the source, spurious signals are produced at the beat frequencies of the true spectral line and the interference. These are called **intermodulation products**. Given the increasingly hostile interference environment at low frequencies, it is important to have receivers with large **dynamic range**, i.e. whose region of linear response is as large as possible. It could often be the case, that it is worth increasing the receiver temperature provided that one gains in dynamic range. For particularly strong and steady sources of interference (such as carriers for nearby TV stations), it is usually the practice to block such signals out using narrow band filters before the first amplifier[14].

3.3.3 Signal to Noise Ratio

Since the signals[15] in a radio telescope are random in nature, the output of a total power detector attached to a radio telescope too will show random fluctuations. Supposing a telescope with system temperature T_{sys}, gain G, and bandwidth $\Delta\nu$ is used to try and detect some astrophysical source. The strategy one could follow is to first look at a 'blank' part of the sky, and then switch to a region containing the source. Clearly if the received power increases, then one has detected radio waves from this source[16]. But given that the output even on a blank region of sky is fluctuating, how can one be sure that the increase in antenna temperature is not a random fluctuation but is indeed due to the astrophysical source? In order to make this decision, one needs to know what the rms is in the fluctuations. It will be shown later[17], that for a total power detector with instantaneous rms T_{sys}, the rms after integrating a signal of bandwidth $\Delta\nu$ Hz for τ seconds is[18] $T_{sys}/\sqrt{\Delta\nu\tau}$. The increase in system temperature is just GS, where S is the flux density of the source. The signal to noise ratio is hence

$$\mathrm{snr} = \frac{\mathrm{GS}\sqrt{\Delta\nu\tau}}{\mathrm{T_{sys}}}$$

This is the fundamental equation for the sensitivity of a single dish telescope. Provided the signal to noise ratio is sufficiently large, one can be confident of having detected the source.

The signal to noise ratio here considers only the 'thermal noise', i.e. the noise from the receivers, spillover, sky temperature etc. In addition there will be random fluctuations from position to position as discussed below because of confusion. For most single dish radio telescopes, especially at low frequencies, the thermal noise reaches the confusion limit (see Section 3.4) in fairly short integration times. To detect even fainter sources, it becomes necessary then to go for higher resolution, usually attainable only through interferometry.

[14] Recall from the discussion above on the effect of introducing lossy elements in the signal path that the price one pays is precisely an increase in receiver temperature

[15] Apart from interference etc.

[16] Assuming of course that you have enough spatial resolution to make this identification

[17] Chapter 5

[18] This can be heuristically understood as follows. For a stochastic proccess of bandwidth $\Delta\nu$, the coherence time is $\sim 1/\Delta\nu$, which means that in a time of τ seconds, one has $\Delta\nu\,\tau$ independent samples. The rms decreases as the square root of the number of independent samples.

3.4 Antenna Patterns

The most important characteristic of an antenna is its ability to absorb radio waves incident upon it. This is usually described in terms of its effective aperture. The effective aperture of an antenna is defined as

$$A_e = \frac{\text{Power density available at the antenna terminals}}{\text{Flux density of the wave incident on the antenna}}$$

The units are

$$\frac{W/Hz}{W/m^2/Hz} = m^2$$

The effective area is a function of the direction of the incident wave, because the antenna works better in some directions than in others. Hence

$$A_e \quad = \quad A_e(\theta, \phi)$$

This directional property of the antenna is often described in the form of a **power pattern**. The power pattern is simply the effective area normalized to be unity at the maximum, i.e.

$$P(\theta, \phi) \quad = \quad \frac{A_e(\theta, \phi)}{A_e^{max}}$$

The other common way to specify the directive property of an antenna is the **field pattern**. Consider an antenna receiving radio waves from a distant point source. The voltage at the terminals of the antenna as a function of the direction to the point source, normalized to unity at maximum, is called the field pattern $f(\theta, \phi)$ of the antenna. The pattern of an antenna is the same regardless of whether it is used as a transmitting antenna or as a receiving antenna, i.e. if it transmits efficiently in some direction, it will also receive efficiently in that direction. This is called **Reciprocity**, (or occassionaly Lorentz Reciprocity) and follows from Maxwell's equations. From reciprocity it follows that the electric field far from a transmitting antenna, normalized to unity at maximum, is simply the Field pattern $f(\theta, \phi)$. Since the power flow is proportional to the square of the electric field, the power pattern is the square of the field pattern. The power pattern is hence real and positive semi-definite.

A typical power pattern is shown in Figure 3.9. The power pattern has a primary maximum, called the **main lobe** and several subsidiary maxima, called *side lobes*. The points at which the main lobe falls to half its central value are called the Half Power points and the angular distance between these points is called the **Half Power Beamwidth (HPBW)**. The minima of the power pattern are called **nulls**. For radio astronomical applications one generally wants the HPBW to be small (so that the nearby sources are not confused with one another), and the sidelobes to be low (to minimize stray radiation). From simple diffraction theory it can be shown that the HPBW of a reflecting telescope is given by

$$\Theta_{HPBW} \sim \lambda / D$$

where D is the physical dimension of the telescope. λ and D must be measured in the same units and Θ is in radians. So the larger the telescope, the better the resolution. For example, the HPBW of a 700 foot telescope at 2380 MHz is about 2 arcmin. This is very poor resolution – an optical telescope ($\lambda \sim 5000\mathring{A}$), a few inches in diameter has a resolution of a few arc seconds. However, the resolution of single dish radio telescopes, unlike optical telescopes, is not limited by atmospheric turbulence. Figure 3.10 shows the power pattern of the (pre-upgrade) Arecibo telescope at 2380 MHz. Although the

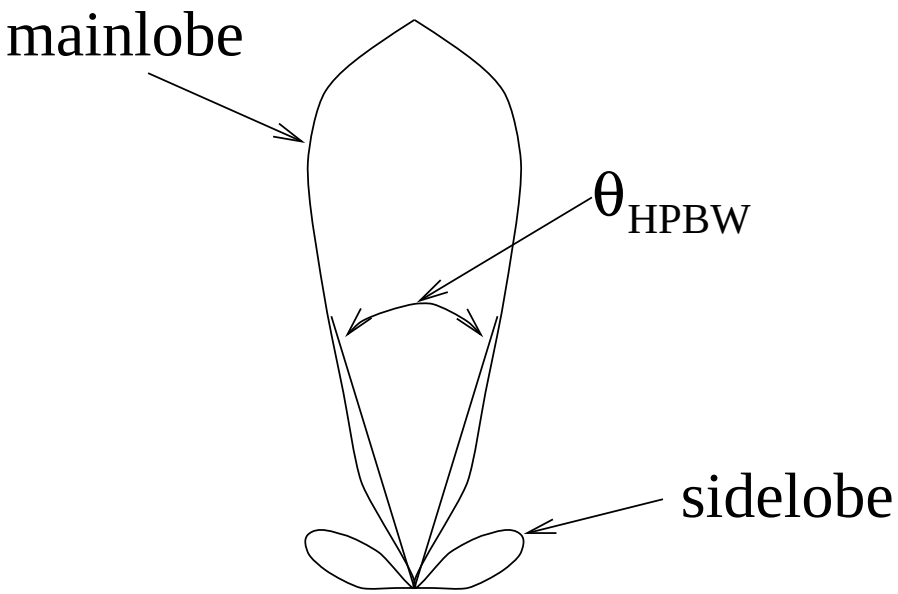


Figure 3.9: Schematic power pattern of an antenna

telescope is 1000 ft in diameter, only a 700 ft diameter aperture is used at any given time, and the HPBW is about 2 arc min. There are two sidelobe rings, which are not quite azimuthally symmetric.

There are two other patterns which are sometimes used to describe antennas. The first is the directivity $D(\theta,\phi)$. The directivity is defined as:

$$D(\theta,\phi) = \frac{\text{Power emitted into } (\theta,\phi)}{(\text{Total power emitted})/4\pi} \tag{3.4.1}$$

$$= \frac{4\pi P(\theta,\phi)}{\int P(\theta,\phi)\, d\Omega} \tag{3.4.2}$$

$$\tag{3.4.3}$$

This is the 'transmitting' pattern of the antenna, and from reciprocity should be the same as the recieving power pattern to within a constant factor. We will shortly work out the value of this constant. The other pattern is the gain $G(\theta,\phi)$. The gain is defined as:

$$G(\theta,\phi) = \frac{\text{Power emitted into } (\theta,\phi)}{(\text{Total power input})/4\pi} \tag{3.4.4}$$

The gain is the same as the directivity, except for an efficiency factor. Finally a figure of merit for reflector antennas is the aperture efficiency, η. The aperture efficiency is defined as:

$$\eta = \frac{A_e^{max}}{A_g} \tag{3.4.5}$$

where A_g is the geometric cross-sectional area of the main reflector. As we shall prove below, the aperture efficiency is at most 1.0.

Consider observing a sky brightness distribution $B(\theta)$ with a telescope with a power pattern like that shown in Figure 3.9. The power available at the antenna terminals is

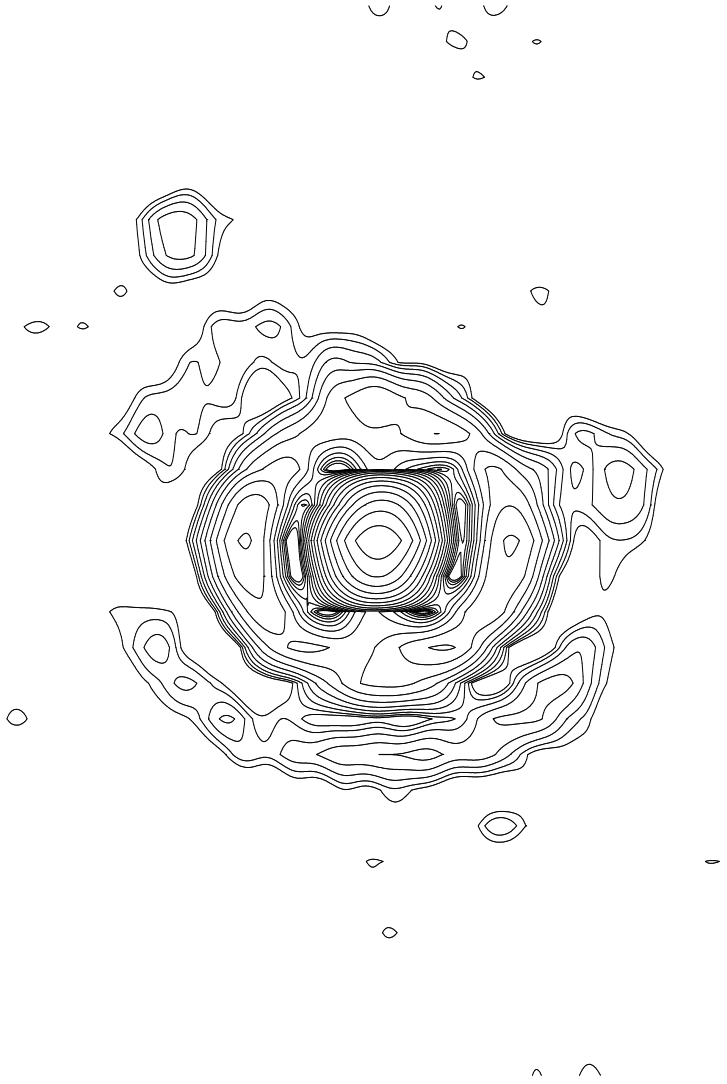


Figure 3.10: The (pre-upgrade) Arecibo power pattern at 2380 MHz. The HPBW is $\sim 2'$.

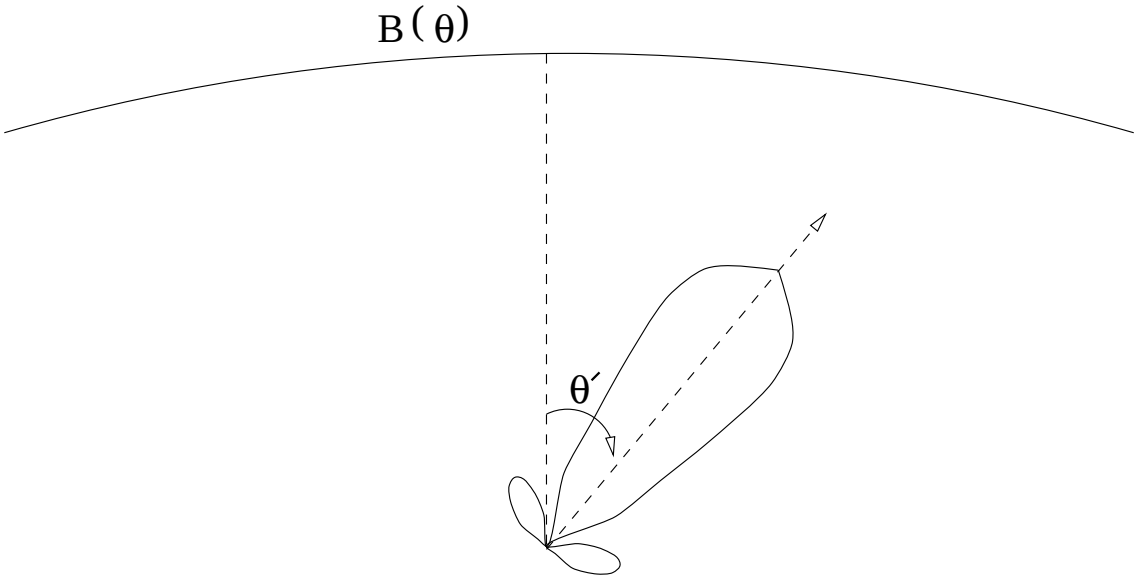


Figure 3.11: The antenna temperature is the convolution of the sky brightness and the telescope beam.

the integral of the brightness in a given direction times the effective area in that direction (Figure 3.11).

$$W(\theta^{'}) = \frac{1}{2}\int B(\theta)A_e(\theta - \theta^{'})d\theta \tag{3.4.6}$$

where the available power W is a function of $\theta^{'}$, the direction in which the telescope is pointed. The factor of $\frac{1}{2}$ is to account for the fact that only one polarization is absorbed by the antenna. In two dimensions, the expression for W is:

$$W(\theta^{'},\phi^{'}) = \frac{1}{2}\int B(\theta,\phi)A_e(\theta - \theta^{'},\phi - \phi^{'})\sin(\theta)d\theta d\phi \tag{3.4.7}$$

in temperature units, this becomes:

$$T_A(\theta^{'},\phi^{'}) = \frac{1}{2}\int \frac{T_B(\theta,\phi)}{\lambda^2}A_e(\theta - \theta^{'},\phi - \phi^{'})\sin(\theta)d\theta d\phi \tag{3.4.8}$$

or

$$T_A(\theta^{'},\phi^{'}) = \frac{A_e^{max}}{\lambda^2}\int T_B(\theta,\phi)P(\theta - \theta^{'},\phi - \phi^{'})\sin(\theta)d\theta d\phi \tag{3.4.9}$$

So the antenna temperature is a weighted average of the sky temperature, the weighting function being the power pattern of the antenna. Only if the power pattern is a single infinitely sharp spike is the antenna temperature the same as the sky temperature. For all real telescopes, however, the antenna temperature is a smoothed version of the sky temperature. Supposing that you are making a sky survey for sources. Then a large increase in the antenna temperature could mean either that there is a source in the main beam, or that a collection of faint sources have combined to give a large total power. From the statistics of the distribution of sources in the sky (presuming you know it) and the power pattern, one could compute the probability of the latter event. This gives a lower limit to the weakest detectable source, below this limit,(called the **confusion limit**), one can no longer be confident that increases in the antenna temperature correspond to a

single source in the main beam. The confusion limit is an important parameter of any given telescope, it is a function of the frequency and the assumed distribution of sources.

Now consider an antenna terminated in a resistor, with the entire system being placed in a black box at temperature T. After thermal equilibrium has been reached, the power flowing from the resistor to the antenna is:

$$P_{R\rightarrow A} = kT$$

The power flow from the antenna to the resistor is (from equation (3.4.9) and using the fact that the sky temperature is the same everywhere)

$$P_{A\rightarrow R} = (\frac{A_e^{max} kT}{\lambda^2}) \int P(\theta,\phi) d\Omega$$

In thermal equilibrium the net power flow has to be zero, hence

$$A_e^{max} = \frac{\lambda^2}{\int P(\theta,\phi) d\Omega}, \tag{3.4.10}$$

i.e. the maximum effective aperture is determined by the *shape* of the power pattern alone. The narrower the power pattern the higher the aperture efficiency. For a reflecting telescope,

$$\int P(\theta,\phi) d\Omega \sim \Theta^2_{HPBW} \sim (\frac{\lambda}{D})^2.$$

so

$$A_e^{max} \sim D^2.$$

The max. effective aperture scales like the geometric area of the reflector, as expected. Also from equation 3.4.10

$$A_e = A_e^{max} P(\theta,\phi) = \frac{\lambda^2 P(\theta,\phi)}{\int P(\theta,\phi) d\Omega}. \tag{3.4.11}$$

Comparing this with equation (3.4.1) gives the constant that relates the effective area to the directivity

$$D(\theta,\phi) = \frac{4\pi}{\lambda^2} A_e(\theta,\phi). \tag{3.4.12}$$

As an application for all these formulae, consider the standard communications problem of sending information from antenna 1 (gain $G_1(\theta,\phi)$, input power P_1) to antenna 2 (directivity $D_2(\theta^{'},\phi^{'})$), at distance R away. What is the power available at the terminals of antenna 2?

The flux density at antenna 2 is given by:

$$S = \frac{P_1}{4\pi R^2} G_1(\theta,\phi)$$

. i.e., the power falls off like R^2, but is not isotropically distributed. (The gain G_1 tells you how collimated the emission from antenna 1 is). The power available at the terminals of antenna 2 is:

$$W = A_{2e} S = \frac{P_1}{4\pi R^2} G_1(\theta,\phi) A_{2e}$$

substituting for the effective aperture from equation (3.4.12)

$$W = (\frac{\lambda}{4\pi R})^2 P_1 G_1(\theta,\phi) D_2(\theta^{'},\phi^{'})$$

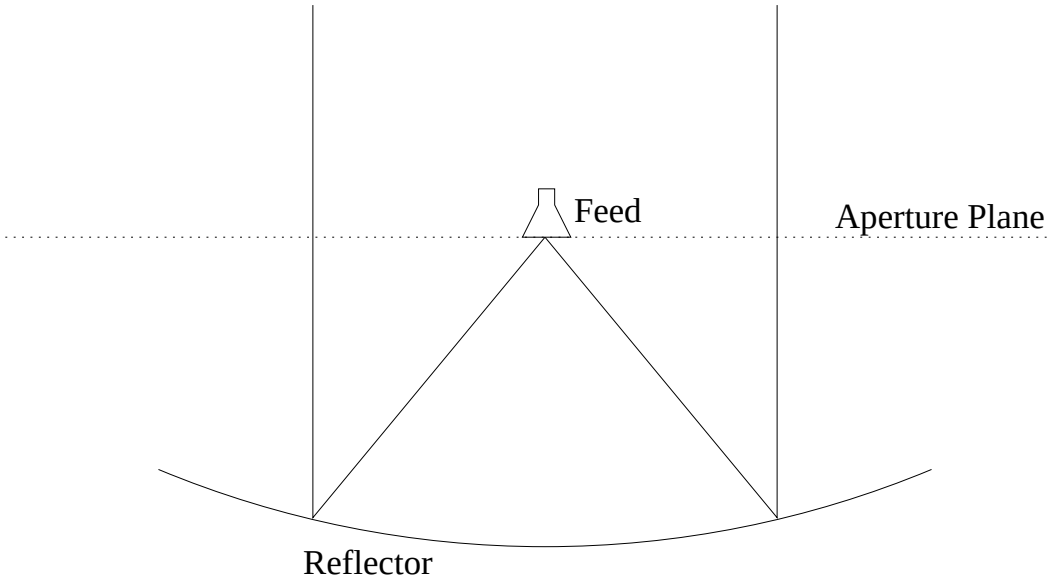


Figure 3.12: Aperture illumination for a parabolic dish.

This is called the **Friis transmission equation**. In Radar astronomy, there is a very similar expression for the power available at an antenna after bouncing off an unresolved target (the **radar range equation**). The major difference is that the signal has to make a round trip, (and the target reradiates power falling on it isotropically), so the received power falls like the fourth power of the distance to the target.

3.5 Computing Antenna Patterns

The next step is to understand how to compute the power pattern of a given telescope. Consider a parabolic reflecting telescope being fed by a feed at the focus. The radiation from the feed reflects off the telescope and is beamed off into space (Figure 3.12). If one knew the radiation pattern of the feed, then from geometric optics (i.e. simple ray tracing, see Chapter 19) one could then calculate the electric field on the plane across the mouth of the telescope (the 'aperture plane'). How does the field very far away from the telescope lookslike? If the telescope surface were infinitely large, then the electric field in the aperture plane is simply a plane wave, and since a plane wave remains a plane wave on propagation through free space, the far field is simply a plane wave traveling along the axis of the reflector. The power pattern is an infinitely narrow spike, zero everywhere except along the axis. Real telescopes are however finite in size, and this results in diffraction. The rigorous solution to the diffraction problem is to find the appropriate Green's function for the geometry, this is often impossible in practise and various approximations are necessary. The most commonly used one is Kirchoff's scalar diffraction theory. However, for our purposes, it is more than sufficient to simply use Huygen's principle.

Huygen's principle states that each point in a wave front can be regarded as an imaginary source. The wave at any other point can then be computed by adding together the contributions from each of these point sources. For example consider a one dimensional aperture, of length l with the electric field distribution ('aperture illumination') $e(x)$. The

field at a point P(R, θ) (Figure 3.13) due to a point source at a distance x from the center of the aperture is (if R is much greater than l) is:

$$dE = \frac{e(x)}{R^2} e^{-j\frac{2\pi x sin\theta}{\lambda}}$$

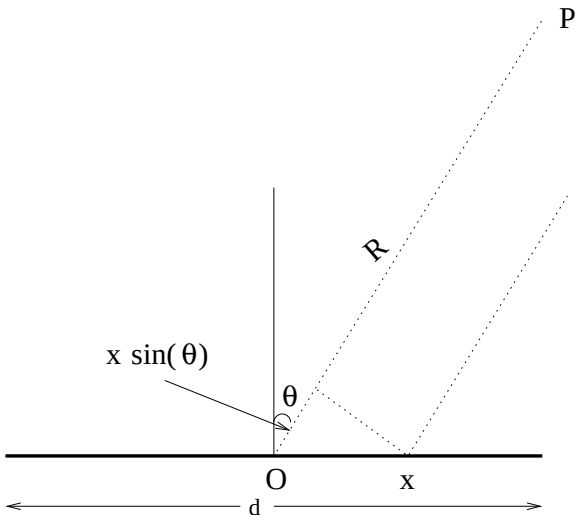


Figure 3.13: The far-field pattern as a function of the aperture illumination.

Where $x \sin\theta$ is simply the difference in path length between the path from the center of the aperture to the point P and the path from point x to point P. Since the wave from point x has a shorter path length, it arrives at point P at an earlier phase. The total electric field at P is:

$$E(R,\theta) = \int_{-l/2}^{l/2} \frac{e(x)}{R^2} e^{-jk\mu x} dx$$

where $k = \frac{2\pi}{\lambda}$ and $\mu = sin\theta$ and x is now measured in units of wavelength. The shape of the distribution is clearly independent of R, and hence the unnormalized power pattern F_U is just:

$$F_U(\mu) = \int_{-\infty}^{\infty} e_1(x) e^{-jk\mu x} dx \tag{3.5.13}$$

where

$$e_1(x) = e(x) \quad if \quad |x| \leq l/2 \quad ; \quad 0 \quad \text{otherwise}$$

The region in which the field pattern is no longer dependent on the distance from the antenna is called the **far field region**. The integral operation in equation (3.5.13) is called the **Fourier transform**. $F_U(\mu)$ is the Fourier transform of $e_1(x)$, which is often denoted as $F_U(\mu) = \mathbf{F}\big[e_1(x)\big]$. The Fourier transform has many interresting properties, some of which are listed below (see also Section 2.5).

1. Linearity

   If $G_1(\mu) = \mathbf{F}\big[g_1(x)\big]$ and $G_2(\mu) = \mathbf{F}\big[g_2(x)\big]$ then $G_1(\mu) + G_2(\mu) = \mathbf{F}\big[g_1(x) + g_2(x)\big]$.

2. Inverse

The Fourier transform is an invertible operation; if

$$G(\mu) = \int_{-\infty}^{\infty} g(x) e^{-j2\pi\mu x} dx$$

then

$$g(x) = \int_{-\infty}^{\infty} G(\mu) e^{j2\pi\mu x} d\mu$$

3. Phase shift

   If $G(\mu) = \mathbf{F}\big[g(x)\big]$ then $G(\mu - \mu_0) = \mathbf{F}\big[g(x)e^{-j2\pi\mu_0 x}\big]$. This means that an antenna beam can be steered across the sky simply by introducing the appropriate linear phase gradient in the aperture illumination.

4. Parseval's theorem

   If $G(\mu) = \mathbf{F}\big[g(x)\big]$, then

$$\int_{-\infty}^{\infty} |G(\mu)|^2 d\mu = \int_{-\infty}^{\infty} |g(x)|^2 dx$$

   This is merely a restatement of power conservation. The LHS is the power outflow from the antenna as measured in the far field region, the RHS is the power outflow from the antenna as measured at the aperture plane.

5. Area

   If $G(\mu) = \mathbf{F}\big[g(x)\big]$, then

$$G(0) = \int_{-\infty}^{\infty} g(x) dx$$

With this background we are now in a position to determine the maximum effective aperture of a reflecting telescope. For a 2D aperture with aperture illumination $g(x, y)$, from equation (3.4.10)

$$A_e^{max} = \frac{\lambda^2}{\int P(\theta, \phi) d\Omega} = \frac{\lambda^2}{\int |F(\theta, \phi)|^2 d\Omega} \tag{3.5.14}$$

But the field pattern is just the normalized far field electric field strength, i.e.

$$F(\theta, \phi) = \frac{E(\theta, \phi)}{E(0, 0)}$$

where $E(\theta, \phi) = \mathbf{F}\big[g(x, y)\big]$. From property (5)

$$E(0, 0) = \int g(x, y) dx dy' \tag{3.5.15}$$

and from Parseval's theorem,

$$\int |E(\theta, \phi)|^2 d\Omega = \int |g(x, y)|^2 dx dy \tag{3.5.16}$$

substituting in equation (3.5.14) using equations (3.5.15), 3.5.16 gives,

$$A_e^{max} = \frac{\lambda^2 \big| \int g(x, y) dx dy \big|^2}{\int |g(x, y)|^2 dx dy}$$

For uniform illumination

$$\frac{A_e^{max}}{\lambda^2} = \frac{A_g^2}{A_g} = A_g$$

Note that since x and y are in units of wavelength, so is A_g. A_e^{max} however is in physical units. Uniform illumination gives the maximum possible aperture efficiency (i.e. 1), because if the illumination is tapered then the entire available aperture is not being used.

As a concrete example, consider a 1D uniformly illuminated aperture of length l. The far field is then

$$E(\mu) = \int_{-l/2}^{l/2} e^{-\frac{j2\pi x\mu}{\lambda}} dx$$

$$= \frac{\lambda \sin(\pi l/\lambda\mu)}{\pi\mu}$$

and the normalized field pattern is

$$F(\mu) = \frac{\sin(\pi l/\lambda\mu)}{(\pi l/\lambda\mu)}$$

This is called a 1D sinc function. The 1st null is at $\mu = \lambda/l$, the 1st sidelobe is at $\mu = 3/2(\lambda/l)$ and is of strength $2/(3\pi)$. The strength of the power pattern 1st sidelobe is $(2/3\pi)^2 = 4.5\%$. This illustrates two very general properties of Fourier transforms:

1. the width of a function is inversely proportional to width of its transform (so large antennas will have small beams and small antennas will have large beams), and

2. any sharp discontinuities in the function will give rise to sidelobes ('ringing') in the fourier transform.

Figure 3.14 shows a plot of the the power and field patterns for a 700 ft, uniformly illuminated aperture at 2380 MHz.

Aperture illumination design hence involves the following following tradeoffs (see also Chapter 19):

1. A more tapered illumination will have a broader main beam (or equivalently smaller effective aperture) but also lower side lobes than uniform illumination.

2. If the illumination is high towards the edges, then unless there is a very rapid cutoff (which is very difficult to design, and which entails high sidelobes) there will be a lot of spillover.

Another important issue in aperture illumination is the amount of aperture blockage. The feed antenna is usually suspended over the reflecting surface (see Figure 3.3) and blocks out part of the aperture. If the illumination is tapered, then the central part of the aperture has the highest illumination and blocking out this region could have a drastic effect on the power pattern. Consider again a 1D uniformly illuminated aperture of length l with the central portion of length d blocked out. The far field of this aperture is (from the linearity of fourier transforms) just the difference between the far field of an aperture of length l and an aperture of length d, i.e.

$$E(\mu) \propto \frac{sin(\pi l\mu/\lambda)}{\pi\mu} - \frac{sin(\pi d\mu/\lambda)}{\pi\mu}$$

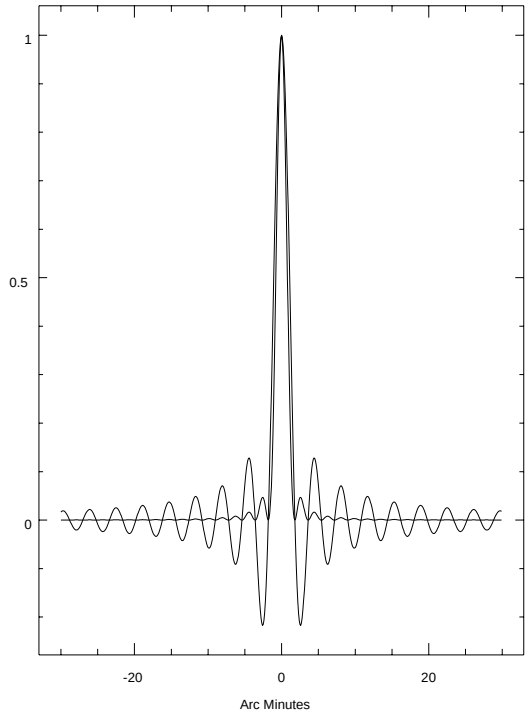


Figure 3.14: Power and field patterns for a 1D uniformly illuminated aperture.

or the normalized field pattern is:

$$F(\mu) = \frac{\lambda}{(l-d)}\left[\frac{sin(\pi l\mu/\lambda)}{\pi\mu} - \frac{sin(\pi d\mu/\lambda)}{\pi\mu}\right]$$

The field pattern of the "missing" part of the aperture has a broad main beam (since $d < l$). Subtracting this from the pattern due to the entire aperture will give a resultant pattern with very high sidelobes. In Figure 3.15 the solid curve is the pattern due to the entire aperture, the dotted line is the pattern of the blocked part and the dark curve is the resultant pattern. (This is for a 100ft blockage of a 700 ft aperture at 2380 MHz). Aperture blockage has to be minimized for a 'clean' beam, many telescopes have feeds offset from the reflecting surface altogether to eliminate all blockage.

As an example of what we have been discussing, consider the Ooty Radio Telescope (ORT) shown in Figure 3.16. The reflecting surface is a cylindrical paraboloid ($530m \times 30m$) with axis parallel to the Earth's axis. Tracking in RA is accomplished by rotating the telescope about this axis. Rays falling on the telescope get focused onto the a line focus, where they are absorbed by an array of dipoles. By introducing a linear phase shift across this dipole array, the antenna beam can be steered in declination (the "phase shift" property of Fourier transforms). The reflecting surface is only part of a paraboloid and does not include the axis of symmetry, the feed is hence completely offset, there is no blockage. The beam however is fan shaped, narrow in the RA direction (i.e. that conjugate to the $530m$ dimension) and broad in the DEC (i.e. that conjugate to the $30m$ dimension).

Aperture blockage is one of the reasons why an antenna's power pattern would deviate from what one would ideally expect. Another common problem that affects the power

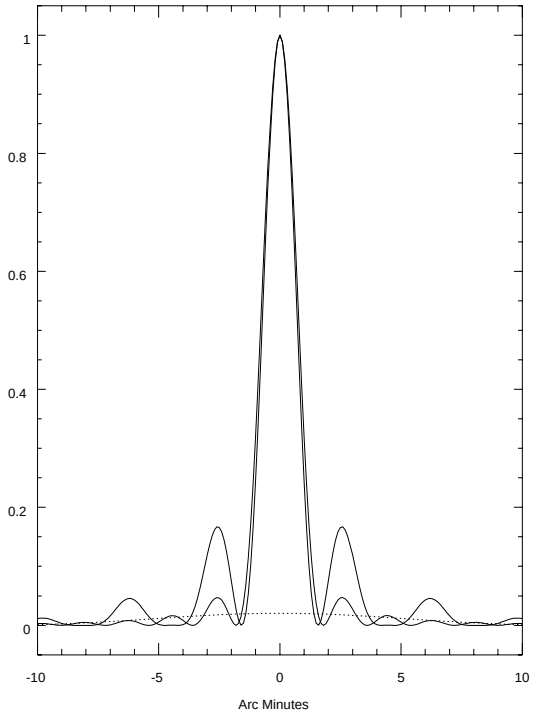


Figure 3.15: Effect of aperture blockage on the power pattern.

pattern is the location of the feed antenna. Ideally the feed should be placed at the focus, but for a variety of reasons, it may actually be displaced from the focus. For example, as the antenna tracks, the reflecting surface gets distorted and/or the feeds legs bend slightly, and for these reasons, the feed is displaced from the actual focal point of the reflector. In an antenna like the GMRT, there are several feeds mounted on a cubic turret at the prime focus, and the desired feed is rotated into position by a servo system (see Chapter 19). Small errors in the servo system could result in the feed pointing not exactly at the vertex of the reflector but along some slightly offset direction. This is illustrated in Figure 3.17. For ease of analysis we have assumed that the feed is held fixed and the reflector as a whole rotates. The solid line shows the desired location of the reflector (i.e. with the feed pointing at its vertex) while the dashed line shows the actual position of the reflector. This displacement between the desired and actual positions of the reflector results in an phase error (produced by the excess path length between the desired and actual reflector positions) in the aperture plane. From the geometry of Figure 3.17 this phase error can be computed, and from it the corresponding distortion in the field and power patterns can be worked out. Figure 3.18[A] shows the result of such a calculation. The principal effect is that the beam is offset slightly, but one can also see that its azimuthal symmetry is lost. Figure 3.18[B] shows the actual measured power pattern for a GMRT antenna with a turret positioning error. As can be seen, the calculated error pattern is a fairly good match to the observed one. Note that in plotting Figure 3.18[B] the offset in the power pattern has been removed (i.e. the power pattern has been measured with respect to its peak position).

Further Reading

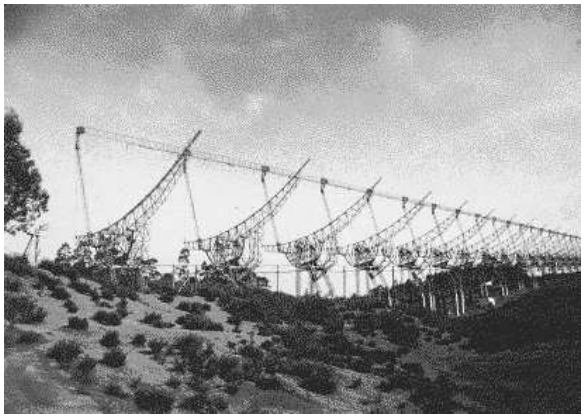
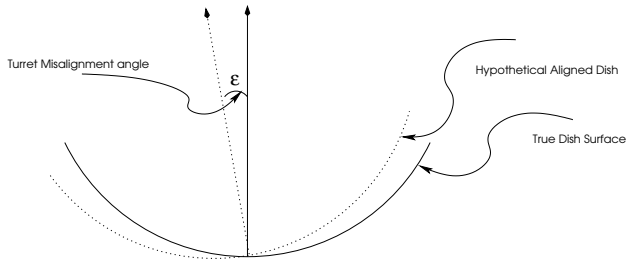


Figure 3.16: The Ooty radio telescope.



[A]

[B]

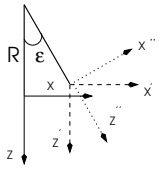


Figure 3.17: Turret positioning error. Ideally the feed should point at the vertex of the reflecting surface, but if the feed turret rotation angle is in error then the feed points along some offset direction.

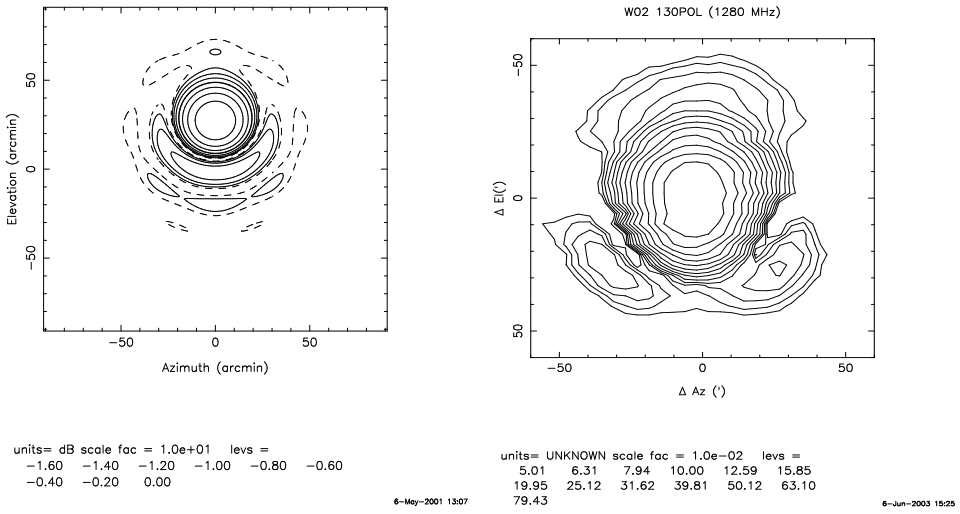
1. Antenna Theory *Analysis and Design* , Constantine A. Balanis, Harper & Row, Publishers, New York.

2. Radio telescopes, *second edition* , W. N. Christiansen & J. A. Hogbom, Cambridge Univ. Press.

3. Microwave Antenna Theory and Design, Samuel Silver (ed.), IEE

4. Reflector Antennas, A. W Love (ed.), IEEE press, *Selected Reprint Series*.

5. Instrumentation and Techniques for Radio Astronomy, Paul F. Goldsmith (ed.), IEEE press *Selected Preprint Series*.



(A) (B)

Figure 3.18: [A] Calculated beam pattern for a turret positioning error. [B] Measured beam pattern for a turret positioning error. The offset in the pattern has been removed, i.e. the power pattern has been measured with respect to its peak position.

Chapter 4

Two Element Interferometers

Jayaram N. Chengalur

4.1 Introduction

From the van-Cittert Zernike theorem (see Chapter 2) it follows that if one knows the mutual coherence function of the electric field, then the source brightness distribution can be measured[1]. The electric field from the cosmic source is measured using an antenna, which is basically a device for converting the electric field into a voltage that can then be further processed electronically (see Chapter 3). In this chapter we will examine exactly how the mutual coherence function is measured.

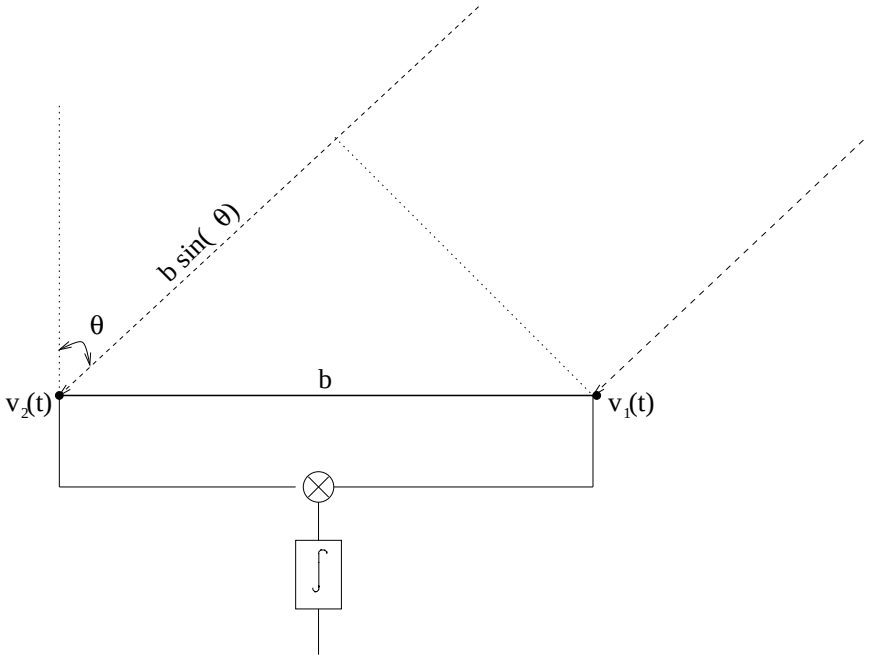


Figure 4.1: A basic two element interferometer

[1]Or in plain english, one make make an image of the source

We start by looking at the relationship between the output of a two element interferometer and the wanted mutual coherence function. Large interferometric arrays can be regarded as collections of two element interferometers, and for this reason it is instructive to understand in detail the working of a two element interferometer.

4.2 A Two Element Interferometer

Consider a two element interferometer shown in Figure 4.1. Two antennas $1, 2$ whose (vector) separation is **b**, are directed towards a point source of flux density S. The angle between the direction to the point source and the normal to the antenna separation vector is θ. The voltages that are produced at the two antennas due to the electric field from this point source are $v_1(t)$ and $v_2(t)$ respectively. These two voltages are multiplied together, and then averaged. Let us start by assuming that the radiation emitted by the source is monochromatic and has frequency ν. Let the voltage at antenna 1 be $v_1(t) = \cos(2\pi\nu t)$. Since the radio waves from the source have to travel an extra distance $b\sin\theta$ to reach antenna 2, the voltage there is delayed by the amount $b\sin\theta/c$. This is called the *geometric delay*, τ_g. The voltage at antenna 2 is hence $v_2(t) = \cos(2\pi\nu(t-\tau_g))$, where we have assumed that the antennas have identical gain. $r(\tau_g)$, the averaged output of the multiplier is hence:

$$\begin{aligned} r(\tau_g) &= \frac{1}{\mathrm{T}}\int_{t-T/2}^{t+T/2}\cos(2\pi\nu t)\cos(2\pi\nu(t-\tau_g))dt \qquad (4.2.1)\\ &= \frac{1}{\mathrm{T}}\int_{t-T/2}^{t+T/2}(\cos(4\pi\nu t - 2\pi\tau_g) + \cos(2\pi\nu\tau_g))dt \\ &= \cos(2\pi\nu\tau_g) \end{aligned}$$

where we have assumed that the averaging time T is long compared to $1/\nu$. The $\cos(4\pi\nu t)$ factor hence averages out to 0. As the source rises and sets, the angle θ changes. If we assume that the antenna separation vector, (usually called the *baseline vector* or just the *baseline*) is exactly east west, and that the source's declination $\delta_0 = 0$, then $\theta = \Omega_E t$, (where Ω_E is the angular frequency of the earth's rotation) we have:

$$r(\tau_g) = \cos(2\pi\nu \ \times \ b/c \ \times \ \sin(\Omega_E(t - t_z))) \qquad (4.2.2)$$

where t_z is the time at which the source is at the zenith. The output $r(\tau_g)$, (also called the *fringe*), hence varies in a quasi-sinusoidal form, with its instantaneous frequency being maximum when the source is at zenith and minimum when the source is either rising or setting (Figure 4.2).

Now if the source's right ascension was known, then one could compute the time at which the source would be at zenith, and hence the time at which the instantaneous fringe frequency would be maximum. If the fringe frequency peaks at some slightly different time, then one knows that assumed right ascension of the source was slightly in error. Thus, in principle at least, from the difference between the actual observed peak time and the expected peak time one could determine the true right ascension of the source. Similarly, if the source were slightly extended, then when the waves from a given point on the source arrive in phase at the two ends of the interferometer, waves arising from adjacent points on the source will arrive slightly out of phase. The observed amplitude of the fringe will hence be less than what would be obtained for a point source of the same total flux. The more extended the source, the lower the fringe amplitude[2]. For a

[2]assuming that the source has a uniform brightness distribution

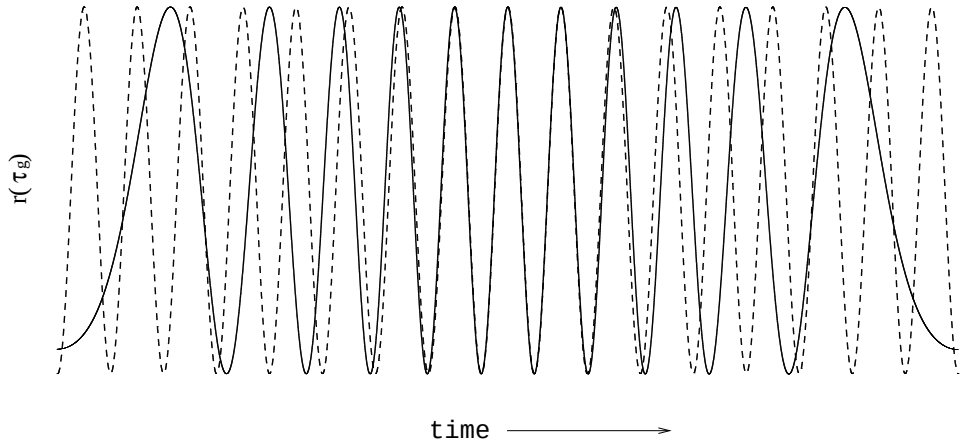


Figure 4.2: The output of a two element interferometer as a function of time. The solid line is the observed qausi-sinosoidal output (the *fringe*), the dotted line is a pure sinusoid whose frequency is equal to the peak instantaneous frequency of the fringe. The instantaneous fringe frequency is maximum when the source is at the zenith (the center of the plot) and is minimum when the source is rising (left extreme) or setting (right extreme).

sufficiently large source with smooth brigtness distribution, the fringe amplitude will be essentially zero[3]. In such circumstances, the interferometer is said to have *resolved out* the source.

Further, two element interferometers cannot distinguish between sources whose sizes are small compared to the fringe spacing, all such sources will appear as point sources. Equivalently when the source size is such that waves from different parts of the source give rise to the same phase lags (within a factor that is small compared to π), then the source will appear as a point source. This condition can be translated into a limit on $\Delta\theta$, the minimum source size that can be resolved by the interferometer, viz.,

$$\pi\nu\Delta\theta b/c \lesssim \pi \quad \Longrightarrow \quad \Delta\theta \lesssim \lambda/b$$

i.e., the resolution of a two element interferometer is $\sim \lambda/b$. The longer the baseline, the higher the resolution.

Observations with a two element interferometer hence give one information on both the source position and the source size. Interferometers with different baseline lengths and orientations will place different constraints on the source brightness, and the Fourier transform in the van Cittert-Zernike theorem can be viewed as a way to put all this information together to obtain the correct source brightness distribution.

4.3 Response to Quasi-Monochromatic Radiation

Till now we had assumed that the radiation from the source was monochromatic. Let us now consider the more realistic case of quasi-monochromatic radiation, i.e. the radiation

[3]This is related to the fact that in the double slit experiment, the interference pattern becomes less distinct and then eventually disappears as the source size is increased (see e.g. Born & Wolf, 'Principles of Optics', Sixth Edition, Section 7.3.4). In fact the double slit setup is mathematically equivalent to the two element interferometer, and much of the terminology in radio interferometry is borrowed from earlier optical terminology.

spectrum[4] contains all frequencies in a band $\Delta\nu$ around ν, with $\Delta\nu$ small compared to ν. If the radiation at some frequency ν arrives in phase at the two antennas in the interferometer, the radiation at some adjacent frequencies will arrive out of phase, and if $\Delta\nu$ is large enough, there will be frequencies at which the radiation is actually 180 degrees out of phase. Intuitively hence one would expect that averaging over all these frequencies would decrease the amplitude of the fringe. More rigorously, we have

$$\begin{aligned} r(\tau_g) &= \frac{1}{\Delta\nu}\int_{\nu-\frac{\Delta\nu}{2}}^{\nu+\frac{\Delta\nu}{2}} \cos(2\pi\nu\tau_g)d\nu \\ &= \frac{1}{\Delta\nu} Re\left[\int_{\nu-\frac{\Delta\nu}{2}}^{\nu+\frac{\Delta\nu}{2}} e^{i2\pi\nu\tau_g} d\nu\right] \\ &= \cos(2\pi\nu\tau_g)\left[\frac{\sin(\pi\Delta\nu\tau_g)}{\pi\Delta\nu\tau_g}\right] \end{aligned} \tag{4.3.3}$$

The quantity in square brackets, the sinc function, decreases rapidly with increasing bandwidth. Hence as one increases the bandwidth that is accepted by the telescope, the fringe amplitude decreases sharply. This is called *fringe washing*. However, since in order to achieve reasonable signal to noise ratio one would require to accept as wide a bandwidth as possible[5], it is necessary to find a way to average over bandwidth without losing fringe amplitude. To understand how this could be done, it is instructive to first look at what the fringe would be for a spatially extended source.

Let the direction vector to some reference point on the source be $\mathbf{s}_0$, and further assume that the source is small that it lies entirely on the tangent plane to the sky at $\mathbf{s_0}$, i.e. that the direction to any point on the source can be written as $\mathbf{s} = \mathbf{s_0} + \boldsymbol{\sigma}$, $\mathbf{s_0}.\boldsymbol{\sigma} = 0$, $\tau_g = \mathbf{s_0}.\mathbf{b}$. Then, from the van Cittert-Zernike theorem we have[6]:

$$\begin{aligned} r(\tau_g) &= Re\left[\int I(\mathbf{s}) e^{\frac{-i2\pi\mathbf{s}.\mathbf{b}}{\lambda}} d\mathbf{s}\right] \\ &= Re\left[e^{\frac{-i2\pi\mathbf{s_0}.\mathbf{b}}{\lambda}} \int I(\mathbf{s}) e^{\frac{-i2\pi\sigma.\mathbf{b}}{\lambda}} d\mathbf{s}\right] \\ &= |\mathcal{V}|\cos(2\pi\nu\tau_g + \Phi_{\mathcal{V}}) \end{aligned} \tag{4.3.4}$$

where $\mathcal{V}$, the complex *visibility* is defined as:

$$\mathcal{V} = |\mathcal{V}|e^{-i\Phi_{\mathcal{V}}} = \int I(\mathbf{s}) e^{\frac{2\pi\sigma.\mathbf{b}}{\lambda}} \tag{4.3.5}$$

The information on the source size and structure is contained entirely in $\mathcal{V}$, the factor $\cos(2\pi\nu\tau_g)$ in eqn. (4.3.4) only contains the information that the source rises and sets as the earth rotates. Since this is trivial and uninteresting, it can safely be suppressed. Conceptually, the way one could suppress this information is to introduce along the electrical signal path of antenna 1 an instrumental delay τ_i which is equal to τ_g. Then we will have $r(\tau_g) = |\mathcal{V}|\cos(\Phi_{\mathcal{V}})$, i.e. the fast fringe oscillation has been suppressed. One can then average over frequency and not suffer from fringe washing. Since τ_g changes with time as the source rises and sets, τ_i will also have to be continuously adjusted. This adjustment

[4] Radiation from astrophysical sources is inherently broadband. Radio telescopes however have narrow band filters which accept only a small part of the spectrum of the infalling radiation.

[5] See Chapter 5

[6] apart from some constant factor related to the gain of the antennas which we have been ignoring throughout.

of τ_i is called *delay tracking*. In most existing interferometers however, the process of preventing fringe washing is slightly more complicated than the conceptual scheme described above. The complication arises because delay tracking is usually done digitally in the baseband, i.e. after the whole chain of frequency translation operations described in Chapter 3. The geometric delay is however suffered by the incoming radiation, which is at the RF frequency.

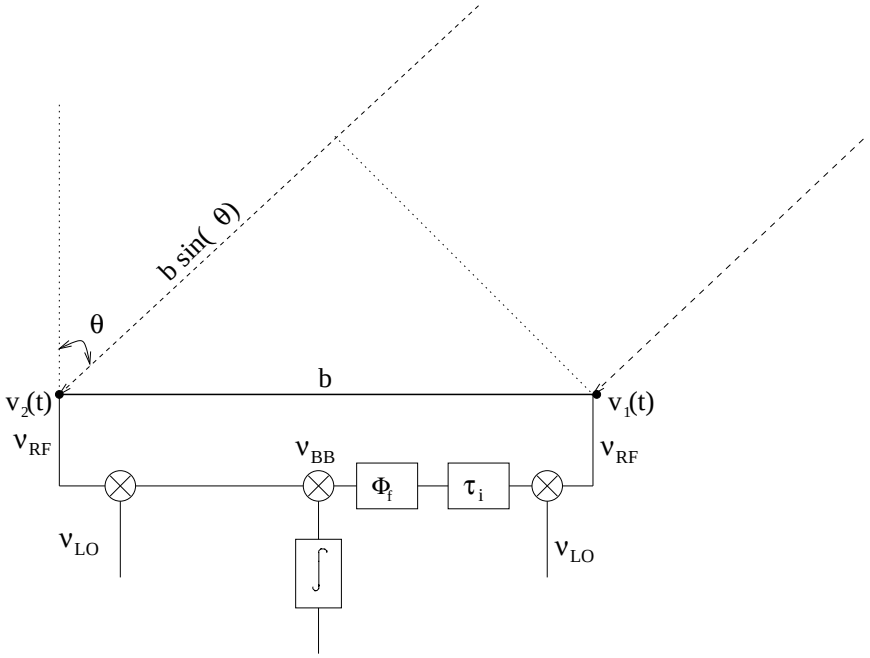


Figure 4.3: A two element interferometer with fringe stopping and delay tracking (see text).

4.4 Two Element Interferometers in Practice

To see this more clearly, let us consider the interferometer shown in Figure 4.3. The signals from antennas $1, 2$ are first converted to a frequency ν_{BB} using a mixer which is fed using a local oscillator of frequency[7] ν_{LO}, i.e. $\nu_{LO} = \nu_{RF} - \nu_{BB}$. Along the signal path for antenna 1 an additional instrumental delay $\tau_i = \tau_g + \Delta\tau$ is introduced, as is also a time varying phase shift Φ_f. The reasons for introducing this phase shift will be clear shortly. Then (see also equations 4.2.1 and 4.3.4) we have:

$$\begin{aligned} r(\tau_g) &= |\mathcal{V}| \left\langle \cos(\Phi_\mathcal{V} + 2\pi\nu_{BB}t - 2\pi\nu_{RF}\tau_g) \cos(2\pi\nu_{BB}(t-\tau_i) + \Phi_f) \right\rangle \quad (4.4.6) \\ &= |\mathcal{V}| cos(\Phi_\mathcal{V} + 2\pi(\nu_{RF} - \nu_{BB})\tau_g - \nu_{BB}\Delta\tau - \Phi_f) \\ &= |\mathcal{V}| \cos(\Phi_\mathcal{V} + 2\pi\nu_{LO}\tau_g - \nu_{BB}\Delta\tau - \Phi_f) \quad (4.4.7) \end{aligned}$$

[7]Note that it is important that the phase of the local oscillator signal be identical at the two antennas, i.e. the local oscillator signal has to be distributed in a phase coherent way to both antennas in the interferometer. Chapter 23 explains how this is acheived at the GMRT.

So, in order to compensate for all time varying phase factors, it is not sufficient to have $\tau_i = \tau_g$, one also needs to introduce a time varying phase $\Phi_f = 2\pi\nu_{LO}\tau_g$. This additional correction arises because the delay tracking is done at a frequency different from ν_{RF}. The introduction of the time varying phase is called *fringe stopping*. Fringe stopping can be achieved in a variety of ways. One common practice is to vary the instantaneous phase of the local oscillator signal in arm 1 of the interferometer by the amount Φ_f. Another possibility (which is the approach taken at the GMRT), is to digitally multiply the signal from antenna 1 by a sinusoid with the appropriate instantaneous frequency.

Another consequence of doing delay tracking digitally is that the geometric delay can be quantized only upto a step size which is related to the sampling interval with which the signal was digitized. In general therefore $\Delta\tau$ is not zero, and is called the *fractional sampling time error*. Correction for this error will be discussed in the Chapter 9.

The delay tracking and fringe stopping corrections apply for a specific point in the sky, viz. the position s_0. This point is called the phase tracking center[8]. Signals, such as terrestrial interference, which enter from the far sidelobes of the antennas do not suffer the same geometric delay τ_g as that suffered by the source. Consequently, delay tracking and fringe stopping introduces a rapidly varying change in the phase of these signals. On long baselines, where the fringe rate is rapid, the terrestrial interference could hence get completely decorrelated. While this may appear to be a terrific added bonus, in principle, terrestrial interference is usually so much stronger than the emission from cosmic sources, that even the residual correlation is sufficient to completely swamp out the desired signal.

We end this chapter by re-establishing the connection between what we have just done and the van Cittert-Zernike theorem. The first issue that we have to appreciate is that the van Cittert-Zernike theorem deals with the complex visibility, $\mathcal{V} = |\mathcal{V}|e^{-i\Phi_\mathcal{V}}$. However, the quantity that has been measured is $r(\tau_g) = |\mathcal{V}|\cos(-\Phi_\mathcal{V})$. If one could also measure $|\mathcal{V}|\sin(-\Phi_\mathcal{V})$, then of course one could reconstruct the full complex visibility. This is indeed what is done at interferometers. Conceptually, one has two multipliers instead of the one in Figure 4.3. The second multiplier is fed the same input as that in Figure 4.3, except that an additional phase difference of $\pi/2$ is introduced in each signal path. As can be easily verified, the output of this multiplier is $|\mathcal{V}|\sin(-\Phi_\mathcal{V})$. Such an arrangement of two multipliers is called a *complex correlator*. The two outputs are called the sine and cosine outputs respectively. For quasi-sinsoidal processes, one has to introduce a $\pi/2$ phase difference at each frequency present in the signal. The corresponding transformation is called a *Hilbert transform*[9]. How the complex correlator is achieved at the GMRT is described in Chapter 9. The output of the complex correlator is hence a single component of the Fourier transform of the source brightness distribution[10]. The component measured depends on the antenna separation as viewed from the source, i.e. $(\mathbf{b}.\mathbf{s_0})/\lambda$, which is also called *the projected baseline length*. For a large smooth source, the Fourier transform will be sharply peaked about the origin, and hence the visibility measured on long baselines will be small.

Further Reading

1. Thompson, R. A., Moran, J. M. & Swenson, G. W. Jr., 'Interferometry & Synthesis in Radio Astronomy', Wiley Interscience.

[8]For maximum sensitivity, one would also point the antennas such that their primary beam maxima are also at s_0.

[9]see Chapter 1

[10]This is true only if the antenna dimensions are neglected. Strictly speaking, the measured visibility is an average over the visibilities in the range $\mathbf{b}+\mathbf{a}$ to $\mathbf{b}-\mathbf{a}$ where a is the diameter of the antennas and b is the separation between their midpoints. As will be seen in Chapter 14 the fact that one has information on visibilities on scales smaller than b is useful when attempting to image large regions of the sky.

2. R. A. Perley, F. R. Schwab, & A. H. Bridle, eds., 'Synthesis Imaging in Radio Astronomy', ASP Conf. Series, vol. 6.

Chapter 5

Sensitivity and Calibration for Interferometers

Jayaram N. Chengalur

5.1 Sensitivity

As we discussed earlier, an aperture synthesis telescope can be regarded as a collection of two element interferometers. Hence, for understanding the sensitivity of such a telescope, it is easier to first start with the case of a two element interferometer. Consider such an interferometer composed of two antennas i, j, (of identical gains, but possibly different system temperatures), looking at a point source of flux density S. We assume that the point source is at the phase center[1] and hence that in the absence of noise the visibility phase is zero. Let the individual antenna gains[2] be G and system temperatures be T_{s_i} and T_{s_j}. If $n_i(t)$ and $n_j(t)$ are the noise voltages of antennas i and j respectively,then $\sigma_i^2 = \left\langle n_i^2(t) \right\rangle = \mathrm{T_{s_i}}$, and $\sigma_j^2 = \left\langle n_j^2(t) \right\rangle = \mathrm{T_{s_j}}$. Similarly if $v_i(t)$ and $v_j(t)$ are the voltages induced by the incoming radiation from the point source, $\left\langle v_i^2(t) \right\rangle = \left\langle v_j^2(t) \right\rangle = \mathrm{GS}$. The instantaneous correlator[3] output is given by:

$$r_{ij}(t) = \left(v_i(t) + n_i(t)\right)\left(v_j(t) + n_j(t)\right)$$

The mean[4] of the correlator output is hence:

$$\begin{aligned} \langle r_{ij}(t) \rangle &= \langle \left(v_i(t) + n_i(t)\right)\left(v_j(t) + n_j(t)\right) \rangle \\ &= \langle v_i(t) v_j(t) \rangle \\ &= \mathrm{GS} \end{aligned} \tag{5.1.1}$$

where we have assumed that the noise voltages of the two antennas are not correlated, and also of course that the signal voltages are not correlated with the noise voltages. $r_{ij}(t)$ is hence an unbiased estimator of the true visibility.

To determine the noise in the correlator output, we would need to compute the rms of $r_{ij}(t)$ for which we need to be able to work out:

[1] See Chapter 4.

[2] Here the gain is taken to be in units of Kelvin per Jansky of flux in the matched polarization

[3] Here we are dealing with an ordinary correlator, not the *complex correlator* introduced in the chapter on two element interferometers.

[4] Note that the average being taken over here is *ensemble* average, and *not* an average over time.

$$\langle r_{ij}(t)r_{ij}(t)\rangle = \langle (v_i+n_i)(v_j+n_j)(v_i+n_i)(v_j+n_j)\rangle$$

where for ease of notation we have stopped explicitly specifying that all voltages are functions of time. This quantity is not trivial to work out in general. However, if we assume that all the random processes involved are Gaussian processes[5] the complexity is considerably reduced because for Gaussian random variables the fourth moment can then be expressed in terms of products of the second moment. In particular[6], if $x_1,\ x_2,\ x_3,\ \&\ x_4$ have a joint gaussian distribution then:

$$\begin{aligned}\langle x_1x_2x_3x_4\rangle &= \langle x_1x_2\rangle\langle x_3x_4\rangle + \langle x_1x_3\rangle\langle x_2x_4\rangle + \\ &\quad \langle x_1x_4\rangle\langle x_2x_3\rangle \end{aligned} \tag{5.1.2}$$

Rather than directly computing $\langle r_{ij}(t)r_{ij}(t)\rangle$, it is instructive first to consider the more general quantity

$$\langle r_{ij}(t)r_{kl}(t)\rangle = \langle (v_i+n_i)(v_j+n_j)(v_k+n_k)(v_l+n_l)\rangle$$

viz. the cross-correlation between the outputs of interferometers (ij) and (kl). We have:

$$\begin{aligned}\langle r_{ij}(t)r_{kl}(t)\rangle &= \langle (v_i+n_i)(v_j+n_j)\rangle\langle (v_k+n_k)(v_l+n_l)\rangle + \\ &\quad \langle (v_i+n_i)(v_k+n_k)\rangle\langle (v_j+n_j)(v_l+n_l)\rangle + \\ &\quad \langle (v_i+n_i)(v_l+n_l)\rangle\langle (v_k+n_k)(v_j+n_j)\rangle \\ \\ &= (\langle v_iv_j\rangle + \langle n_i^2\rangle\delta_{ij})(\langle v_kv_l\rangle + \langle n_k^2\rangle\delta_{kl}) + \\ &\quad (\langle v_iv_k\rangle + \langle n_i^2\rangle\delta_{ik})(\langle v_jv_l\rangle + \langle n_j^2\rangle\delta_{jl}) + \\ &\quad (\langle v_iv_l\rangle + \langle n_i^2\rangle\delta_{il})(\langle v_kv_j\rangle + \langle n_k^2\rangle\delta_{kj}) \\ \\ &= (\mathrm{GS})^2 + \mathrm{GS}(\sigma_i^2\delta_{ij} + \sigma_k^2\delta_{kl}) + \sigma_i^2\delta_{ij}\sigma_k^2\delta_{kl} + \\ &\quad (\mathrm{GS})^2 + \mathrm{GS}(\sigma_i^2\delta_{ik} + \sigma_j^2\delta_{jl}) + \sigma_i^2\delta_{ik}\sigma_j^2\delta_{jl} + \\ &\quad (\mathrm{GS})^2 + \mathrm{GS}(\sigma_i^2\delta_{il} + \sigma_k^2\delta_{kj}) + \sigma_i^2\delta_{il}\sigma_k^2\delta_{kj}\end{aligned} \tag{5.1.3}$$

The case we are currently interested in is $\langle r_{ij}(t)r_{ij}(t)\rangle$, which from eqn(5.1.3) is:

$$\begin{aligned}\langle r_{ij}(t)r_{ij}(t)\rangle &= 3(\mathrm{GS})^2 + (\sigma_i^2+\sigma_j^2)\mathrm{GS} + \sigma_i^2\sigma_j^2 \\ &= 2(\mathrm{GS})^2 + (\mathrm{GS}+\mathrm{T}_{s_i})(\mathrm{GS}+\mathrm{T}_{s_j})\end{aligned} \tag{5.1.4}$$

To get the variance of $r_{ij}(t)$ we need to subtract the square of the mean of $r_{ij}(t)$ from the expression in eqn(5.1.4). Substituting for $\langle r_{ij}(t)\rangle^2$ from eqn(5.1.1) we have:

$$\sigma_{ij}^2 = (\mathrm{GS})^2 + (\mathrm{GS}+\mathrm{T}_{s_i})(\mathrm{GS}+\mathrm{T}_{s_j}) \tag{5.1.5}$$

Note that the angular brackets denote ensemble averaging. In real life of course one cannot do an ensemble average. Instead one does an average over time, i.e. we work in

[5] Recall from the discussion of sensitivity of a single dish telescope that the central limit theorem ensures that the signal and noise statistics will be well approximated by a Gaussian. This of course does not include 'systematics', like eg. interference, or correlator offsets because of bit getting stuck in the on or off mode etc.

[6] The derivation of this expression is particularly straightforward if one works with the moment generating function; see also the derivation sketched in Chapter 1.

terms of a time averaged correlator output $\bar{r}_{ij}(t)$, defined as

$$\bar{r}_{ij}(t) = \frac{1}{\mathrm{T}} \int_{t-\mathrm{T}/2}^{t+\mathrm{T}/2} r_{ij}(t^{'})dt^{'}$$

As can easily be verified, $\langle \bar{r}_{ij} \rangle = \langle r_{ij} \rangle$. However, computing the second moment, viz., $\bar{\sigma}_{ij}^2 = \langle \bar{r}_{ij} \bar{r}_{ij} \rangle - \langle \bar{r}_{ij} \rangle^2$ is slightly more tricky. It can be shown[7] that if $x(t)$ is a zero mean stationary process and that $\bar{x}(t)$ is the time average of $x(t)$ over the interval $(t-T/2, t+T/2)$, then

$$\bar{\sigma}_x^2 = \frac{1}{T} \int_{-\mathrm{T}/2}^{\mathrm{T}/2} \left(1 - \frac{|\tau|}{\mathrm{T}}\right) R_{xx}(\tau)\ d\tau \tag{5.1.6}$$

where $R_{xx}(\tau)$ is the auto-correlation function of $x(t)$, and $\bar{\sigma}$ is the variance of $x(t)$. Now, if $x(t)$ is a quasi-sinusoidal process with bandwidth $\Delta\nu$, then the integral of $R_{xx}(\tau)$ will be negligible outside the coherence time $1/\Delta\nu$. Further, if $T >> 1/\Delta\nu$, then the factor in parenthesis in eqn(5.1.6) can be taken to be ~ 1 for $\tau < 1/\Delta\nu$. Hence we have:

$$\begin{aligned} \bar{\sigma}_x^2 &\simeq \frac{1}{T} \int_{-\mathrm{T}/2}^{\mathrm{T}/2} R_{xx}(\tau)\ d\tau \quad \simeq \frac{1}{T} \int_{-\infty}^{\infty} R_{xx}(\tau)\ d\tau \\ &= \frac{1}{T} \mathrm{S}_{xx}(0) \qquad\qquad = \frac{1}{T} \frac{\sigma_x^2}{2\Delta\nu} \end{aligned} \tag{5.1.7}$$

where $\mathrm{S}_{xx}(\nu) = \sigma_x^2/2\Delta\nu$ is the power spectrum[8] of $x(t)$. From eqn(5.1.7) and eqn(5.1.5) we hence have

$$\bar{\sigma}_{ij}^2 = \frac{1}{2\mathrm{T}\Delta\nu} \left((\mathrm{GS})^2 + (\mathrm{GS} + \mathrm{T}_{s_i})(\mathrm{GS} + \mathrm{T}_{s_j}) \right) \tag{5.1.8}$$

Putting all this together we get that the signal to noise ratio of a two element interferometer is given by:

$$\mathrm{snr} = \frac{(\sqrt{2\mathrm{T}\Delta\nu}\mathrm{GS})}{\sqrt{(\mathrm{GS})^2 + (\mathrm{GS} + \mathrm{T}_{s_i})(\mathrm{GS} + \mathrm{T}_{s_j})}} \tag{5.1.9}$$

There are two special cases which often arise in practice. The first is when the source is weak, i.e. $\mathrm{GS} \ll \mathrm{T}_s$. In this case the snr becomes

$$\mathrm{snr} = \frac{(\sqrt{2\mathrm{T}\Delta\nu}\mathrm{GS})}{\sqrt{\mathrm{T}_{s_i}\,\mathrm{T}_{s_j}}} \tag{5.1.10}$$

For a single dish with the collecting area equal to the sum of the collecting areas of antennas i and j (i.e. with gain 2G), and with system temperature $\mathrm{T}_s = \sqrt{\mathrm{T}_{\mathrm{s_i}}\,\mathrm{T}_{\mathrm{s_j}}}$ the signal to noise would have been a factor of $\sqrt{2}$ better[9]. The loss of signal to noise in the two element interferometer is because one does not measure the auto-correlations of antennas i and j. Only their cross-correlation has been measured. In a sigle dish one would have effectively measured the cross-correlation as well as the auto-correlations.

[7] Papoulis, 'Probability, Random Variables & Stochastic Processes', Third Edition, Chapter 10

[8] Where we have made the additional assumption that $x(t)$ is a white noise process, i.e. that its spectrum is flat. The power spectrum for such processes is easily derived from noting that $\int_{-\infty}^{\infty} \mathrm{S}_{xx}(\nu)d\nu = \sigma_x^2$, and that for a quasi-sinosoidal proccess of bandwidth $\Delta\nu$, the integrand is non zero only over an interval $2\Delta\nu$ (including the negative frequencies).

[9] As you can easily derive from eqns 5.1.1 and 5.1.3 by putting $i = j = k = l$. Note that in this case eqn 5.1.1 becomes $\langle r_{ii}(t) \rangle = (v_i(t) + n_i(t))\ (v_i(t) + n_i(t)) = 2\mathrm{GS} + \mathrm{T}_s$

The other special case of interest is when the source is extremely bright, i.e. $\mathrm{GS} \gg \mathrm{T}_s$. In this case, the signal to noise ratio is:

$$\mathrm{snr} = \frac{(\sqrt{2\mathrm{T}\Delta\nu}\mathrm{GS})}{\sqrt{2(\mathrm{GS})^2}} = \sqrt{\mathrm{T}\Delta\nu} \tag{5.1.11}$$

This is as expected, because for very bright sources, one is limited by the Poisson fluctuations of the source brightness, and hence one would expect the signal to noise ratio to go as the square root of the number of independent measurements. Since one gets an independent measurement every $1/\Delta\nu$ seconds, the total number of independent measurements in a time T is just $\mathrm{T}\Delta\nu$.

Having derived the signal to noise ratio for a two element interferometer, let us now consider the case of an N element interferometer. This can be considered as ${}^{N}C_2$ two element interferometers. Let us take the case where the source is weak. Then from eqn(5.1.3) the correlation between $r_{12}(t)$ and $r_{13}(t)$ is given by

$$\begin{aligned} \langle r_{12}(t) r_{13}(t) \rangle &= \sigma_1^2 \delta_{12} \sigma_1^2 \delta_{13} + \sigma_1^2 \delta_{13} \sigma_1^2 \delta_{21} + \sigma_1^2 \delta_{11} \sigma_2^2 \delta_{23} \\ &= 0 \end{aligned} \tag{5.1.12}$$

The outputs are uncorrelated, even though these two interferometers have one antenna in common[10]. Similarly, one can show that (as expected) the outputs of two two-element interferometers with no antenna in common are uncorrelated. Since the r_{ij}'s are all uncorrelated with one another, the rms noise can simply be added in quadrature. In particular, for an N element array, where all the antennas are identical and have the same system temperature, the signal to noise ratio while looking at a weak source is:

$$\underline{\mathrm{snr} = \sqrt{\mathrm{N}(\mathrm{N}-1)\mathrm{T}\Delta\nu}\ \mathrm{GS}} \over \mathrm{T_s} \tag{5.1.13}$$

This is the fundamental equation[11] that is used to estimate the integration time required for a given observation. The signal to noise ratio for an N element interferometer is less than what would have been expected for a single dish telescope with area N times that of a single element of the interferometer, but only by the factor $\mathrm{N}/\sqrt{\mathrm{N}(\mathrm{N}-1)}$. The lower sensitivity is again because the N auto-correlations have not been measured. For large N however, this loss of information is negligible. For the GMRT, $\mathrm{N} = 30$ and $\mathrm{N}/\sqrt{\mathrm{N}(\mathrm{N}-1)} = 1.02$, hence the snr is essentially the same as that of a single dish with 30 times the collecting area of a single GMRT dish.

For a complex correlator[12], the analysis that we have just done holds separately for the cosine and sine channels of the correlator. If we call the outputs of such a correlator r_{ij}^c and r_{ij}^s then it can be shown that the noise in r_{ij}^c and r_{ij}^s is uncorrelated. Further since the time averaging can be regarded as the adding together of a large number of independent samples ($\sim \sqrt{T\Delta\nu}$), from the central limit theorem, the statistics of the noise in $\bar{r}_{ij}^c$ and $\bar{r}_{ij}^s$ are well approximated as Gaussian. It is then possible to derive the statistics of functions of $\bar{r}_{ij}^c$ and $\bar{r}_{ij}^s$, such as the visibility amplitude ($\sqrt{\bar{r}_{ij}^c + \bar{r}_{ij}^s}$) and the visibility phase ($\tan^{-1} \bar{r}_{ij}^s / \bar{r}_{ij}^c$). For example, it can be shown that the visibility amplitude has a Rice distribution[13]

[10] This may seem counter intuitive, but note that the outputs are only uncorrelated, they are not independent.

[11] In some references, an efficiency factor η is introduced to account for degradation of signal to noise ratio because of the noise introduced by finite precision digital correlation etc. This factor has been ignored here, or equivalently one can assume that it has been absorbed into the system temperature.

[12] See the chapter on two element interferometers

[13] Papoulis, 'Probability, Random Variables & Stochastic Processes', Third Edition, Chapter 6.

For an extended source, the entire analysis that we have done continues to hold, with the exception that S should be treated as the correlated part of the source flux density. For example, at low frequencies, the Galactic background is often much larger than the receiver noise and one would imagine that the limiting case of large source flux density (i.e. eqn(5.1.11) is applicable. However, since this background is largely resolved out at even modest spacings, its only effect is an increase in the system temperature.

Finally we look at the noise in the image plane, i.e. after Fourier transformation of the visibilities. Since most of the astronomical analysis and interpretation will be based on the image, it is the statistics in the image plane that is usually of interest. The intensity at some point (l, m) in the image plane is given by:

$$I(l,m) = \frac{1}{\mathrm{M}} \sum_p w_p \mathcal{V}_p e^{-i2\pi(lu_p + mv_p)}$$

where w_p is the weight[14] given to the pth visibility measurement $\mathcal{V}_p$, and there are a total of M independent measurements. The cross-correlation function in the image plane, $\left\langle I(l,m) I(l^{'}, m^{'}) \right\rangle$ is hence:

$$\left\langle I(l,m) I(l^{'}, m^{'}) \right\rangle = \frac{1}{\mathrm{M}^2} \sum_p \sum_q w_p w_q \left\langle \mathcal{V}_p \mathcal{V}_q^* \right\rangle e^{-i2\pi(lu_p + mv_p)} e^{i2\pi(l^{'} u_q + m^{'} v_q)}$$

In the absence of any sources, the visibilities are uncorrelated with one another, and hence, we have

$$\left\langle I(l,m) I(l^{'}, m^{'}) \right\rangle = \frac{1}{\mathrm{M}^2} \sum_m w_p^2 \sigma_p^2 e^{-i2\pi((l-l^{'})u_p + (m-m^{'})v_p)}$$

Hence in the case that all the noise on each measurement is the same, and that the weights given to each visibility point is also the same, (i.e. uniform tapering), the correlation in the map plane has exactly the same shape as the dirty beam. Further the variance in image plane would then be $\sigma_{\mathcal{V}}^2/\mathrm{M}$, where $\sigma_{\mathcal{V}}^2$ is the noise on a single visibility measurement. This is equivalent to eqn(5.1.13), as indeed it should be.

Because the noise in the image plane has a correlation function shaped like the dirty beam, one can roughly take that the noise in each resolution element is uncorrelated. The expected statistics after simple image plane operations (like smoothing) can hence be worked out. However, after more complicated operations, like the various possible deconvolution operations, the statistics in the image plane are not easy to derive.

5.2 Calibration

We have assumed till now that we have been working with calibrated visibilities, i.e. free from all instrumental effects (apart from some additive noise component). In reality, the correlator output is different from the true astronomical visibility for a variety of reasons, to do with both instrumental effects as well as propagation effects in the earth's atmosphere and ionosphere.

At low frequencies, it is the effect of the ionosphere that is most dominant. As is discussed in more detail in Chapter 16, density irregularities cause phase irregularities in the wavefront of the incoming radio waves. One would expect therefore that the image

[14] As discussed in Chapter 11, this weight is in general a combination of weights chosen from signal to noise ratio considerations and from synthesized beam shaping considerations.

of the source would be distorted in the same way that atmospheric turbulence ('seeing') distorts stellar images at optical wavelengths. To first order this is true, but for the ionosphere the 'seeing disk' is generally smaller than the diffraction limit of typical interferometers. There are two other effects however which are more troublesome. The first is 'scintillation', where because of diffractive effects the flux density of the source changes rapidly – the flux density modulation could approach 100%. The other is that slowly varying, large scale refractive index gradients cause the apparent source position to wander. At low frequencies, the source position could often wander by several arc minutes, i.e. considerably more than the synthesized beam. As we shall see below, provided the time scale of this wander is slow enough, it can be corrected for.

Let us take the case where the effect of the ionosphere is simply to produce an excess path length, i.e. for an antenna i let the excess phase[15] for a point source at sky position (l, m) be $\phi_i(l, m, t)$, where we have explicitly put in a time dependence. Then the observed visibility on a baseline (i, j) would be

$$\widetilde{\mathcal{V}}_{ij}(t) = G_{ij}(t) \int e^{-i(\phi_i(l,m,t)-\phi_j(l,m,t))} I(l,m) e^{-i2\pi(lu_{ij}+mv_{ij})} \tag{5.2.14}$$

where $I(l, m)$ is the sky brightness distribution and we have ignored the primary beam[16]. $G_{ij}(t)$ is 'instrumental phase', i.e. the phase produced by the amplifiers, transmission lines, or other instrumentation along the signal path. If $\phi_i(l, m, t)$ were some general, unknown function of (l, m, t) it would not be possible to reconstruct the true visibility from the measured one. However, since the size scale of ionospheric disturbances is $\sim$ a few hundred kilometers, it is often the case that $\phi_i(l, m, t)$ is constant over the entire primary beam, i.e. there is no (l, m) dependence. The source is then said to lie within a single *iso-planatic patch*. In such situations, the ionospheric phase can be taken out of the integral, and eqn(5.2.14) reduces to:

$$\widetilde{\mathcal{V}}_{ij}(t) = G_{ij}(t) e^{-i(\phi_i(t)-\phi_j(t))} \int I(l,m) e^{-i2\pi(lu_{ij}+mv_{ij})} \tag{5.2.15}$$

If it also the case that the ionospheric and instrumental gains are changing slowly, then they can be calibrated in the following manner. Suppose that close to the source of interest, there is a calibration source whose true visibility $\mathcal{V}^c_{ij}$ is known. Then one could intersperse observations of the target source with observations of the calibrator. For the calibrator, dividing the observed visibility $\widetilde{\mathcal{V}}^c_{ij}(t)$ by the (known) true visibility, $\mathcal{V}^c_{ij}(t)$ one can measure the factor $G_{ij}(t)e^{-i(\phi_i(t)-\phi_j(t))}$. This can then be applied as a correction to the visibilities of the target source. For slightly better corrections, one could interpolate in time between calibrator observations. This is the basis of what is sometimes called 'ordinary' calibration. The calibrator source is usually an isolated point source, although this is not, strictly speaking, necessary. It is sufficient to know the true visibilities $\mathcal{V}^c_{ij}(t)$. Note that if the calibrators absolute flux is also known, then this calibration procedure will also calibrate the amplitude scale of the target source[17].

In the approach outlined above, in order to calibrate the data one needs to solve for an unknown complex number per baseline, (i.e. N(N-1)/2 complex numbers for an N element interferometer). If we assume that the correlator itself does not produce any errors[18], i.e. that all the instrumental errors occur in the antennas or the transmission lines, then the

[15] by which we mean the phase difference over what would have been obtained in the absence of the ionosphere

[16] i.e. we have set the factor $B(l, m)/\sqrt{1 - l^2 - m^2}$ to 1.

[17] provided, as we will discuss in more detail later, that the system temperature does not differ for the target source and the calibrator

[18] which is often a good assumption for digital correlators

instrumental gain can be written out as antenna based terms, i.e.

$$G_{ij}(t) = g_i(t)g_j^*(t) \tag{5.2.16}$$

where $g_i(t)$ and $g_j(t)$ are the complex gains along the signal paths from antennas 1 and 2. But the ionospheric phase can also be decomposed into antenna based quantities (see eqn 5.2.15), and can hence be lumped together with the instrumental phase. Consequently the total unknown complex gains that have to be solved for reduces from N(N-1)/2 to N, which can be a dramatic reduction for large N. (For the GMRT it is a reduction from 435 unknowns to 30 unknowns).

However to appreciate the real power of this decomposition into antenna based gains, consider the following quantities. First let us look at the sum of the phases of the raw visibilities $\widetilde{\mathcal{V}}_{12}$, $\widetilde{\mathcal{V}}_{23}$ and $\widetilde{\mathcal{V}}_{31}$. If we call the true visibility phase $\psi_{\mathcal{V}_{ij}}$, the raw visibility phase $\psi_{\widetilde{\mathcal{V}}_{ij}}$ and the sum of the instrumental and ionospheric phases χ_i, then we have

$$\begin{aligned} \psi_{\widetilde{\mathcal{V}}_{12}} + \psi_{\widetilde{\mathcal{V}}_{23}} + \psi_{\widetilde{\mathcal{V}}_{31}} &= \chi_1 - \chi_2 + \psi_{\mathcal{V}_{12}} + \chi_2 - \chi_3 + \psi_{\mathcal{V}_{12}} + \chi_3 - \chi_1 + \psi_{\mathcal{V}_{31}} \\ &= \psi_{\mathcal{V}_{12}} + \psi_{\mathcal{V}_{23}} + \psi_{\mathcal{V}_{31}} \end{aligned} \tag{5.2.17}$$

i.e. over any triangle of baselines the sum of the phases of the raw visibilities is the true source visibility. This is called *phase closure*. Similarly it is easy to show that for any baselines 1,2,3,4, the ratio of the raw visibilities will be the same as the true visibilities, i.e.

$$\frac{|\widetilde{\mathcal{V}}_{12}||\widetilde{\mathcal{V}}_{34}|}{|\widetilde{\mathcal{V}}_{23}||\widetilde{\mathcal{V}}_{41}|} = \frac{|\mathcal{V}_{12}||\mathcal{V}_{34}|}{|\mathcal{V}_{23}|\mathcal{V}_{41}|} \tag{5.2.18}$$

This is called *amplitude closure*. For an N element interferometer, we have $1/2N(N-1) - (N-1)$ constraints on the phase and $1/2N(N-1) - N$ constraints on the amplitude. For large N, this is considerably more than the N unknown gains that one is solving for. The large number of available constraints means that the following iterative scheme would work.

1. Choose a suitable starting model for the brightness distribution. Compute the model visibilities.
2. For this model, solve for the antenna gains, subject to the closure constraints.
3. Apply these gain corrections to the visibility data, use the corrected data to make a fresh model of the brightness distribution.

For arrays with sufficient number of antennas, convergence is usually rapid. Note however, for this to work, the signal to noise ratio per visibility point[19] has to be reasonable, i.e. 2-3. This is often the case at low frequencies, and this technique of determining antenna gains (which is called *self calibration*) is usually highly successful.

Note that if one adds a phase $\chi_i = 2\pi(l_0u_i + m_0v_i)$ to each antenna (where $l_0,\ m_0$ are arbitrary and (u_i, v_i) are the (u,v) co-ordinates of the ith antenna), the phase closure constraints (eqn 5.2.17) continue to be satisfied. That means that in self calibration the phases can be solved only upto a constant phase gradient across the uv plane, i.e. the absolute source position is lost. Similarly, it is easy to see that the amplitude closure constraints will be satisfied even if one multiplies all the gains by a constant number, i.e. in self calibration one loses information on the absolute source flux density . The only way to determine the absolute source flux density is to look at a calibrator of known flux.

[19]Actually strictly speaking one means the signal to noise ratio over an interval for which the ionospheric phase can be assumed to be constant

Since antenna gains and system temperatures are usually stable over several hours[20], it is usually sufficient to do this calibration only once during an observing run. A more serious problem at low frequencies is that the Galactic background (whose strength varies with location on the sky) makes a significant contribution to the system temperature. Hence, when attempting to measure the source flux density, it is important to correct for the fact that the system temperature is different for the calibrator source as compared to the target source. The system temperature can typically be measured on rapid time scales by injecting a noise source of known strength at the front end amplifier.

Another related way (to selfcal) of solving for the system gains is the following. Suppose that the visibility on baselines (i, j) and (k, l) are identical. Then the ratio of the measured visibilities is directly related to the ratio of the complex instrumental gains of antennas i, j, k & l. If there are enough number of such 'redundant' baselines, one could imagine solving for the instrumental gains. Some arrays, like the WSRT have equispaced antennas, giving rise to a very large number of redundant baselines, and this technique has been successfuly used to calibrate complex sources[21] For a simple source, like a point source, all possible baselines are redundant, and this technique reduces essentially to self-calibration.

At the very lowest frequencies ($\nu < 200$ MHz, roughly for the GMRT) the assumption that the source lies within the iso-planatic patch probably begins to break down. The simple self calibration scheme outlined above will stop working in that regime. A possible solution then, is to solve (roughly speaking) for the phase changes produced by each iso-planatic patch. Often the primary beams of several antennas will pass through the same iso-planatic patches, so the extra number of degrees of freedom introduced will not be substantial, and an iterative approach to solving for the unknowns will probably converge[22].

5.3 Further Reading

1. Hamaker J. P., O'Sullivan, J. D. & Noordam, J. E., Journal of the Opt. Soc. Of America, **67**, 1122.

2. Thompson, R. A., Moran, J. M. & Swenson, G. W. Jr., 'Interferometry & Synthesis in Radio Astronomy', Wiley Interscience.

3. R. A. Perley, F. R. Schwab, & A. H. Bridle, eds., 'Synthesis Imaging in Radio Astronomy'

[20] Or change in a predictable manner with changing azimuth and elevation of the antennas

[21] see Noordam, J. E. & de Bruyn A. G., 1982, Nature **299**, 597.

[22] See Subrahmanya, C. R., (in 'Radio Astronomical Seeing', J. E. Baldwin & Wang Shouguan eds.) for more details

Chapter 6

Phased Arrays

Yashwant Gupta

6.1 Introduction

A single element telescope with a steerable paraboloidal reflecting surface is the simplest kind of radio telescope that is commonly used. Such a telescope gives an angular resolution $\sim \lambda/D$, where D is the diameter of the aperture and λ is the wavelength of observation. For example, for a radio telescope of 100 m diameter (which is about the largest that is practically feasible for a mechanically steerable telescope), operating at a wavelength of 1 m, the resolution is $\sim 30\ arc\ min$. This is a rather coarse resolution and is much less than the resolution of ground based optical telescopes.

Use of antenna arrays is one way of increasing the effective resolution and collecting area of a radio telescope. An array usually consists of several discrete antenna elements arranged in a particular configuration. Most often this configuration produces an unfilled aperture antenna, where only part of the overall aperture is filled by the antenna structure. The array elements can range in complexity from simple, fixed dipoles to fully steerable, parabolic reflector antennas. The outputs (voltage signals) from the array elements can be combined in various ways to achieve different results. For example, the outputs may be combined, with appropriate phase shifts, to obtain a single, total power signal from the array – such an array is generally referred to as a phased array. If the outputs are multiplied in distinct pairs in a correlator and processed further to make an image of the sky brightness distribution, the array is generally referred to as a correlator array (or an interferometer). Here we will primarily be concerned with the study of phased arrays, with direct comparison of the performance with correlator arrays, where relevant.

6.2 Array Theory

6.2.1 The 2 Element Array

We begin by deriving the far field radiation pattern for the case of the simplest array, two isotropic point source elements separated by a distance d, as shown in Figure 6.1. The net far field in the direction θ is given as

$$E(\theta) \;=\; E_1\, e^{j\psi/2} \;+\; E_2\, e^{-j\psi/2} \quad , \tag{6.2.1}$$

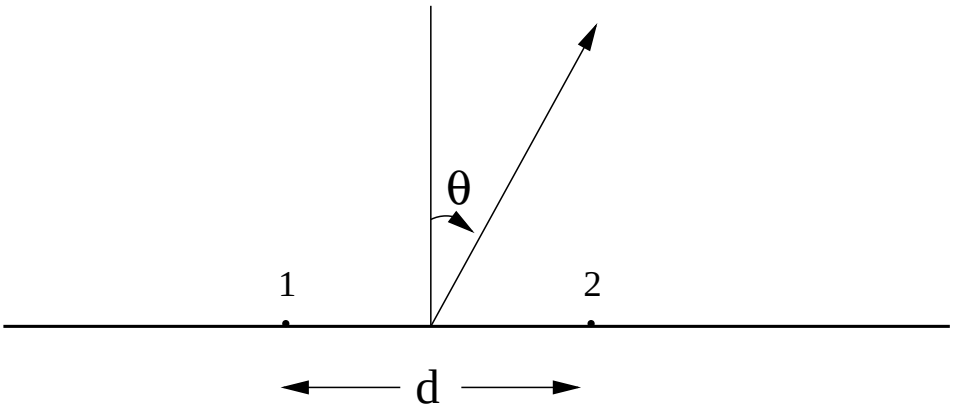


Figure 6.1: Geometry for the 2 element array.

where $\psi = k\,d\,\sin\theta + \delta$, $k = 2\pi/\lambda$ is the wavenumber and δ is the intrinsic phase difference between the two sources. E_1 and E_2 are the amplitudes of the electric field due to the two sources, at the distant point under consideration. The reference point for the phase, referred to as the phase centre, is taken halfway between the two elements. If the two sources have equal strength, $E_1 = E_2 = E_0$ and we get

$$E(\theta) \;=\; 2\,E_0\,\cos(\psi/2) \tag{6.2.2}$$

The power pattern is obtained by squaring the field pattern. By virtue of the reciprocity theorem[1], $E(\theta)$ also represents the voltage reception pattern obtained when the signals from the two antenna elements are added, after introducing the phase shift δ between them.

For the case of $\delta = 0$ and $d \gg \lambda$, the field pattern of this array shows sinusoidal oscillations for small variations of θ around zero, with a period of $2\lambda/d$. Non-zero values of δ simply shift the phase of these oscillations by the appropriate value.

If the individual elements are not isotropic but have identical directional patterns, the result of eqn 6.2.2 is modified by replacing E_0 with the element pattern, $E_i(\theta)$. The final pattern is given by the product of this element pattern with the $\cos(\psi/2)$ term which represents the array pattern. This brings us to the important principle of pattern multiplication which can be stated as : The total field pattern of an array of nonisotropic but similar elements is the product of the individual element pattern and the pattern of an array of isotropic point sources each located at the phase centre of the individual elements and having the same relative amplitude and phase, while the total phase pattern is the sum of the phase patterns of the individual elements and the array of isotropic point sources. This principle is used extensively in deriving the field pattern for complicated array configurations, as well as for designing array configurations to meet specified field pattern requirements (see the book on "Antennas" by J.D. Kraus (1988) for more details).

6.2.2 Linear Arrays of n Elements of Equal Amplitude and Spacing :

We now consider the case of a uniform linear array of n elements of equal amplitude, as shown in Figure 6.2. Taking the first element as the phase reference, the far field pattern is given by

$$E(\theta) \;=\; E_0\,\left[1 \;+\; e^{j\psi} \;+\; e^{j2\psi} \;+\; \ldots \;+\; e^{j(n-1)\psi}\right] \quad , \tag{6.2.3}$$

[1] see Chapter 3

where $\psi = k\,d\,\sin\theta + \delta$, $k = 2\pi/\lambda$ is the wavenumber and δ is the progressive phase difference between the sources. The sum of this geometric series is easily found to be

$$E(\theta) \;=\; E_0\,\frac{\sin(n\psi/2)}{\sin(\psi/2)}\,e^{j(n-1)\psi/2}. \tag{6.2.4}$$

If the centre of the array is chosen as the phase reference point, then the above result does not contain the phase term of $(n-1)\psi/2$. For nonisotropic but similar elements, E_0 is replaced by the element pattern, $E_i(\theta)$, to obtain the total field pattern.

The field pattern in eqn 6.2.4 has a maximum value of nE_0 when $\psi = 0, 2\pi, 4\pi, \ldots$. The maxima at $\psi = 0$ is called the main lobe, while the other maxima are called grating lobes. For $d < \lambda$, only the main lobe maxima maps to the physically allowed range of $0 \le \theta \le 2\pi$. By suitable choice of the value of δ, this maxima can be "steered" to different values of θ, using the relation $k\,d\,\sin\theta = -\delta$. For example, when all the elements of the array are in phase ($\delta = 0$), the maximum occurs at $\theta = 0$. This is referred to as a "broadside" array. For a maximum along the axis of the array ($\theta = 90^o$), $\delta = -k\,d$ is required, giving rise to an "end-fire" array. The broadside array produces a disc or fan shaped beam that covers a full 360^o in the plane normal to the axis of the array. The end-fire array produces a cigar shaped beam which has the same shape in all planes containing the axis of the array. For nonisotropic elements, the element pattern also needs to be steered (electrically or mechanically) to match the direction of its peak response with that of the peak of the array pattern, in order to achieve the maximum peak of the total pattern.

For the case of $d > \lambda$, the grating lobes are uniformly spaced in $\sin\theta$ with an interval between adjacent lobe maxima of λ/d, which translates to $\ge \lambda/d$ on the θ axis (see Figure 6.3).

The uniform, linear array has nulls in the radiation pattern which are given by the condition $\psi = \pm 2\pi l/n, \quad l = 1, 2, 3, \ldots$ which yields

$$\theta \;=\; \sin^{-1}\left[\frac{1}{kd}\left(\pm\frac{2\pi l}{n} - \delta\right)\right] \quad . \tag{6.2.5}$$

For a broadside array ($\delta = 0$), these null angles are given by

$$\theta \;=\; \sin^{-1}\left(\pm\frac{2\pi l}{nkd}\right) \quad . \tag{6.2.6}$$

Further, if the array is long ($nd \gg l\lambda$), we get

$$\theta \;\simeq\; \pm\,\frac{\lambda l}{nd} \;\simeq\; \pm\,\frac{l}{L_\lambda} \quad , \tag{6.2.7}$$

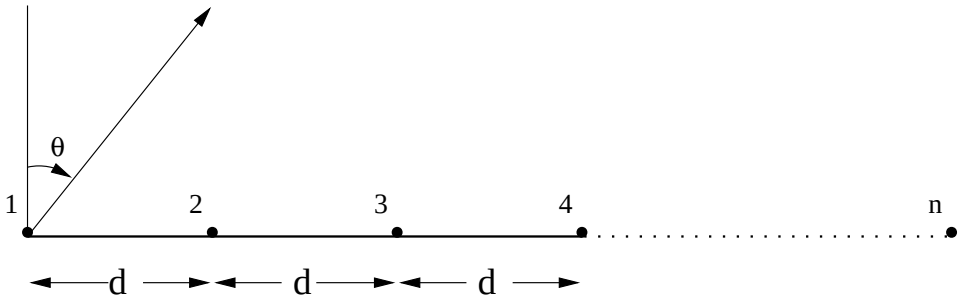


Figure 6.2: Geometry for the n element array

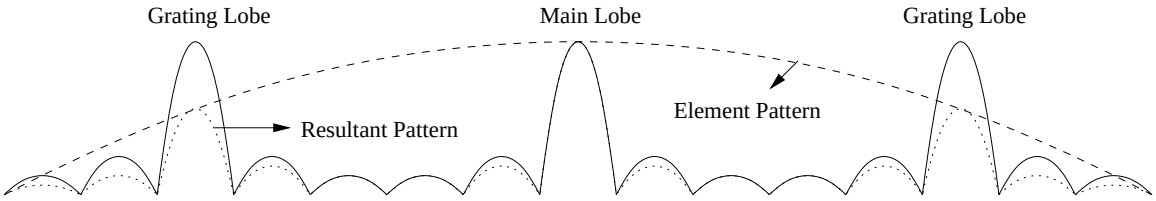


Figure 6.3: Grating lobes for an array of n identical elements. The solid line is the array pattern. The broad, dashed line curve is an example of the element pattern. The resultant of these two is shown as the dotted pattern.

where L_λ is the length of the array in wavelengths and $L_\lambda = (n-1)d/\lambda \simeq nd/\lambda$ for large n. The first nulls occur at $l = \pm 1$, and the beam width between first nulls (BWFN) for such an array is given by

$$BWFN = \frac{2}{L_\lambda}\, rad = \frac{114.6}{L_\lambda}\, deg \ . \tag{6.2.8}$$

The half-power beam width (HPBW) is then given by

$$HPBW \simeq \frac{BWFN}{2} = \frac{57.3}{L_\lambda}\, deg \ . \tag{6.2.9}$$

Similarly, it can be shown that the HPBW of an end-fire array is $\sqrt{2/L_\lambda}$ (see "Antennas" by J.D. Kraus (1988) for more details).

Such linear arrays are useful for studying sources of size $< \lambda/d$ radians, as only one lobe of the pattern can respond to the source at a given time. Also, the source should be strong enough so that confusion due to other sources in the grating lobes is not significant. Linear grating arrays are particularly useful for studying strong isolated sources such as the Sun.

The presence of grating lobes (with amplitude equal to the main lobe) in the response of an array is usually an unwanted feature, and it is desirable to reduce their levels as much as possible. For non-isotropic elements, the taper in the element pattern provides a natural reduction of the amplitude of the higher grating lobes. This is illustrated in Figure 6.3. To get complete cancellation of all the grating lobes starting with the first one, requires an element pattern that has periodic nulls spaced λ/d apart, with the first null falling at the location of the first grating lobe. This requires the elements to have an aperture of $\sim d$, which makes the array equivalent to a continuous or filled aperture telescope. This can be seen mathematically by replacing E_0 in eqn 6.2.4 by the element pattern of an antenna of aperture size d and showing that it reduces to the expression for the field pattern of a continuous aperture of size nd.

The theoretical treatment given above is easily extended to two dimensional antenna arrays.

6.2.3 The Fourier Transform Approach to Array Patterns

So far we have obtained the field pattern of an array by directly adding the electric field contributions from different elements. Now, it is well established that for a given aperture, if the electric field distribution across the aperture is known, then the radiation pattern can be obtained from a Fourier Transform of this distribution (see, for example, Christiansen & Hogbom 1985). This principle can also be used for computing the field pattern of an array. Consider the case of the array pattern for the 2-element array discussed earlier, as an example. The electric field distribution across the aperture can be

taken to be zero at all points except at the location of the two elements, where it is a delta function for isotropic point sources. The Fourier Transform of this gives the sinusoidal oscillations in $\sin\theta$, which have also been inferred from eqn 6.2.2.

Using the Fourier Transform makes it easy to understand the principle of pattern multiplication described above. When the isotropic array elements are replaced with directional elements, it corresponds to convolving their delta function electric field distribution with the electric field distribution across the finite apertures of these directional elements. Since convolution of two functions maps to multiplication of their Fourier Transforms in the transform domain, the total field pattern of the array is naturally the product of the field pattern of the array with isotropic elements with the field pattern of a single element. The computational advantages of the Fourier Transform makes this approach the natural way to obtain the array pattern of two dimensional array telescopes having a complicated distribution of elements.

6.3 Techniques for Phasing an Array

The basic requirement for phasing an array is to combine the signals from the elements with proper delay and phase adjustments so that the beam can be pointed or steered in the chosen direction. Some of the earliest methods employed techniques for mechanically switching in different lengths of cables between each element and the summing point, to introduce the delays required to phase the array for different directions. The job became somewhat less cumbersome with the use of electronic switches, such as PIN diodes. However, the complexity of the cabling and switching network increases enormously with the increase in number of elements and the number of directions for which phasing is required.

Another method of phasing involves the use of phase shifters at each element of the array. For example, this can be achieved by using ferrite devices or by switching in incremental lengths of cable (or microstrip delay lines), using electronic switches. The phase increments are usually implemented in binary steps (for example $\lambda/2, \lambda/4, \lambda/8, \ldots$). In this scheme, the value of the smallest incremental phase difference controls the accuracy of the phasing that can be achieved.

In most modern radio telescopes, digital electronic techniques are used for processing the signals. The output from an antenna is usually down-converted to a baseband frequency in a heterodyne receiver after which it is Nyquist sampled for further processing. Techniques for introducing delays and phase changes in the signal in the digital domain, using computers or special purpose hardware, are fairly easy to implement and flexible.

The description of phasing techniques given above applies when the delay compensation of the signals from the different elements of the array is carried out at the radio frequency of observation. When this delay compensation is carried out at the intermediate or baseband frequency of the heterodyne receiver, the signals pick up an extra phase term of $2\pi\,\nu_{LO}\,\tau_g$, where ν_{LO} is the local oscillator frequency used for the down conversion and τ_g is the delay (with respect to the phase centre of the array) suffered for the element (see for example Thompson, Moran & Swenson, 1986). To obtain the optimum phased array signal, these phase terms have to be compensated before the signals from array elements with different values of τ_g are added. Furthermore, τ_g for an array element varies with time for observations of a given source and this also needs to be compensated.

For an array with similar elements, the amplitude of the signals from the elements is usually kept constant at a common value, while the phase is varied to phase the array. However, in the most general case, the amplitude of the signals from different elements can be adjusted to enhance some features of the array response. This is most often used

to reduce the sidelobe levels of the telescope or shift the nulls of the array pattern to desired locations, such as directions from which unwanted interference signals may be coming. Arrays where such adjustments are easily and dynamically possible are called adaptive beam-forming arrays, and are discussed further in Chapter 7.

6.4 Coherently *vs* Incoherently Phased Array

Normally, the signals from an n-element phased array are combined by adding the voltage signals from the different antennas after proper delay and phase compensation. This summed voltage signal is then put through a square-law detector and an output proportional to the power in the summed signal is obtained. For identical elements, this phased array gives a sensitivity which is n times the sensitivity of a single element, for point source observations. The beam of such a phased array is much narrower than that of the individual elements, as it is the process of adding the voltage signals with different phases from the different elements that produces the narrow beam of the array pattern. For some special applications, it is useful to first put the voltage signal from each element of the array through a square-law detector and then add the powers from the elements to get the final output of the array. This corresponds to an incoherent addition of the signals from the array elements, whereas the first method gives a coherent addition. In the incoherent phased array operation, the beam of the resultant telescope has the same shape as that of a single element, since the phases of the voltages from individual elements are lost in the detection process. This beam width is usually much more than the beam width of the coherent phased array telescope. The sensitivity to a point source is higher for the coherent phased array telescope as compared to the incoherent phased array telescope, by a factor of $\sqrt{n}$.

The incoherent phased array mode of operation is useful for two kinds of astronomical obervations. The first is when the source is extended in size and covers a large fraction of the beam of the element pattern. In this case, the incoherent phased array observation gives a better sensitivity. The second case is when a large region of the sky has to be covered in a survey mode (for example, in a survey of the sky in search for new pulsars). Here, the time taken to cover the same area of sky to equal senstivity level is less for the incoherent phased array mode. Only for a filled aperture phased array telescope are these times the same. For a sparsely filled physical aperture such as an earth rotation aperture synthesis telescope, this distinction between the coherent and incoheret phased array modes is an important aspect of phased array operation.

6.5 Comparison of Phased Array with a Multi-Element Interferometer

As has been mentioned in Section 1, the basic distinction between a phased array and a multi-element interferometer is that in a phased array the signals from all the elements are added in phase before (or after) being put through a square-law detector, where as in a multi-element interferometer, the signals from the elements are correlated in pairs for each possible combination of two elements and these outputs are further processed to make a map of the brightness distribution. Thus, if the signal from element i is given by V_i, the output of the (coherent) phased array can be written as

$$V_{PA} = \left\langle \left(\sum_{i=1}^{n} V_i \right)^2 \right\rangle \tag{6.5.10}$$

whereas the interferometer output is given by

$$V_{ij} \quad = \quad \langle V_i V_j \rangle \quad i,j = 1,2,\ldots,n\,;\, i \neq j \tag{6.5.11}$$

Expansion of the right hand side of eqn 6.5.10 produces terms of the kind $< V_i V_j >$ and V_i^2. The first kind are all available from the correlator outputs and, if the correlator also records the self products of all the elements, the second kind are also provided by the correlator. Thus, by appropriate combinations of the outputs of the correlator used in the multi-element interferometer, the phased array output can be synthesised. Even the steering of the beam of the phased array can be achieved by combining the visibilities from the correlator after multiplying with appropriate phase factors. Also, the incoherently phased array output can be synthesised by combining only the self product outputs from the correlator.

However, the network of multipliers required to implement the correlator is a much more complicated hardware than the adder and square law detector needed for the phased array. Further, the net data rate out of the correlator is much higher than that from the phased array output, for data with the same time resolution. Thus, the interferometer achieves the phased array response in a very expensive manner. This is especially true for very compact, point-like sources where observations with an interferometer do not provide any extra information about the nature of the source. For example, observations of pulsars are best suited to a phased array, as these are virtually point sources for the interferometer and the requirement for high time resolution that is relevant for their studies is more easily met with a phased array output.

6.6 Further Reading

1. Kraus, J.D. “Radio Astronomy”, Cygnus-Quasar Books, Ohio, USA, 1986

2. Kraus, J.D. “Antennas”, McGraw-Hill Book Company, New York, USA, 1988

3. Thompson, A.R., Moran, J.M. & Swenson, G.W. “Interferometry and Synthesis in Radio Astronomy”, John Wiley & Sons, New York, USA, 1986

4. Christiansen, W.N. & Hogbom, J.A. “Radio Telescopes”, Cambridge University Press, Cambridge, UK, 1985

Chapter 7

Imaging With Dipolar Arrays

N. Udaya Shankar

In this lecture we will discuss the radio telescopes in which a beamforming network is used to combine signals from the antenna elements and may also provide the required aperture distribution for beam shaping and side lobe control.

7.1 Early History of Dipole Arrays

Radiotelescopes with a variety of antennas of different forms have been built to suit the large range of wavelengths over which radio observations are made[1]. Quasi-optical antennas such as parabolic reflectors are considered more appropriate for milli-meter and centi-meter wavelengths. At the other end of the radio spectrum, multi element arrays of dipole antennas have been preferred for meter and deca-meter wavelengths.

Early observations in radio astronomy were made using one of the two methods, either pencil beam aerials of somewhat lower resolution to investigate the distribution of radio emission over the sky, or interferometers to observe bright sources of small angular size. However, the observations made during the early 1950's, showed that to determine the real nature of the radio brightness distribution it is necessary to construct pencil beam radio telescopes having beam widths of the same order as the separation between the lobes of the interferometers then in use $(\sim 1')$. An important step towards such modern high-resolution radiotelescopes was the realisation that in many cases even unfilled apertures, which contain all the relative positions of a filled aperture, ("skeleton telescopes") can be used to measure the brightness distribution. A cross-type radio telescope, pioneered by Mills was the first to demonstrate the principle of skeleton telescopes.

A cross consists of two long and relatively narrow arrays arranged as a symmetrical cross, usually in the $N-S$ and the $E-W$ directions, intersecting at right angles at their centers (Figure 7.1). Each array has a fan beam response, narrow along its length and wide in a perpendicular direction[2]. The outputs from both the arrays are amplified and multiplied together; only sources of radiation that lie within the cross hatched portion of Figure 7.1(b) produce a coherent signal. Thus an effective pencil beam is produced of

[1] see the illustrations in Chapter 3

[2] See Section 6.2.2

angular size determined solely by the length of the two arrays. A substantial number of telescopes were constructed based on this principle.

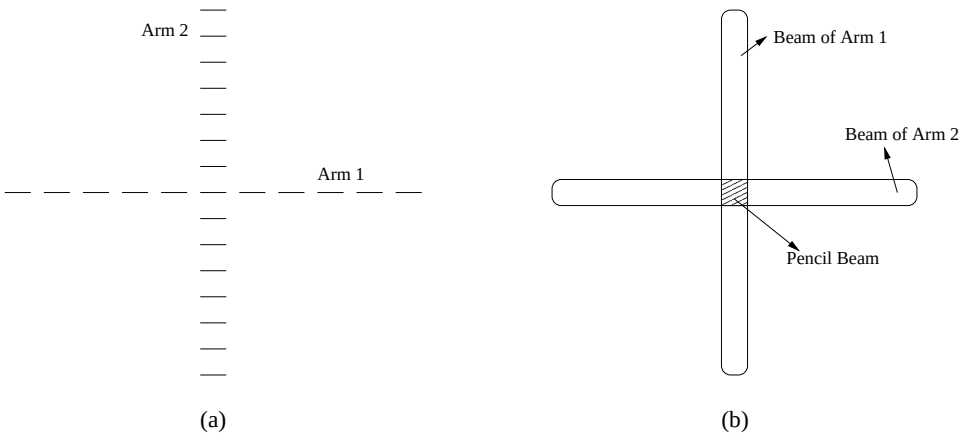


Figure 7.1: A cross type telescope. The arrays in Panel (a) produce the fan beams shown in Panel (b). When the outputs of these two arrays are multiplied together, only signals originating from the cross hatched region common to both beams produce a coherent output. The resolution of such a telescope hence depends only on the lengths of the arms.

The Sydney University telescope was constructed as a cross with aerials of overall dimensions approximately 1 mile long and 40 ft wide (Mills et al 1963). The mile-long reflectors are in the form of cylindrical parabolas, with a surface of wire mesh. Line feeds for two operating frequencies of 408MHz and 111.5MHz were provided at their foci. The $N-S$ arm employs a fixed reflector pointing vertically upwards and the beam is directed in the meridian plane by phasing the dipoles of the feed. The $E-W$ arm is tiltable about its long axis to direct the beam, also in the meridian plane, to intersect the $N-S$ response pattern. No phasing was employed in this aerial. The angular coverage was 55^0 on either side of the zenith. The $E-W$ aperture is divided into two separate halves through which the continuous $N-S$ arm passes. The total collecting area is $400,000$ sq.ft. This instrument had a resolution of approximately $2'.8$ at 408 MHz. This later came to be known as the "Mills Cross" and is one of the earliest cross type radio telescope built. In order to reduce cost, this telescope was built as a meridian transit instrument.

Note that in a cross antenna, one quarter of the antenna provides redundant information, since all element spacings of a filled aperture are still present if half of one array is removed. In fact, it can be shown that the cosine response of a T array is similar to that of a full cross. Thus a survey carried out using a T array has the same resolution as that of a survey carried out using a cross. However it has a collecting area $\sqrt{2}$ times lower than the corresponding cross and hence a lower sensitivity.

7.2 Image Formation

An array can be considered as a sampled aperture. When an array is illuminated by a source, samples of the source's wavefront are recorded at the location of the antenna elements. The outputs from the elements can be subjected to various forms of signal processing, where in phase and amplitude adjustments are made to produce the desired outputs. If the voltages from elemental antennas are simply added (as in the phased

arrays discussed in Chapter 6), the energy received from a large portion of the sky will be rejected. When the array is illuminated by a point source this gives the beam of the array which is the Fourier transform of the aperture current distribution. A single beam instrument can use only a part of the total available time to observe each beam width of the sky. One can generate multiple independent beams in the sky by amplifying the signals from element separately and combining them with different phase shifts. Such a multiple-beam or image forming instrument can observe different directions in the sky simultaneously.

A simple linear array, which generates a single beam, can be converted to a multiple beam antenna by attaching phase shifters to the output of each element. Each beam to be formed requires one additional phase shifter per element. Thus an N element array needs N squared phase shifters. Since the formation of a beam is Fourier transforming the aperture distribution, this requirement of N squared phase shifters is very similar to the requirement of N squared multipliers for an N point Fourier transform. Such a network is known as a Blass network (Figure 7.2). Similar to the fast Fourier transform, we also have a Butler beam-forming matrix, which needs only $N \times logN$ elements for beam forming. The Butler matrix uses 90^0 phase-lag hybrid junctions with 45^0 fixed-phase shifters. Blass and Butler networks for a four-element array are shown in the Figure 7.2. If the elemental spacing is $\lambda/2$, the butler matrix produces four beams. Although these beams overlap, they are mutually orthogonal. Surprisingly the Butler matrix was developed before the development of the FFT.

There are a number of drawbacks with multiple-beam formers, viz.

1. It is difficult to reconfigure the beam former. Most multiple beam formers can only produce fixed beams.

2. The separation between the multiple beams cannot be any less than that for orthogonal beams.

3. As the number of beams is increased, one has to keep track of the signal to noise ratio (SNR) of the individual beams.

4. As the array length becomes longer and the total span of the multiple beams increases, the difference between the arrival times of the wave-front from the source to the ends of the array become comparable to the inverse of the bandwidth of the signal used and the loss of SNR due to bandwidth effects becomes large.

7.3 Digital Beam Forming

Digital Beam Forming (DBF) is a marriage between the antenna technology and digital technology. Workers in Sonar and Radar systems first developed the early ideas of digital beam forming. This coupled with the development of aperture synthesis techniques in radio astronomy lead to the development of the modern dipolar arrays.

An antenna can be considered to be a device that converts spatio temporal signals into strictly temporal signals, there by making them available to a wide variety of signal processing techniques. From a conceptual point of view, its sampled outputs represent all of the data arriving at the antenna aperture. No information is destroyed, at least not until the processing begins and any compromises that are made in the processing stages can be noted and estimates made of the divergence of the actual system from the ideal.

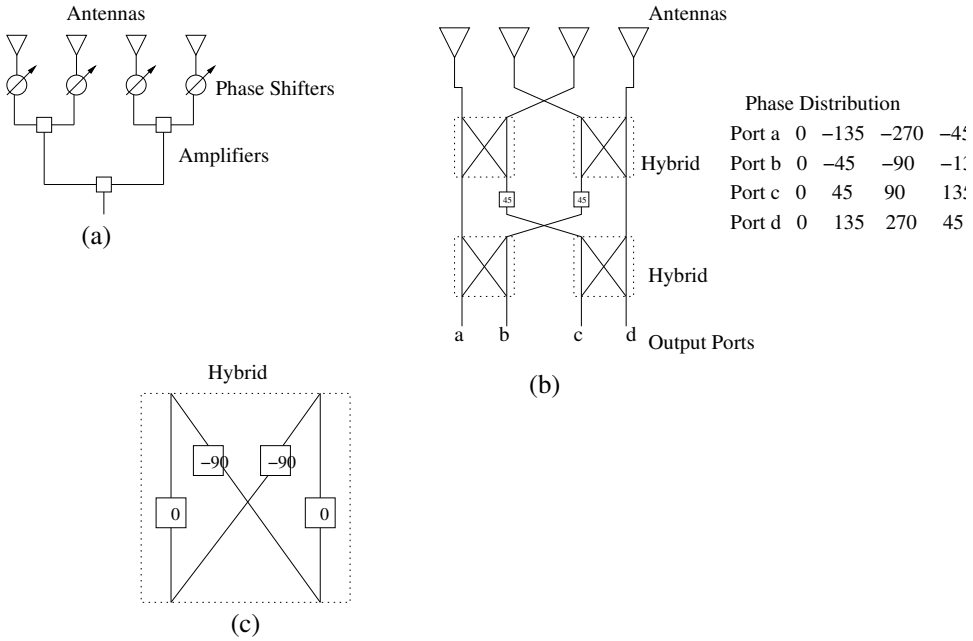


Figure 7.2: A Blass beam forming networks (Panel (a)). Such a network requires N^2 phase shifters to form N beams from N antennas. On the other hand, the Butler beam forming network (Panels (b) and (c)) requires only $N\log(N)$ phase sifters to achieve the same result.

Digital beam forming is based on the conversion of the RF signal at each antenna elements into two streams of binary baseband signals representing cos and sin channels[3]. These two digital baseband signals can be used to recover both the amplitudes and phases of the signals received at each element of the array. The process of digital beamforming implies weighting by a complex weighting function and then adding together to form the desired output. The key to this technology is the accurate translation of the analog signal into the digital regime. Close matching of several receivers is not achieved in hardware, but rather by applying a calibration process. It is expected that more and more of receiver functions will be implemented using software. Eventually one would expect that the receiver would be built using software rather than hardware. We shall get back to this aspect later.

Figure 7.3 depicts a simple structure that can be used for beamforming. The process represented in Figure 7.3(a) is referred to as element-space beamfroming, where the data signals from the array elements are directly multiplied by a set of weights to form the desired beam. Rather than directly weighting the outputs from the array elements, they can be first processed by a multiple-beam beamformer to form a suite of orthogonal beams. The output of each beam can then be weighted and the result combined to produce a desired output. This process is often referred to as the beam-space beamforming (Fig. 7.3(b).

[3]See Section 4.4

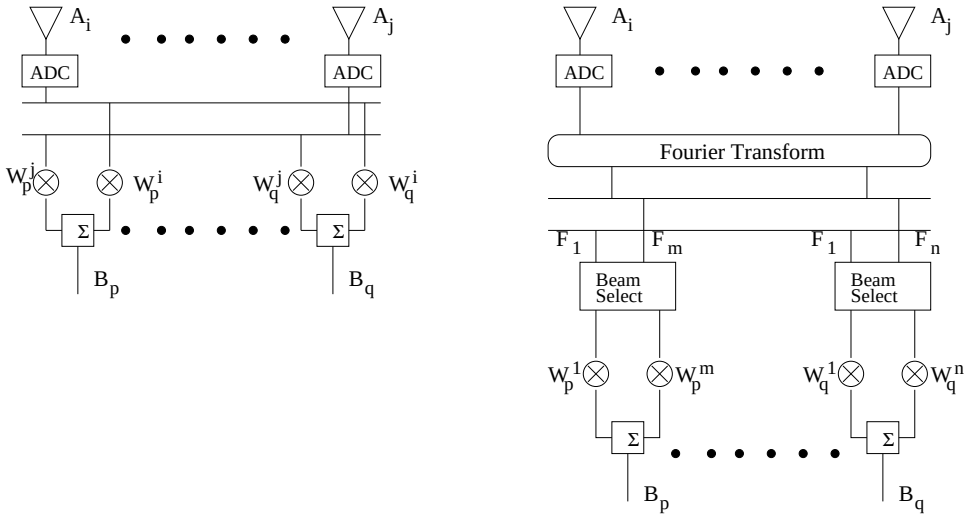


Figure 7.3: Digital beam forming networks. Panel (a) shows an element space beam former while Panel (b) shows a beam space digital beam former.

7.4 Radio Telescopes with Digital Beam Forming Networks

7.4.1 The Clark Lake TEE-PEE-TEE Telescope

This telescope is no more existent. I am using it here as a good example of a telescope which uses a combination of beam forming and synthesis-imaging techniques. This was a fully steerable deca-metric array. This was a T array of 720 conical spiral antennas, 3.0 km by 1.8 km. It had the best sensitivity in the 25 MHz to 75 MHz. Both its operating frequency and beam position were adjustable in less than 1 ms (see Erickson et al. 1982).

The basic element is a long spiral element utilising eight wires wound around a support system that consists of eight parallel filaments. Each element is circularly polarised with a diode switch at its apex that rotates its excitation and thus adjusts its phase. Steering of the array is accomplished by putting a linear phase gradient across groups of 15 elements, called banks. There are 16 banks in the 1800 m $N-S$ arm and 32 banks in the 3000 m $E-W$ arm. The output of each bank is brought separately to the central observatory building.

A separate receiver channel is attached to the output of each of the 48 banks. Each channel employs a superheterodyne receiver[4] to down convert the signal to 10 MHz. The 10 MHz output of each of the receiver channel is sampled at a frequency of 12 MHz digitally delayed and then cross-correlated in a 512 channel two-bit three level complex correlator. An off-line processor removes the fringe rotation[5] introduced by the earth's rotation and integrates the data for periods up to 5 minutes. A Fourier transform then produces a map of the area of the sky under observation. These maps may be averaged to effectively integrate the signal for periods of hours.

It's total collecting area was $250\lambda^2$. The synthesised beam at 30.9 MHz had a width of $13'.0 \times 11'.1$ at the zenith. The confusion limit of the telescope was around 1Jy. It produced

[4]See Section 3.1
[5]See Section 4.4

1024 picture elements in a field of view roughly $6^0 \times 4^0$.

7.4.2 GEETEE: The Gauribidanur T Array

GEETEE is a low frequency radiotelescope operating at 34.5 MHz. It is situated near Gauribidanur, ~ 80 km from Bangalore, India. The antenna system is a T shaped array with 1000 dipoles, 640 in the 1.4 km long $E-W$ array and 360 in the 0.45 km long S array. It's collecting area in the $EW \times S$ correlation mode is 18,000 Sq m and has a resolution of $26' \times 42'\ Sec(\delta - 14^0.1)$. The EW array consists of four rows of dipoles in the NS direction, with 160 dipoles in each row. The S array consists of 90 rows in the NS direction with four dipoles each placed in the EW direction.

A multibeam-forming receiver has been built for $GEETEE$ to obtain long periods of interference free observation over as large a patch of sky as possible in one day. A short observing time for a wide field survey at low frequencies minimises the effects of the ionosphere. For multibeam operation a single row of EW is used in the meridian transit mode. Single row was chosen to maximise the coverage in declination. A single beam in the EW direction was considered sufficient, as the images are confusion limited. 90 outputs of the S array are transmitted to the observatory in 23 open-wire transmission lines using time division multiplexing. In the observatory building, the signals from the EW and S arrays are down-converted to an intermediate frequency of 4 MHz. Then each of the S array output is correlated with the EW array output using one-bit correlators. This gives 90 visibilities sampled at 5 m intervals along the NS direction. The Fourier transform of these visibilities gives 90 multiple beams in the NS direction covering a span of $\pm 47^0$ of Zenith angle along the meridian. A two dimensional image of the sky is obtained by stacking successive scans across the meridian.

7.4.3 MOST: The Molonglo Observatory Synthesis Telescope

A severe disadvantage of the original Mills Cross was that it could make only transit observations. It was recognized that a steerable telescope was necessary to obtain extended observing times and greater sensitivity. To achieve this at a reasonable cost it was decided to abandon the NS arm of the cross and provide a new phased system for the EW arm only. With this a two dimensional aperture is synthesised using earth rotation synthesis. If linear polarisation is used, the position angle of the feeds with respect to the sky will also rotate. Hence, the existing linear feeds were replaced by a circularly polarised feeds.

The usual aperture synthesis procedure accumulates data as points in the spatial frequency (u, v) plane and then interpolates them onto a rectangular grid[6]. The map in the (θ, ϕ) domain is produced by a fast Fourier transform. An important requirement of this method is that the primary beam shape must not vary throughout the observation. This makes it unsuitable for the Molonglo telescope where the primary beam is derived from a rectangular aperture. Because of the mutual coupling problems together with the foreshortening of the effective aperture, the gain of the telescope can vary by over a factor of five as the pointing moves from the meridian to 60^0 from the meridian. This gain variation can be removed from the sampled data, but, the change in beam widths during observations leading to a large variation in the relative gain, between the center of the map and map edges, cannot be corrected for.

The problem of non-circularity and variability of the primary beam may be overcome by the fan beam synthesis or the beam space beam forming. For this the E and the W reflector, each 778 m long and 11.6 m wide (separated by a gap of 15 m) are divided into

[6]See Chapter 11

44 sections of length 17.7 m. The E and W reflectors are tilted about an EW axis by a shaft extending the whole length. To control the direction of response in an east-west direction a phase gradient is set up between the feed elements by differential rotation. Each module output is heterodyned to 11 MHz. A phase controlled transmission line running the length of each antenna distributes the Local Oscillator. One of these lines is phase switched at 400 Hz.

The detection and synthesis process involves the formation of a set of contiguous fan beams in each antenna. The 44 signals are added together in a resistance array to produce 64 real time fan beams. Signals from corresponding beams from each antenna are multiplied to produce 64 real time interferometer beams. By switching the phase gradient by a small amount every second, these 64 beams are time multiplexed to produce either 128, 256, or 384 beams in each 24-second sample. Each beam has an EW width of 43" and at meridian passage a NS width of $2^0.3$. The hardware beams have a separation of 22" and the time multiplexed beams 11", which is just under half the Nyquist sampling requirement.

If observations of a particular field extend over hour angles of ± 6 h, the fan beam rotates through all position angles and synthesis may be performed. The field is represented by a square array of points corresponding to the projection of the celestial sphere onto a plane normal to the earth's rotation axis. Every 24 seconds, the accumulated signal at each of the 4x63 fan beam response angles are added to the nearest (l, m) array points. This process continues throughout the 12 hours of synthesis. The computation apart from summation includes gain, pointing, and phase corrections; cleaning to improve the map; to locate the sources and to measure their flux densities and position.

7.4.4 Summary

These three radio telescopes illustrate different methods of imaging using dipolar arrays as applied to radioastronomy. GEETEE: One-dimensional image synthesis on the meridian with the entire aperture being present at the same time; CLARK LAKE: A two dimensional image synthesis which gave periods of integration much larger than the meridian transit time. The entire aperture was present during an observation schedule; MOST: Rotational synthesis which is used to synthesise a large two dimensional array, using a linear array. All of them use principles of beam forming. GEETEE and CLARK LAKE use the method of measurement of visibilities in the (u, v) domain, while MOST employs the method of direct fan beam synthesis.

We see that the dipolar arrays are used in the meter wavelength ranges more often than at high frequencies. They have very wide fields of view (GEETEE, almost 100^0) and are very good workhorses for surveying the sky. They are good imaging instruments also since they combine the phased array techniques with the principles of synthesis imaging to make images. Unfortunately most of the arrays are equipped with a limited number of correlators and cannot measure all the possible "$n(n-1)/2$" baselines with "n" aperture elements. Thus they are not well suited for applications of self-calibration. Being skeleton telescopes, they have no redundancy in the imaging mode and redundant baseline calibration is not easily applicable. (See Chapter 5 for a discussion on self-calibration and redundant baseline calibration). This has resulted in surveys with limited dynamic range capability. None of these low frequency arrays are equipped with feeds with orthogonal polarisation. So they are not suitable for polarisation studies.

While combining the beam forming techniques with the synthesis techniques, one has to be very careful about the sampling requirement of the spatial frequencies; otherwise one will end up with grating lobes in the synthesised image, even while using linear arrays with contiguous elements spaced $\lambda/2$ apart. Since the dipolar arrays are employed

generally as correlation telescopes and do not have a common collecting area in the arms used for correlation, they suffer from the "zero-spacing problem[7]". Most often today's receivers employ bandpass sampling[8] and if the sampling frequency is not properly chosen one will lose signal to noise. While imaging with arrays it is not un-common, one confronts conflicting requirements between surveying sensitivity and the field of view.

A question may arise in your minds at this stage - with a handful of telescopes using the phased array approach, is there any future for them in radio astronomy? In the remainder of this chapter, I will discuss the possible future of dipolar arrays for radio astronomy.

7.5 Square Kilometer Array (SKA) Concept

In one way or another, all of the various research directions in radioastronomy are limited by our current instrumental sensitivities. Only by ensuring the continued access to order-of-magnitude improvements in our capabilities, can we ensure a continued high rate of discovery! The sensitivity of radio telescopes, in the time between 1940 and 1980, have shown an exponential improvement, over at least 6 orders of magnitude (10^0 mJy to 0.1 mJy for 1 minute integration time). The radio astronomers are toying with the idea of building a telescope with an improvement in sensitivity by a factor of 100 and are hoping that it will lead to fundamental scientific advances (Braun, 1996)

Consideration of the many varied scientific drivers suggests the following basic technical specifications for the instrument:

1. A frequency range of 200 to 2000 MHz.
2. A total collecting area of 1 km^2
3. Distribution over at least 32 elements.

The NFRA in their study of the SKA concept suggest that a broad-band, highly integrated phased array antennas should be adopted for such an array. Some of the advantages are:

1. Phased arrays give "complete" control of beam. The main application considered being the adaptive suppression of RFI environment.
2. Multiple independent beams possible resulting in multiple programs and rapid surveys.

They are planning development work in this direction in several steps: Adaptive array demo, one sq. meter array and a thousand element array and proof of principal arrays. Discussion of all these aspects is beyond the scope of this chapter. Instead we end with the principle of an adaptive array.

7.6 Adaptive Beam Forming

An adaptive beam former is a device that is able to separate signals co-located in the frequency band but separated in the spatial domain. This provides a means for separating the desired signal from interfering signals. An adaptive beam former is able to

[7]The zero spacing problem refers to the difficulty in imaging very large sources, (whose visibilities peak near the origin of the u-v plane) with arrays which provide few to no samples near the u-v plane origin. See Section 11.6 for a more detailed discussion.

[8]See Chapter 1

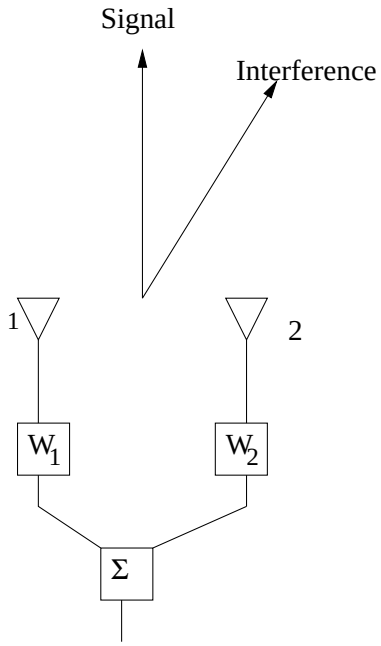


Figure 7.4: A two element adaptive array for interference suppression. The array simultaneously accepts a signal coming from the zenith, while rejecting an interfering signal 30^o from the zenith by a suitable choice of the weights W_i.

automatically optimise the array pattern by adjusting the elemental control weights until a prescribed objective function is satisfied. An algorithm designed for that purpose specifies the means by which the optimisation is achieved. These devices use far more of the information available at the antenna aperture than does a conventional beamformer.

The procedure used for steering and modifying an array's beam pattern in order to enhance the reception of a desired signal, while simultaneously suppressing interfering signals through complex weight selection is illustrated by the following example. Let us consider the array shown in Figure 7.4. The array consists of two antennas with a spacing of $\lambda/2$. Let the signal $S(t)$ arriving from a radio source at zenith is the desired signal. Let $I(t)$ be an interfering signal arriving from a direction $\theta = \pi/6$ radians. The signal from each element is multiplied by a variable complex weight (w_1, w_2) and the weighted signals are then summed to form the array output. The array output due to the desired signal is

$$Y(t) = A\ e^{j2\pi ft}[w_1 + w_2]. \tag{7.6.1}$$

For the $Y(t)$ to be equal to $S(t)$, it is necessary that

$$RP[w_1] + RP[w_2] = 1 \tag{7.6.2}$$

and

$$IP[w_1] + IP[w_2] = 0. \tag{7.6.3}$$

Where RP and IP denote real and imaginary parts of the complex weights. The interfering signal arrives at the element 2 with a phase lead of $\pi/2$ with respect to the element 1. Consequently the array output due to the interfering signal is given by

$$Y_i(t) = [Ne^{j2\pi ft}]w_1 + [Ne^{j2\pi ft+\pi/2}]w_2. \tag{7.6.4}$$

For the array response to the interference to be zero, it is necessary that

$$RP[w_1] + RP[jw_2] = 0 \tag{7.6.5}$$

and

$$IP[w_2] + IP[jw_2] = 0. \tag{7.6.6}$$

The requirement that the array has to respond to only the radio source and not to the interfering signal leads to the solution

$$w_1 = 1/2 - j1/2 \tag{7.6.7}$$

and

$$w_2 = 1/2 + j1/2. \tag{7.6.8}$$

With these weights, the array will accept the desired signal while simultaneously rejecting the interference.

The method used in the above example exploits the fact that there is only one directional interference source and uses the a priori information concerning the frequency and the directions of both of the signals. A more practical processor should not require such a detailed a priori information about the location, number and nature of sources. But this example has demonstrated that a system consisting of an array, which is configured with complex weights, provides numerous possibilities for realising array system objectives. We need to only develop a practical processor for carrying out the complex weight adjustment. In such a processor the choice of the weighting will be based on the statistics of the signal of interest received at the array. Basically the objective is to optimise the beamformer response with respect to a prescribed criterion, so that the output contains minimal contribution from the interfering signal.

There can be no doubt about the worsening observing situation in radio astronomy due to the increased use of frequency space for communications. But a pragmatic view is that it is hopeless to resist the increased use of frequency space by others and we must learn to live with it. The saving grace is that the requirements of mobile cellular, satellite and personal communication services systems are pushing the advancement in technology to provide increasingly faster and less expensive digital hardware. The present trend is to replace the analog functions of a radio receiver with software or digital hardware. The ultimate goal is to directly digitise the RF signal at the output of the receiving antenna and then implement the rest of the radio functions in either digital hardware or software. Trends have evolved toward this goal by incorporating digitisation closer and closer to the antenna at increasingly higher frequencies and wider bandwidths. It is appropriate that the radio astronomer uses this emerging technology to make the future radio telescopes interference free. Adaptive arrays hold the key to this endeavour.

7.7 Further Reading

1. Braude S. Ya., Megn A.V., Ryabon B.P., Sharykin N.K., Zhonck I.N. 1978, *Decametric survey of discrete sources in the Northern sky.*, Astrophysics and Space Science, 54 3-36

2. Braun R. 1996, *In the Westerbork Observatory: Continuing Adventure in Radio Astronomy eds.* Raimond E. and Genee R.O. Kluwer Dordrecrt.

3. Erickson W.C. Mahoney M.J., Erb K. 1982, *The Clark Lake Teepee - tee telescope*; Astrphys. J Suppl., 50 403-419

4. Mills B.Y., Aitchison R.E. Little A.G., February 1963, *The Sydney University cross-type Radio Telescope*; Proceedings of the I.R.E. Australia, 156-165.

5. Udaya Shankar N., Ravi Shankar T.S. 1990, A *Digital Correlaiton Receiver for the GEETEE Radio Telescope*; Journal of Astrophysics and Astronomy, 11, 297-310.

Chapter 8

Correlator I. Basics

D. Anish Roshi

8.1 Introduction

A radio interferometer measures the mutual coherence function of the electric field due to a given source brightness distribution in the sky. The antennas of the interferometer convert the electric field into voltages. The mutual coherence function is measured by cross correlating the voltages from each pair of antennas. The measured cross correlation function is also called *Visibility*. In general it is required to measure the visibility for different frequencies (spectral visibility) to get spectral information for the astronomical source. The electronic device used to measure the spectral visibility is called a *spectral correlator*. These devices are implemented using digital techniques. Digital techniques are far superior to analog techniques as far as stability and repeatability is concerned.

The first of these two chapters on correlators covers some aspects of digital signal processing used in digital correlators. Details of the hardware implementation of the GMRT spectral correlator are presented in the next lecture.

8.2 Digitization

The signals[1] at the output of the antenna/receiver system are analog voltages. Measurements using digital techniques require these voltages to be sampled and quantized.

8.2.1 Sampling

A band limited signal $s(t)$ with bandwidth $\Delta\nu$ can be uniquely represented by a time series obtained by periodically sampling $s(t)$ at a frequency f_s (the sampling frequency) which is greater than a critical frequency $2\Delta\nu$ (Shannon 1949). The signal is said to be 'Nyquist sampled' if the sampling frequency is exactly equal to the critical frequency $2\Delta\nu$.

The spectrum of signals sampled at a frequency $< 2\ \Delta\nu$ (i.e. under sampled) is distorted. Therefore the time series thus obtained is not a true representation of the band limited signal. The spectral distortion caused by under sampling is called *aliasing*.

[1]For all the analysis presented here we assume that radio astronomy signals are stationary and ergodic stochastic processes with a gaussian probability distribution. We also assume that the signals have zero mean.

8.2.2 Quantization

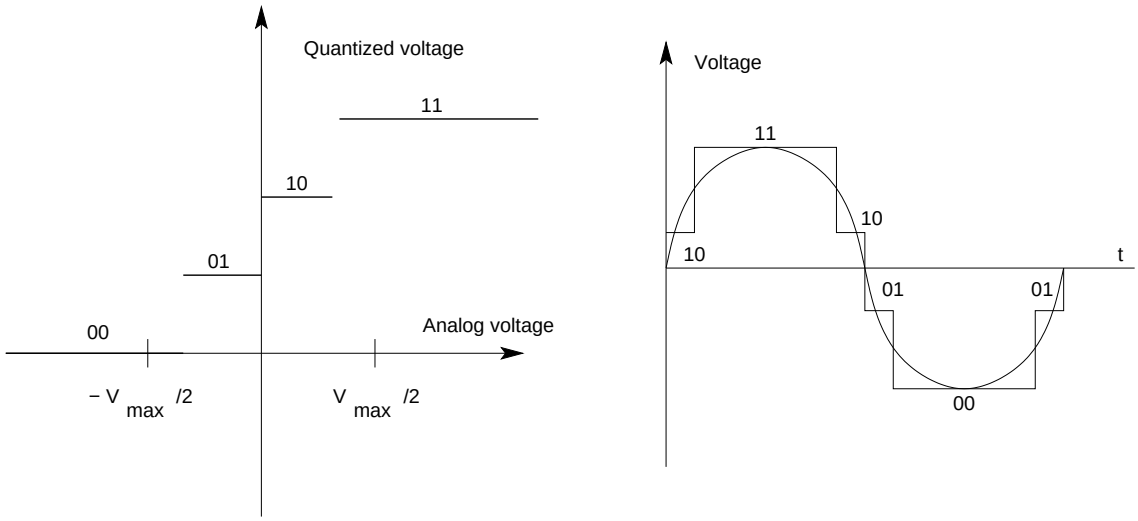


Figure 8.1: Transfer function of a two bit four level quantizer. The *binary* numbers corresponding to the quantized voltage range from 00 to 11. Quantization of a sine wave with such a quantizer is also shown.

The amplitude of the sampled signal is a continuous value. Digital systems represent values using a finite number of bits. Hence the amplitude has to be approximated and expressed with these finite number of bits. This processes is called *quantization*. The quantized values are integer multiple of a quantity q called the *quantization step*. An example of two bit (or equivalently four level) quantization is shown in Fig. 8.1. For the quantizer $q = V_{max}/2^2$, where V_{max} is the maximum voltage (peak-to-peak) that can be expressed within an error of $\pm q/2$.

Quantization distorts the sampled signal affecting both the amplitude and spectrum of the signal. This is evident from Fig. 8.1 for the case of a two bit four level quantized sine wave. The amplitude distortion can be expressed in terms of an error function $e(t) = s(t) - s_q(t)$, which is also called the *quantization noise*. Here $s_q(t)$ is the output of the quantizer. The variance of quantization noise under certain restricted conditions (such as uniform quantization) is $q^2/12$. The spectrum of quantization noise extends beyond the bandwidth $\Delta\nu$ of $s(t)$ (see Fig. 8.2). Sampling at the Nyquist rate ($2\Delta\nu$) therefore aliases the power of the quantization noise outside $\Delta\nu$ back into the spectral band of $s(t)$. For radio astronomy signals, the spectral density of the quantization noise within $\Delta\nu$ can be considered uniform and is $\sim q^2/12\Delta\nu$ (assuming uniform quantization). Reduction in quantization noise is hence possible by oversampling $s(t)$ (i.e. $f_s > 2\Delta\nu$) since it reduces the aliased power. For example, the signal to noise ratio of a digital measurement of the correlation function of s(t) (see Section 8.5) using a Nyquist sampling and a two bit four level quantizer is 88% of the signal to noise ratio obtained by doing analog correlation for Nyquist sampling and 94% if one were to sample at twice the Nyquist rate.

The largest value that can be expressed by a quantizer is determined by the number of bits (M) used for quantization. This value is $2^M - 1$ for binary representation. The finite number of bits puts an upper bound on the amplitude of input voltage that can be expressed within an error $\pm q/2$. Signals with amplitude above the maximum value will be 'clipped', thus producing further distortion. This distortion is minimum if the

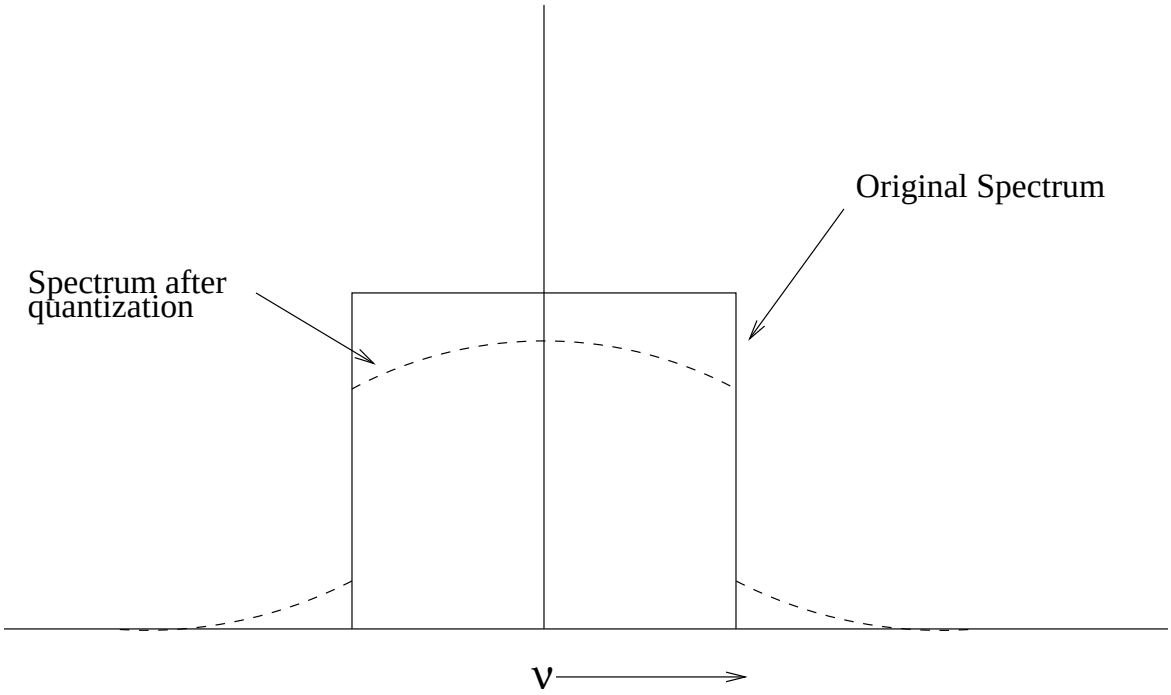


Figure 8.2: Power spectrum of band limited gaussian noise after one bit quantization. The spectrum of the original analog signal is shown with a solid line, while that of the quantized signal is shown with a dotted line.

probability of amplitude of the signal exceeding $+V_{max}/2$ and $-V_{max}/2$ is less than 10^{-5}. For a signal with a gaussian amplitude distribution this means that $V_{max} = 4.42\sigma$, σ being the standard deviation of $s(t)$.

8.2.3 Dynamic Range

As described above, the quantizer degrades the signal if its (peak-to-peak) amplitude is above an upper bound V_{max}. The minimum change in signal amplitude that can be expressed is limited by the quantization step q. Thus a given quantizer operates over a limited range of input voltage amplitude called its *dynamic range*. The Dynamic range of a quantizer is usually defined by the ratio of the power of a sinusoidal signal with peak-to-peak amplitude = V_{max} to the variance of the quantization noise. For an ideal quantizer with uniform quantization the dynamic range is $\frac{3}{2}2^{2M}$. Thus the dynamic range is larger if the number of bits used for quantization is larger.

8.3 Discrete Fourier Transform

The Fourier Transform (FT) of a signal $s(t)$ is defined as

$$S(w) = \int_{-\infty}^{+\infty} s(t)e^{-j\omega t}\mathrm{d}t \tag{8.3.1}$$

Discrete Fourier Transform (DFT) is an operation to evaluate the FT of the sampled signal $s(n)$ ($\equiv s(n\frac{1}{f_s})$) with a finite number of samples (say N). It is defined as

$$S(k) = \sum_{n=0}^{N-1} s(n)e^{-j2\pi nk/N};\ \ 0 \leq k \leq N-1 \tag{8.3.2}$$

The relationship between FT and DFT and some properties of DFT are discussed here.

Consider a time series $s(n)$, which is obtained by sampling a continuous band limited signal $s(t)$ at a rate f_s (see Fig. 8.3). The sampling function is a train of delta function $\mathrm{III}(t)$. The length of the series is restricted to N samples by multiplying with a rectangular window function $\Pi(t)$. The modification of the signal $s(t)$ due to these operations and the corresponding changes in the spectrum are shown in Fig. 8.3. The spectral modifications can be understood from the properties of Fourier transforms. The FT of the time series can now be written as a summation (assuming N is even)

$$\begin{aligned} S(\omega) &= \int_{-\infty}^{+\infty} s(t) \sum_{n=-N/2}^{N/2-1} \delta(t - \frac{n}{f_s})e^{-j\omega t}\mathrm{d}t \\ &= \sum_{n=-N/2}^{N/2-1} s(\frac{n}{f_s})e^{-\frac{j\omega n}{f_s}} \end{aligned} \tag{8.3.3}$$

What remains is to quantize the frequency variable. For this the frequency domain is sampled such that there is no *aliasing in the time domain* (see Fig. 8.3). This is satisfied if $\Delta\omega = 2\pi f_s/N$. Thus Eq. 8.3.3 can be written as

$$S(k\Delta\omega) = \sum_{n=-N/2}^{N/2-1} s(\frac{n}{f_s})e^{-\frac{jk\Delta\omega n}{f_s}} \tag{8.3.4}$$

Using the relation $\Delta\omega/f_s = 2\pi/N$ and writing the variables as discrete indices we get the DFT equation. The cyclic nature of DFT (see below) allows n and k to range from 0 to $N-1$ instead of $-N/2$ to $N/2-1$.

Some properties that require attention are:

1. The spectral values computed for $N/2 \geq k \geq 3N/2-1$ are identical to those for $k = -N/2$ to $N/2-1$. In fact the computed values have a periodicity equal to $N\Delta\omega$ which makes the DFT cyclic in nature. This periodicity is a consequence of the sampling done in the time and frequency domain (see Fig. 8.3).

2. The sampling interval of the frequency variable $\Delta\omega$ (= $2\pi f_s/N$) is inversely proportional to the total number of samples used in the DFT. This is discussed further in Section 8.3.1.

There are several algorithms developed to reduce the number of operations in the DFT computation, which are called Fast Fourier Transform (FFT) algorithms. These algorithms reduce the time required for the computation of the DFT from $O(N^2)$ to $O(N\log(N))$. The FFT implementation used in the GMRT correlator uses Radix 4 and Radix 2 algorithms.

In the digital implementation of FFTs the quantization of the coefficients $e^{-j2\pi nk/N}$ degrades the signal to noise ratio the of spectrum. This degradation is in addition to the quantization noise introduced by the quantizer. Thus the dynamic range reduces further due to coefficient quantization. Coefficient quantization can also produce systematics in the computed spectrum. This effect also depends on the statistics of the input signal, and in general can be reduced only by using a larger number of bits for coefficient representation.

8.3.1 Filtering and Windowing

The Fourier transform of a signal $s(t)$ is a decomposition into frequency or spectral components. The DFT also performs a spectral decomposition but with a finite *spectral resolution*. The spectrum of a signal $s(t)$ obtained using a DFT operation is the convolution of the true spectrum of the signal $S(f)$ convolved by the FT $W(f)$ of the window function, and sampled at discrete frequencies. Thus a DFT is equivalent to a filter bank with filters spaced at $\Delta\omega$ in frequency. The response of each filter is the Fourier transform of the *window function* used to restrict the number of samples to N. For example, in the above analysis (see Section 8.3) the response of each 'filter' is the *sinc* function, (which is the FT of the rectangular window $\Pi(t)$). The spectral resolution (defined as the full width at half maximum (FWHM) of the filter response) of the sinc function is $\frac{1.21\Delta\omega}{2\pi}$. Different window functions $w(n)$ give different 'filter' responses, i.e. for

$$S(k) = \sum_{n=0}^{N-1} w(n)s(n)e^{-j2\pi nk/N} \tag{8.3.5}$$

the Hanning window

$$\begin{aligned} w(n) &= 0.5(1+\cos(2\pi n/N)) \text{ for } -N/2 \leq n \leq N/2-1 \\ &= 0 \text{ elsewhere} \end{aligned} \tag{8.3.6}$$

has a spectral resolution $\frac{2\Delta\omega}{2\pi}$. Side lobe reduction and resolution are the two principal considerations in choosing a given window function (or equivalently a given filter response). The rectangular window (i.e. sinc response function) has high resolution but a peak sidelobe of 22% while the Hanning window has poorer resolution but peak sidelobe level of only 2.6%.

8.4 Digital Delay

In interferometry the geometric delay suffered by a signal (see Chapter 4) has to be compensated before correlation is done. In an analog system this can be achieved by adding or removing cables from the signal path. An equivalent method in digital processing is to take sampled data that are offset in time. Mathematically, $s(n-m)$ is the sample delayed by $m \times 1/f_s$ with respect to $s(n)$ (where f_s is the sampling frequency). In such an implementation of delay it is obvious that the delay can be corrected only to the nearest integral multiple of $1/f_s$.

A delay less than $1/f_s$ (called *fractional delay*) can also be achieved digitally. A delay τ introduced in the path of a narrow band signal with angular frequency ω produces a phase $\phi = \omega\tau$. Thus, for a broad band signal, the delay introduces a phase gradient across the spectrum. The slope of the phase gradient is equal to the delay or $\tau = \frac{\mathrm{d}\phi}{\mathrm{d}\omega}$. This means that introducing a phase gradient in the FT of $s(t)$ is equivalent to introducing a delay is $s(t)$. Small enough phase gradients can be applied to realize a delay $< 1/f_s$. In the GMRT correlator, residual delays $\tau < 1/f_s$ is compensated using this method. This correction is called the Fractional Sampling Time Correction or FSTC.

8.5 Discrete Correlation and the Power Spectral Density

The cross correlation of two signals $s_1(t)$ and $s_2(t)$ is given by

$$R_c(\tau) = \ < s_1(t)s_2(t+\tau) > \tag{8.5.7}$$

where τ is the time delay between the the two signals. In the above equation the angle bracket indicates averaging in time. For measuring $R_c(\tau)$ in practice an estimator is defined as

$$R(m) = \frac{1}{N}\sum_{n=0}^{N-1} s_1(n)s_2(n+m) \;\; 0 \leq m \leq M \tag{8.5.8}$$

where m denotes the number of samples by which $s_2(n)$ is delayed, M is the maximum delay ($M \ll N$). By definition $R(m)$ is a random variable. The expectation value of $R(m)$ converges to $R_c(\tau = \frac{m}{f_s})$ when $N \to \infty$. The autocorrelation of the time series $s_1(n)$ is also obtained using a similar equation as Eq. 8.5.8 by replacing $s_2(n+m)$ by $s_1(n+m)$.

The correlation function estimated from the quantized samples in general deviates from the measurements taken with infinite amplitude precision. The deviation depends on the true correlation value of the signals. The relationship between the two measurement can be expressed as

$$\hat{R}_c(m/f_s) = \mathrm{F}(\hat{R}(m)) \tag{8.5.9}$$

where $\hat{R}_c(m/f_s)$ and $\hat{R}(m)$ are the normalized correlation functions (normalized with zero lag correlation in the case of autocorrelation and with square root of zero lag autocorrelations of the signal $s_1(t)$ and $s_2(t)$ in the case of cross correlation) and F is a correction function. It can be shown that the correction function is monotonic (Van Vleck & Middelton 1966, Cooper 1970, Hagan & Farley 1973, Kogan 1998). For example, the functional dependence for a one-bit quantization (the 'Van Vleck Correction') is

$$\hat{R}_c(m/f_s) = \sin(\frac{\pi}{2}\hat{R}(m)) \tag{8.5.10}$$

Note that the correction function is non-linear and hence this correction should be applied before any further operation on the correlation function. If the number of bits used for quantization is large then over a large range of correlation values the correction function is approximately linear.

The power spectral density (PSD) of a stationary stochastic process is defined to be the FT of its auto-correlation function (the Wiener-Khinchin theorem). That is if $R_c(\tau) = \; < s(t)s(t-\tau) >$ then the PSD, $S_c(f)$ is

$$S_c(f) = \int_{-\infty}^{\infty} R_c(\tau)e^{-j2\pi f\tau}d\tau \tag{8.5.11}$$

From the properties of Fourier transforms we have

$$R_c(0) = \; < s(t)s(t) > \; = \; \int_{-\infty}^{\infty} S_c(f)df \tag{8.5.12}$$

i.e. the function $S_c(f)$ is a decomposition of the variance (i.e. 'power') of $s(t)$ into different frequency components.

For sampled signals, the PSD is estimated by the Fourier transform of the discrete auto-correlation function. In case the signal is also quantized before the correlation, then one has to apply a Van Vleck correction *prior* to taking the DFT. Exactly as before, this estimate of the PSD is related to the true PSD via convolution with the window function.

One could also imagine trying to determine the PSD of a function $s(t)$ in the following way. Take the DFTs of the sampled signal $s(n)$ for several periods of length N and average them together and use this as an estimate of the PSD. It can be shown that this process is exactly equivalent to taking the DFT of the discrete auto-correlation function.

The cross power spectrum of the two signals is defined as the FT of the cross correlation function and the estimator is defined in a similar manner to that of the auto-correlation case.

8.6 Further Reading

1. Cooper, B.F.C. 1970, Aust. J. Phys. 23, 521
2. Hagen, J.B., Farley, D.T. 1973, Radio Science, 8, 775
3. Kogan, L. 1998, Radio Science, 33, N5, p 1289
4. Thompson, R.A., Moran, J.M., Swenson, Jr. G.W., "Interferometry and Synthesis in Radio Astronomy", Chapter 8, John Wiley & Sons, 1986.
5. Oppenheim, A.V. & Schafer, R.W., "Digital Signal Processing", Prentice-Hall, Englewood Cliffs, New Jersey, 1975.
6. Rabiner L.R. & Gold B, "Theory and Application of DSP"
7. Shannon, C.E. 1949, Proc. IRE, 37, 10
8. Thompson, A.R. & D'Addario, L.R. in "Synthesis Imaging in Radio Astronomy", R.A. Perley, F.R. Schwab, & A.H. Bridle, eds., ASP Conf. Series, vol. 6.
9. Van Vleck, J. H., Middelton, D. 1966, Proc IEEE, 54, 2

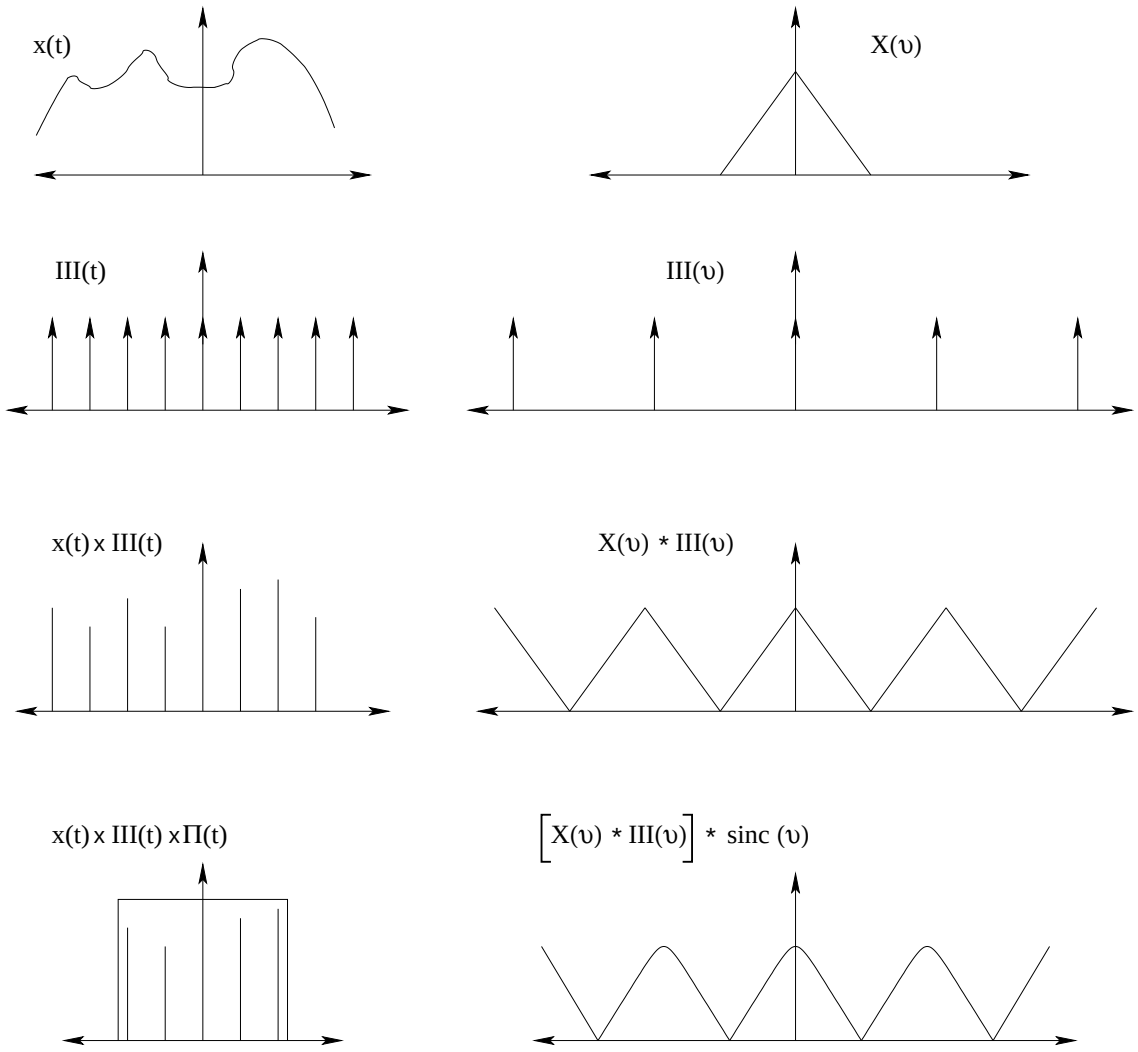


Figure 8.3: The relation between the continuous Fourier transform and the discrete Fourier transform. The panels on the left show the time domain signal and those on the right show the corresponding spectra.

Chapter 9

Correlator – II: Implementation

D. Anish Roshi

The visibility measured by an interferometer is characterized by the amplitude and phase of the fringe at different instants. For simplicity first consider the output of a two element interferometer. In the quasi monochromatic approximation the multiplier output can be written as (see Chapter 4)

$$r_R(\tau_g) = \mathrm{Re}[v_1(\nu,t)v_2^*(\nu,t)] = |\mathcal{V}|\cos(2\pi\nu\tau_g + \Phi_\mathcal{V}), \tag{9.0.1}$$

where $v_1(\nu,t)$ and $v_2^*(\nu,t)$ are the voltages at the outputs of the receiver systems of the two antennas, $|\mathcal{V}|$ and $\Phi_\mathcal{V}$ are the amplitude and the phase of the visibility and τ_g is the geometric delay. The quantities required for mapping a source are $|\mathcal{V}|$ and $\Phi_\mathcal{V}$ for all pairs of antennas of the interferometer. These quantities are measured by first canceling the $2\pi\nu\tau_g$ term in Eq. 9.0.1 by *delay tracking and fringe stopping*. In general, one needs to know the amplitude and phase of the visibility as a function of frequency. This chapter covers the implementation of a spectral correlator to measure the visibility amplitude and phase. Further since the delay tracking (and fringe stopping for some cases) is usually also done by the correlator, these issues are also discussed.

9.1 Delay Tracking and Fringe Stopping

Signals received by antennas are down converted to baseband by mixing with a local oscillator of frequency ν_{LO}. The geometric delay compensation is usually done by introducing delays in the baseband signal. The output of a correlator after introducing a delay τ_i can be written as (see Chapter 4)

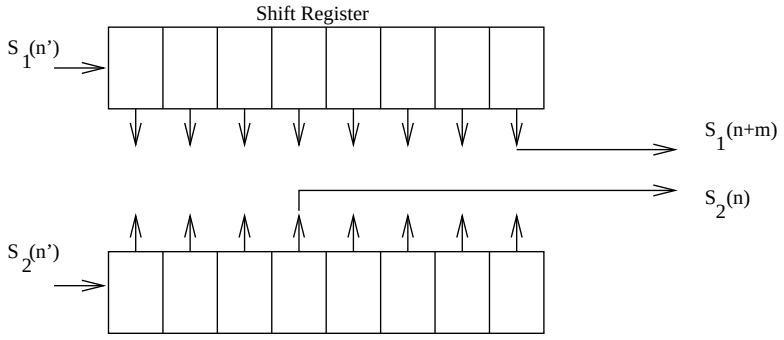
$$\begin{aligned} r_R(\tau_g) &= |\mathcal{V}|\cos(2\pi\nu\tau_g - 2\pi\nu_{BB}\tau_i + \Phi_\mathcal{V}) && (9.1.2)\\ &= |\mathcal{V}|\cos(2\pi\nu_{LO}\tau_g - 2\pi\nu_{BB}\Delta\tau_i + \Phi_\mathcal{V})\,, && (9.1.3)\end{aligned}$$

where ν_{BB} is the baseband frequency and $\Delta\tau_i = \tau_g - \tau_i$ is the residual delay. There are two terms that arise in the equation due to delay compensation:

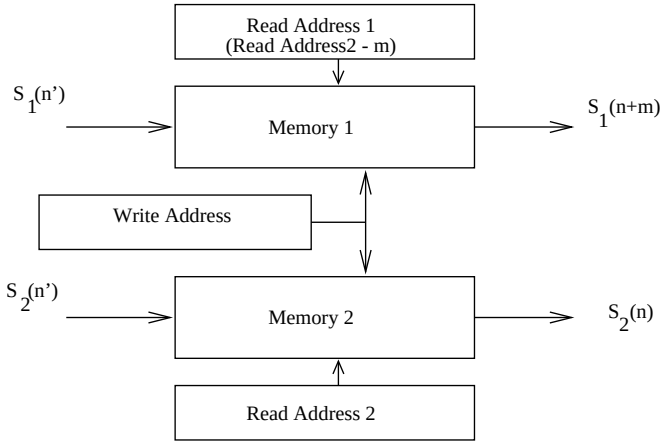
1. $2\pi\nu_{BB}\Delta\tau_i$, and

2. $2\pi\nu_{LO}\tau_g$.

The first term is due to finite precision of delay compensation and the later is a consequence of the delay being compensated in the baseband (as opposed to the RF, which is where the geometric delay is suffered, see Chapter 4). The phase $2\pi\nu_{BB}\Delta\tau_i$ depends on ν_{BB}. For observations with a bandwidth $\Delta\nu$ this term produces a phase gradient across $\Delta\nu$. The phase gradient is a function of time since the delay error changes with time. The phase $2\pi\nu_{LO}\tau_g$ is independent of ν_{BB}, thus is a constant across the entire band. This phase is also a function of time due to time dependence of τ_g. Thus both these quantities have to be dynamically compensated.



Delay implementation using shift registers



Delay implementation using Memory

Figure 9.1: Digital implementation of delay tracking in units of the sampling period using shift registers (top) and random access memory (bottom).

Delay compensation in multiples of sampling interval $1/f_s$ can be achieved by shifting the sampled data (see Chapter 8). This is schematically shown in Fig. 9.1. The digitized samples are passed through shift registers. The length of the shift registers are adjusted to introduce the required delay between the signals. Another way of implementing delay is by using random access memory (RAM). In this scheme, the data from the antennas are written into a RAM (Fig. 9.1). The data is then read out from this memory for further

proccessing. However, the read pointer and the write pointer are offset, and the offset between the two can be adjusted to introduce exactly the required delay. In the GMRT correlator, the delay compensation is done using such a high speed dual port RAM.

A fractional delay can be introduced by changing the phase of the sampling clock. The phase is changed such that signals from two antennas are sampled with a time difference equal to the fractional delay. A second method is to introduce phase gradients in the spectrum of the signal (see Chapter 8). This phase gradient can be introduced after taking Fourier Transforms of signals from the antennas (see Section 9.2.1).

Compensation of $2\pi\nu_{LO}\tau_g$, (called *fringe stopping*, can be done by changing the phase of the local oscillator signal by an amount ϕ_{LO} so that $2\pi\nu_{LO}\tau_g - \phi_{LO} = 0$. Alternatively, this compensation can be achieved digitally by multiplying the sampled time series by $e^{-j\phi_{LO}}$. (Recall from above that the fringe rate is the same for all frequency channels, so this correction can be done in the time domain). The fringe

$$\phi_{LO}(t) = 2\pi\nu_{LO}\tau_g = 2\pi\nu_{LO}\frac{b\sin(\Omega t)}{c} \tag{9.1.4}$$

is a non-linear function of time (see Chapter 4). Here Ω is the rate at which the source is moving in the sky (i.e. the angular rotation speed of the earth), b is the baseline length and c is the velocity of light. For a short time interval Δt about t_0 the time dependence can be approximated as

$$\phi_{LO}(t) = \phi_{LO}(t_0) + 2\pi\nu_{LO}\frac{b\Omega\cos(\Omega t_0)}{c}\Delta t. \tag{9.1.5}$$

i.e. $\phi_{LO}(t)$ is the phase of an oscillator with frequency

$$\nu_{LO}\frac{b\Omega\cos(\Omega t_0)}{c} \tag{9.1.6}$$

After a time interval Δt the frequency of the oscillator has to be updated. Digital implementation of an oscillator of this type is called a *Number controlled oscillator* (NCO). The frequency of the oscillator is varied by loading a control $number$ to the device. The initial phase of the NCO can also be controlled which is used to introduce $\phi_{LO}(t_0)$. In the GMRT correlator, fringe stopping is done using an NCO.

9.2 Spectral Correlator

The output of a simple multiplier of the two element interferometer after delay compensation can be written as:

$$r_R = |\mathcal{V}|\cos(\Phi_{\mathcal{V}}). \tag{9.2.7}$$

To separate $|\mathcal{V}|$ and $\Phi_{\mathcal{V}}$ a second product is measured after introducing a phase shift of $90\,\mathrm{deg}$ in the signal path (see Fig 9.2). Introducing a $90\,\mathrm{deg}$ shift in the path of one of the signals will result in (se Eq. 9.0.1)

$$r_I(\tau_g) = |\mathcal{V}|\cos(2\pi\nu\tau_g + \Phi_{\mathcal{V}} + \pi/2), \tag{9.2.8}$$

and after compensating for $2\pi\nu\tau_g$

$$\begin{aligned} r_I &= |\mathcal{V}|\cos(\Phi_{\mathcal{V}} + \pi/2) \\ &= |\mathcal{V}|\sin(\Phi_{\mathcal{V}}). \end{aligned} \tag{9.2.9}$$

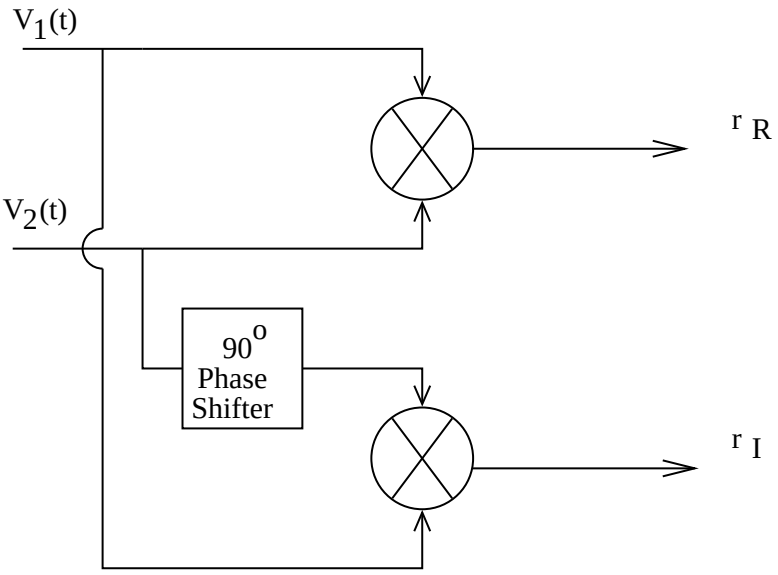


Figure 9.2: Block diagram of a complex multiplier.

From these two measurement we get

$$|\mathcal{V}| = \sqrt{r_R^2 + r_I^2} \tag{9.2.10}$$

$$\Phi_{\mathcal{V}} = \tan^{-1}(\frac{r_I}{r_R}). \tag{9.2.11}$$

Alternatively, for mathematical convenience, the two measurements can be considered as the real and imaginary part of a complex number, i.e.

$$\mathcal{V} = r_R + jr_I \tag{9.2.12}$$

Thus the pair of multipliers together with an integrator (to get the time average) form the basic element of a *complex correlator*.

In the above analysis a narrow band signal (quasi monochromatic) is considered. In an actual interferometer the observations are made over a finite bandwidth $\Delta\nu$ and one requires the complex visibilities to be measured as a function of frequencies within $\Delta\nu$. This can be achieved in one of the two ways described below.

9.2.1 FX Correlator

The band limited signal can be decomposed into spectral components using a filter bank. The spectral visibility is then obtained by separately cross correlating each filter output using a complex correlator (see Fig. 9.3). The digital implementation of this method is called an *FX correlator* (F for Fourier Transform and X for multiplication or correlation). The GMRT correlator is an FX correlator. A schematic of an FX correlator is shown in Fig. 9.4. The analog voltages $V_1(t)$ and $V_2(t)$ are digitized first using ADCs. The geometric delay in steps of the sampling intervals (integral delay) are then compensated for. The

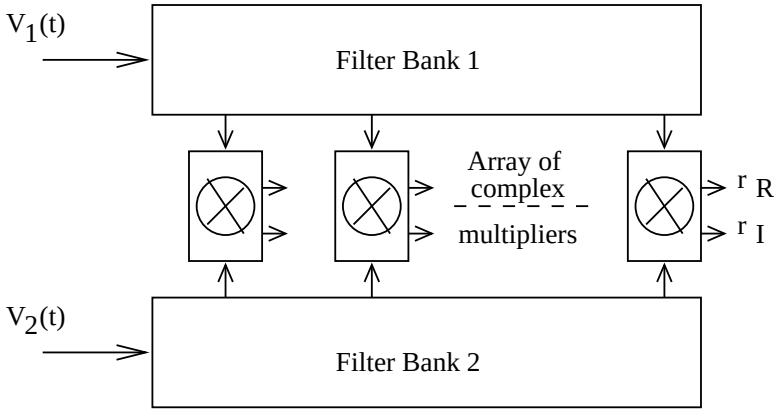


Figure 9.3: A spectral correlator using filter bank and complex multipliers.

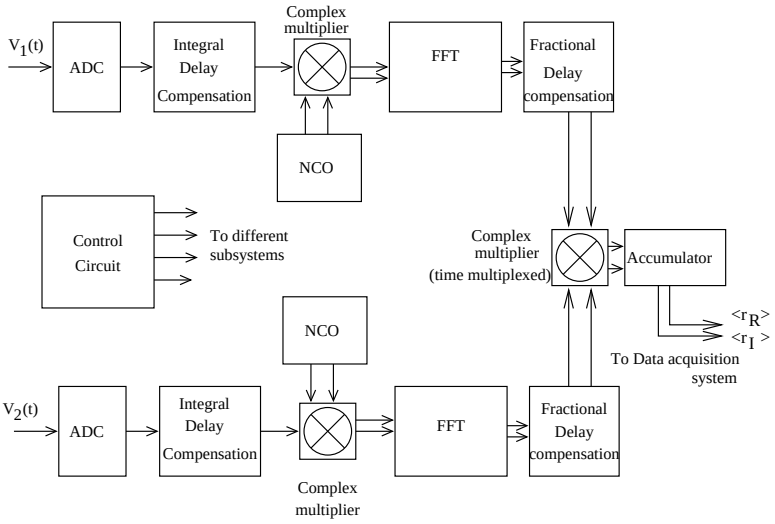


Figure 9.4: Block diagram of an FX correlator.

integral delay compensated samples are multiplied by the output of NCO for fringe stopping. The samples from each antenna are then passed through an FFT block to realize a filter bank. Phase gradients are applied after taking the Fourier Transform for fractional delay compensation. The spectral visibility is then measured by multiplying the spectral components of one antenna with the corresponding spectral components of other antennas. These are then integrated for some time to get an estimate of the cross correlation. Since the Fourier transform is taken before multiplication it is called an FX correlator. For continuum observations with an FX correlator the visibility measured from all spectral components can be averaged after bandpass calibration.

9.2.2 XF Correlator

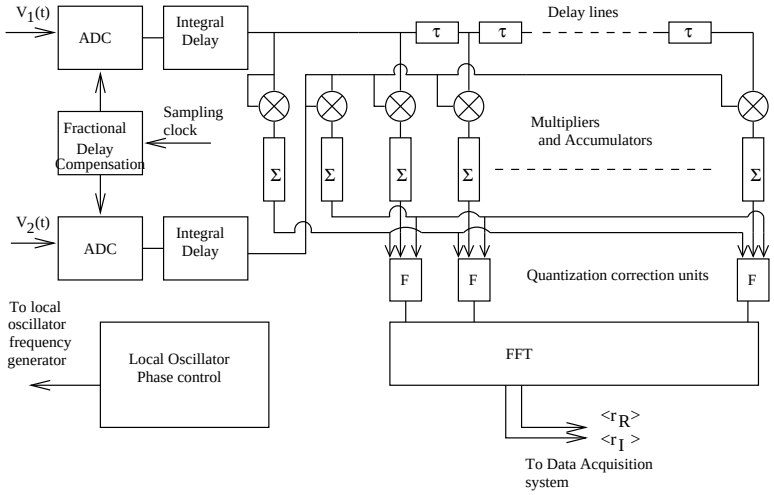


Figure 9.5: Block diagram of a XF correlator.

Eq. 9.0.1 for a broadband signal after delay compensation and integration (time average) can be written as

$$< r_R >= \mathrm{Re}\Big[\int_{-\infty}^{+\infty} < v_1(\nu,t)v_2^*(\nu,t) > \mathrm{d}\nu\Big] \ , \tag{9.2.13}$$

where $v_1(\nu,t)$ and $v_2(\nu,t)$ can be considered as the spectral components of the signals from the antennas. Introducing a delay of τ to one of the signals $v_1(\nu,t)$ modifies the above equation to

$$< r_R(\tau) >= \mathrm{Re}\Big[\int_{-\infty}^{+\infty} < v_1(\nu,t)v_2^*(\nu,t) > e^{-j2\pi\nu\tau}\mathrm{d}\nu\Big] \tag{9.2.14}$$

The above equation is a Fourier Transform equation; the Fourier Transform of the cross spectrum $< v_1(\nu,t)v_2^*(\nu,t) >$ (averaging over t). Thus $< r_R(\tau) >$ is the cross correlation *measured as a function of* τ. Since $v_1(\nu,t)$ and $v_2^*(\nu,t)$ are Hermitian functions, as they are spectra of real signals, their product is also hermitian. Therefore $< r_R(\tau) >$ is a real function and hence it can be measured with a simple correlator (not a complex correlator). Thus the second method of measuring spectral visibility is to measure $< r_R(\tau) >$ for each pair of antennas as a function of τ and later perform an Fourier Transform to get the cross spectrum. The digital implementation of this method is called an *XF correlator*.

A block diagram of an XF correlator is shown in Fig. 9.5. In this diagram, fractional delays are compensated for by changing the phase of the sampling clock. After delay compensation, the cross correlations for different delay are measured using delay lines and multipliers, which are followed by integrator. Since the cross correlation function in general is not an even function of τ, the delay compensation is done such that the correlation function is measured for both positive and negative values of τ in the correlator. The zero lag autocorrelations of the signals are also measured, which is used to normalize the cross correlation. The quantization correction (block marked as F) is then applied to the normalized cross correlations. The cross spectrum is obtained by performing a DFT

on the corrected cross correlation function. A peculiarity of this implementation is that the correlations are measured first and the Fourier Transform is taken later to get the spectral information. Hence it is called an XF correlator.

9.3 Further Reading

1. Thompson, R.A., Moran, J.M., Swenson, Jr. G.W., "Interferometry and Synthesis in Radio Astronomy", Chapter 8, John Wiley & Sons, 1986.

2. Thompson, A.R. & D'Addario, L.R. in "Synthesis Imaging in Radio Astronomy", R.A. Perley, F.R. Schwab, & A.H. Bridle, eds., ASP Conf. Series, vol. 6.

Chapter 10

Mapping I

Sanjay Bhatnagar

In the Chapters 2 & 4, the conceptual basis and formulation of aperture synthesis in Radio Astronomy has been described. In particular, it has been shown that (1) an interferometer records the mutual coherence function, also called the visibility of the signals from the sky, and (2) the visibility is the Fourier transform of the sky brightness distribution. This chapter describes the coordinate systems used in practical aperture synthesis in Radio Astronomy and presents the derivation of the 2D Fourier transform relation between the visibility and the brightness distribution.

10.1 Coordinate Systems

10.1.1 Angular Co-ordinates

As described in Chapter 4, the response of an interferometer to quasi-monochromatic radiation from a point source located at the phase center is given by

$$r(\tau(t)) = cos(2\pi\nu_o\tau), \tag{10.1.1}$$

where $\tau = \tau_o = (D/c)sin(\theta(t))$ is the geometrical delay, θ is the direction which the antennas are tracking with respect to the vertical direction, λ is the wavelength, ν_o is the center frequency of the observing band and D is the separation between the antennas. As the antennas track the source, the geometrical delay changes as a function of time. This changing τ is exactly compensated with a computer controlled delay and for a point source at the phase center, the output of the interferometer is the amplitude of the fringe pattern.

For a source located at an angle $\theta = \theta_o + \Delta\theta$, for small $\Delta\theta$, $\tau = \tau_o + \Delta\theta(D/c)cos(\theta(t))$. Since fringe stopping compensates for τ_o, the response of the interferometer for a source $\Delta\theta$ away from the phase center is $cos(2\pi\Delta\theta D_\lambda cos(\theta))$ where $D_\lambda = D/\lambda$. If the phase center is shifted by equivalent of $\lambda/4$, the interferometer will pick up an extra phase of $\pi/2$ and the response will be sinusoidal instead of co-sinusoidal. Hence, an interferometer responds to both even and odd part of the brightness distribution. Interferometer response can then be written in complex notation as

$$r(\tau(t)) = e^{2\pi\iota\Delta\theta D_\lambda cos(\theta)}. \tag{10.1.2}$$

Writing $u = D_\lambda cos(\theta)$, which is the projected separation between the antennas in units of wavelength in the direction normal to the phase center and $l = sin(\Delta\theta) \approx \Delta\theta$, we get

$$r(u,l) = e^{2\pi\iota ul} = e^{2\pi\iota u\Delta\theta} \tag{10.1.3}$$

as the complex response of a two element interferometer for a point source of unit flux located $\Delta\theta$ away from the phase center given by the direction θ_o.

Usually the phase center coincides with the center of the field being tracked by all the antennas. Let the normalized power reception patter of antennas (which are assumed to be identical) at a particular frequency be $B(\Delta\theta)$ and the surface brightness of an extended source be represented by $I(\Delta\theta)$. The response of the interferometer to a point source located $\Delta\theta$ away from the phase center would then be $I(\Delta\theta)B(\Delta\theta)e^{2\pi\iota u\Delta\theta}$. For an extended source with a continuous surface brightness distribution, the response is given by

$$V(u) = \int B(\Delta\theta)I(\Delta\theta)e^{2\pi\iota u\Delta\theta}d\Delta\theta = \int B(l)I(l)e^{2\pi\iota ul}dl. \tag{10.1.4}$$

The above equation is a 1D Fourier transform relation between the source brightness distribution and the output of the visibility function V. The integral is over the entire sky visible to the antennas but is finite only for a range of l limited by the antenna primary reception pattern $B(l)$. In practice, u is calculated as a function of the source position in the sky, specified in astronomical co-ordinate system, as seen by the observer on the surface of the earth.

l in the above equation is the direction of the elemental source flux relative to the pointing center. u then has the interpretation of spatial frequency and $V(u)$ represents the 1D spatial frequency spectrum of the source.

10.1.2 Astronomical Co-ordinate System

The position of a source in the sky can be specified in various spherical coordinates systems in astronomy, differing from each other by the position of the origin and orientation of the axis. The position of the sources are specified using the azimuth and elevation angles in these coordinate systems. In the Equatorial Co-ordinate system the source position is specified by the Declination (δ) which is the elevation of the source from the normal to the celestial equator and the Right Ascension (RA), which is the azimuthal anlge from a reference position ("the first point of Aries"). The reference direction for RA is line of intersection of the equatorial and Ecliptic planes. The position of the source in the sky, in this coordinate system, remains constant as earth rotates. The azimuth and elevation of the antennas, which rotate with earth, are constantly adjusted to track a point in the sky specified by (RA, δ) coordinates. The changing position of the sources in the sky, as seen by the observer on the surface of earth is specified by replacing RA by Hour Angle (HA), which is the azimuth of the source measured in units of time, with respect to the local meridian of the source with $HA = -6^h$ pointing due East.

10.1.3 Physical Coordinate System

The antennas are located on the surface and rotate with respect to a source in the sky due the rotation of the earth. For aperture synthesis the antenna positions are specified in a co-ordinate system such that the separation of the antennas is the projected separation in plane normal to the phase center. In other words, in such a co-ordinate system the separation between the antennas is as seen by the observer sitting in the source reference frame. This system, shown in Fig 10.1, is the right-handed (u, v, w) coordinate system

fixed on the surface of the earth at the array reference point, with the (u, v) plane always parallel to the tangent plane in the direction of phase center on the celestial sphere and the w axis along the direction of phase center. The u axis is along the astronomical E-W direction and v axis is along the N-S direction. The (u, v) co-ordinates of the antennas are the E-W and N-S components of position vectors. As the earth rotates, the (u, v) plane rotates with the source in the sky, changing the (u, v, w) coordinates of the antennas, generating tracks in the uv-plane.

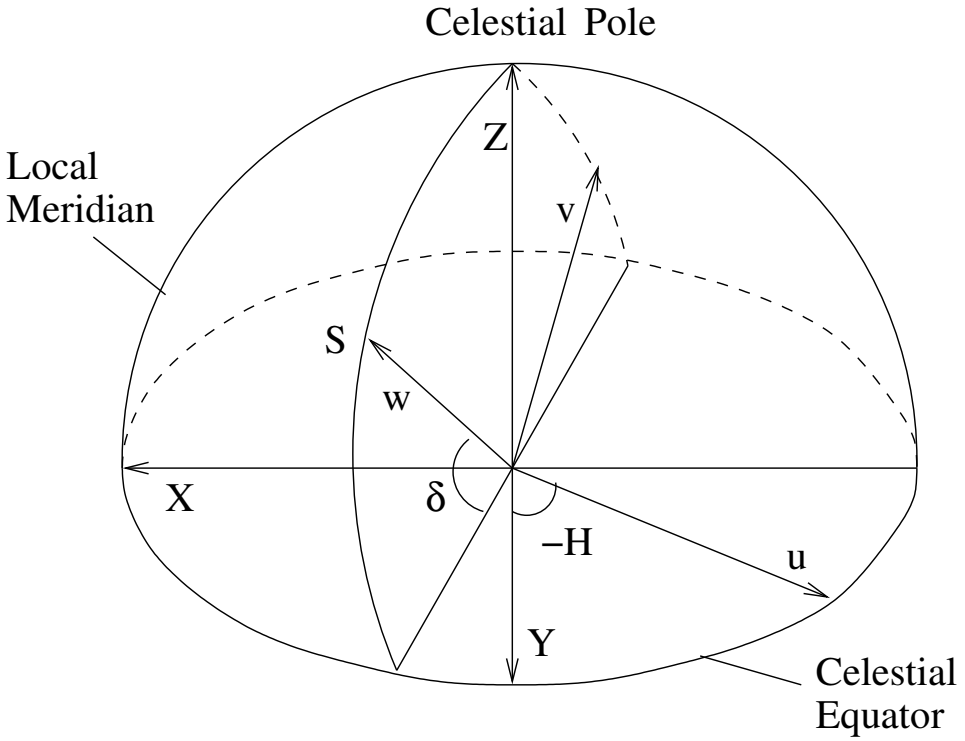


Figure 10.1: Relationship between the terrestrial co-ordinates (X,Y,Z) and the (u, v, w) co-ordinate system. The (u, v, w) system is a right handed system with the w axis pointing to the source S.

In the above formulation, the u co-ordinate of one antenna is with respect to the other antenna making the interferometer, which is located at the origin. If the origin is arbitrarily located and the co-ordinates of the two antennas are u_1 and u_2, Eq. 10.1.3 becomes

$$r(u, l) = e^{2\pi\iota(u_1 - u_2)l}. \tag{10.1.5}$$

Since only the relative positions of the antennas with respect to each other enter the equations, it is only useful to work with difference between the position vectors of various antennas in the (u, v, w) co-ordinate system. The relative position vectors are called "Baseline vectors" and their lengths referred to as "baseline length".

The source surface brightness distribution is represented as a function of the direction cosines in the (u, v, w) coordinate system. In Eq. 10.1.4 above, l is the direction cosine. The source coordinate system, which is flat only for small fields of view, is represented by (l, m, n). Since this coordinate system represents the celestial sphere, n is not an independent coordinate and is constrained to be $n = \sqrt{1 - l^2 - m^2}$.

10.1.4 Coordinate Transformation

To compute the (u, v, w) co-ordinates of the antennas, the antenna locations must first be specified in a terrestrial co-ordinate system. The terrestrial coordinate system generally used to specify the position of the antennas is a right-handed Cartesian coordinate system as shown in Figure 10.2. The (X, Y) plane is parallel to the earth's equator with X in the meridian plane and Y towards east. Z points towards the north celestial pole. In terms of the astronomical coordinate system (HA, δ), $X = (0^h, 0^o)$, $Y = (-6^h, 0^o)$ and $Z = (\delta = 90^o)$. If the components of $\bar{D}_\lambda$ are $(X_\lambda, Y_\lambda, Z_\lambda)$, then the components in the (u, v, w) system can be expressed as

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} sin(HA) & cos(HA) & 0 \\ -sin(\delta)cos(HA) & sin(\delta)sin(HA) & cos(\delta) \\ cos(\delta)cos(HA) & -cos(\delta)sin(HA) & sin(\delta) \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}. \tag{10.1.6}$$

As earth rotates, the HA of the source changes continuously, generating different set of (u, v, w) co-ordinates for each antenna pair at each instant of time. The locus of projected antenna-spacing components u and v defines an ellipse with hour angle as the variable given by

$$u^2 + \left(\frac{v - Zcos\delta_o}{sin\delta_o}\right)^2 = X^2 + Y^2, \tag{10.1.7}$$

where (HA_o, δ_o) defines the direction of phase center. In the uv-plane, this is an ellipse, referred to as the uv-track with HA changing along the ellipse. The pattern generated by all the uv points sampled by the entire array of antennas over the period of observation is referred to as the uv-*coverage* and as is clear from the above transformation matrix, is different for different δ. Examples of uv-coverage for a few declinations for full synthesis with GMRT array are shown in Figure 10.4.

The uv domain is the spatial frequency domain and uv-coverage represent the spatial frequencies sampled by the array. The shorter baselines (uv points closer to the origin) provide the low resolution information about the source structure and are sensitive to the large scale structure of the source while the longer baselines provide the high resolution information. GMRT array configuration was designed to have roughly half the antennas in a compact "Central Square" to provide the shorter spacings information, which is crucial mapping extended source and large scale structures in the sky. The uv-coverage of the central square antennas is shown in Figure 10.5. Notice that there are no measurements for $(u = 0, v = 0)$. $V(0, 0)$ represents the total integrated flux received by the antennas and is absent in the visibility data. Effect of this on the image will be discussed later.

The astronomical coordinates depend on the line of intersection of the ecliptic and equatorial planes. The uv-coverage in turn depends on the position of the source in the astronomical coordinate system. Since the reference line of the this coordinate system changes because of the well known precession of the earth's rotation axis, the uv-coverage also becomes a function of the reference epoch for which the source position is specified. For the purpose of comparison and consistence in the literature, all source positions are specified in standard epochs (B1950 or J2000). Since each point in the (u, v, w) plane measures a particular spatial frequency and this spatial frequency coverage differs from one epoch to another, it's necessary to precess the source coordinates to the current epoch (also called the "date coordinates") prior to observations and all processing of the visibility data for the purpose of mapping must be done with (u, v, w) evaluated for the epoch of observations. Precessing the visibilities to the standard epoch prior to inverting the Eq. 10.2.10 will require specifying the real and imaginary parts of the visibility at

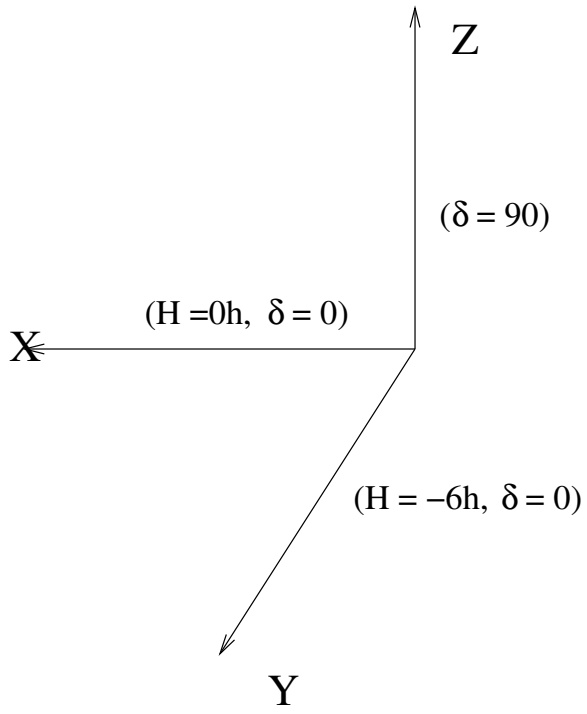


Figure 10.2: The (X,Y,Z) co-ordinate system used to specify antenna locations.

(u, v, w) coordinates which are in fact not measured (since the uv-coverage changes with epoch) introducing errors in the mapping procedure.

10.2 2D Relation Between Sky and Aperture Planes

Below, we derive the generalized 2D Fourier transform relation between the visibility and the source brightness distribution in the (u, v, w) system. The geometery for this derivation is shown in Fig 10.3.

Let the vector $\bar{L}_o$ represent the direction of the phase center and the vector $\bar{D}_\lambda$ represent the location of all antennas of an array with respect to a reference antenna. Then $\tau_g = \bar{D}_\lambda.\bar{L}_o$. Note that $2\pi\bar{D}_\lambda.\bar{L}_o = 2\pi w$ is phase by which the visibility should be rotated to stop the fringe. For any source in direction $\bar{L} = \bar{L}_o + \bar{\sigma}$, the output of an interferometer after fringe stopping will be

$$V(\bar{D}_\lambda) = \int_{4\pi} I(\bar{L})B(\bar{L})e^{2\pi\iota\bar{D}_\lambda.(\bar{L}-\bar{L}_o)}d\Omega. \tag{10.2.8}$$

The vector $\bar{L} = (l, m, n)$ is in the sky tangent plane, $\bar{L}_o$ is the unit vector along the w axis and $\bar{D}_\lambda = (u, v, w)$. The above equation can then be written as

$$V(u, v, w) = \int\int I(l, m)B(l, m)e^{2\pi\iota(ul+vm+w(\sqrt{1-l^2-m^2}-1))}\frac{dldm}{\sqrt{1-l^2-m^2}}. \tag{10.2.9}$$

If the array is such that all antennas are exactly located in the (u, v) plane, w is exactly zero and the above equation reduces to an exact 2D Fourier transform relation between

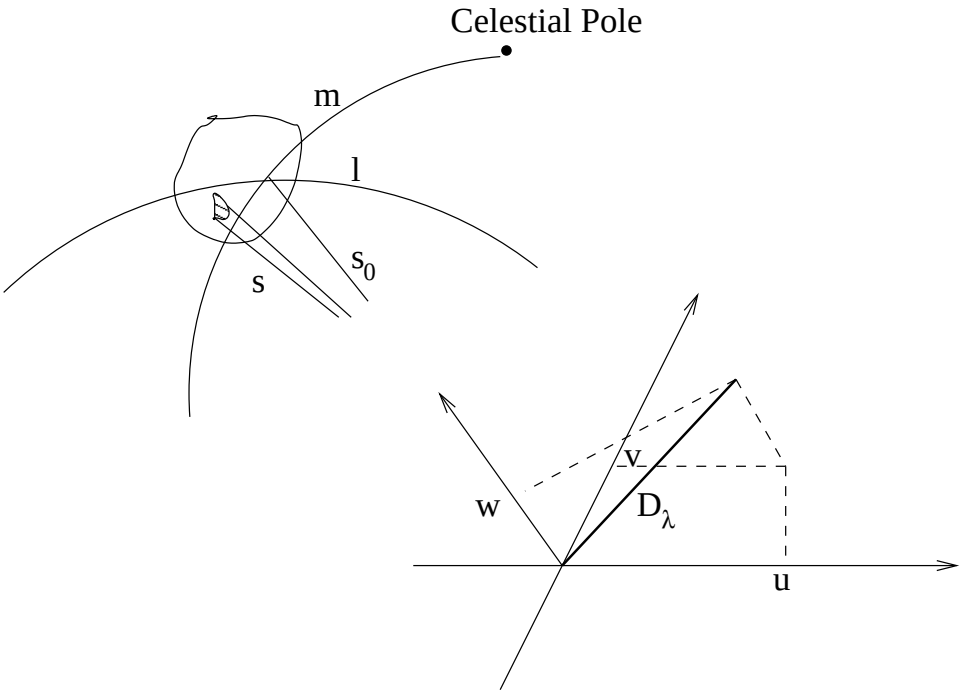


Figure 10.3: Relationship between the (l, m) co-ordiantes and the (u, v, w) co-ordinates

the source brightness distribution and the visibility. This is true for a perfect east-west array (like WSRT or ATCA). However to maximize the uv-coverage arrays like GMRT or VLA are not perfectly east-west. As mentioned earlier, the integrals in the above equation are finite for a small portion of the sky (being limited by the primary beam patter of the antennas). If the field of view being mapped is small (ie. for small l and m) $\sqrt{1-l^2+m^2}-1 \approx -\frac{1}{2}(l^2+m^2)$ and can be neglected. Eq. 14.1.1 becomes

$$V(u,v,w) \approx V(u,v,0) = \int\int I(l,m)B'(l,m)e^{2\pi\iota(ul+vm)}dldm. \tag{10.2.10}$$

where $B' = B/\sqrt{1-l^2-m^2}$. Neglecting the w-term puts restrictions on the field of view that can be mapped without being effected by the phase error which is approximately equal to $\pi(l^2+m^2)w$. Eq. 10.2.10 shows the 2D Fourier transform relation between the surface brightness and visibility.

Since there are finite number of antennas in an aperture synthesis array, the uv-coverage is not continuous. Let

$$S(u,v) = \begin{array}{l} 1, \text{ for all measured (u,v) points} \\ 0, \text{ every where else.} \end{array} \tag{10.2.11}$$

Then to represent the real life situation, assuming that $B(l,m) = 1$ over the extent of the source, Eq. 10.2.10 becomes

$$V(u,v)S(u,v) = \int\int I(l,m)e^{2\pi\iota(ul+vm)}dldm. \tag{10.2.12}$$

Inverting the above equation and using the convolution theorem, we get $I^D = I * DB$ where DB is the Fourier transform of S. DB is the transfer function of the the telescope

for imaging and is referred to as the *Dirty Beam*. I^D represents the raw image produced by an earth rotation aperture synthesis telescope and is referred to as the *Dirty Map*. Contribution of *Dirty Beam* to the map and methods of removing these these effects will be discussed in greater detail in later lectures.

In all the above discussion, we have assumed the observations are monochromatic with negligible frequency bandwidth and that the (u, v) measurements are instantaneous measurements. None of these assumptions are true in real life. Observations for continuum mapping are made with as large a frequency bandwidth as possible (to maximize the sensitivity) and the data is recorded after finite integration. Both result into degradation in the map plane and these effects will be discussed in the later chapters.

Neglecting the w-term essentially implies that the source brightness distribution is approximated to be restricted to the tangent plane at the phase center in the sky rather than on the surface of the celestial sphere. At low frequencies, where the antenna primary beams are larger and the radio emission from sources is also on a larger scale, this assumption restricts the mappable part of the sky to a fraction of the primary beam. Methods to relax this assumption will also be discussed in a later lecture.

10.3 Further Reading

1. Interferometry and Synthesis in Radio Astronomy; Thompson, A. Richard, Moran, James M., Swenson Jr., George W.; Wiley-Interscience Publication, 1986.

2. Synthesis Imaging In Radio Astronomy; Eds. Perley, Richard A.,Schwab, Frederic R., and Bridle, Alan H.; ASP Conference Series, Vol 6.

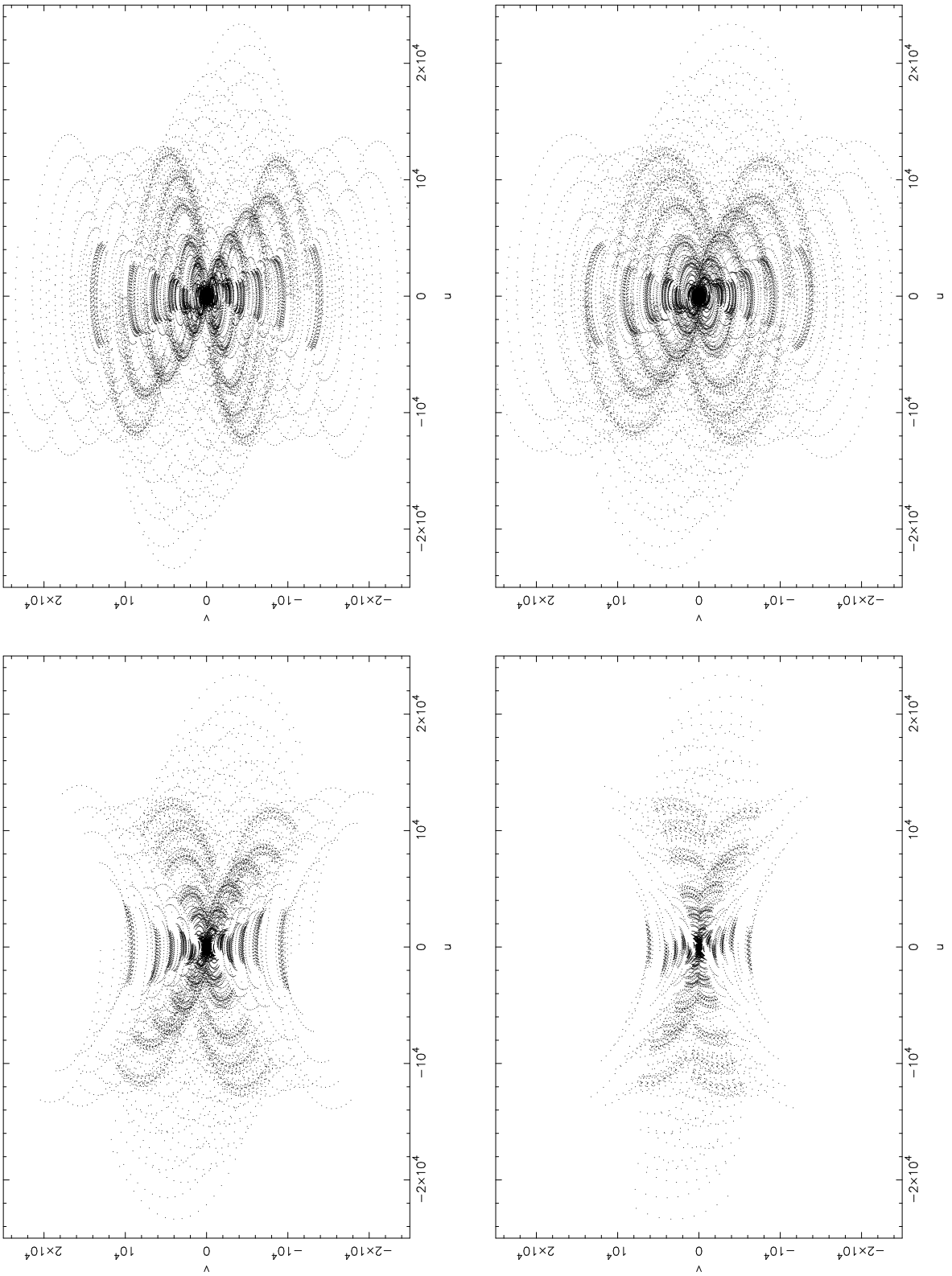


Figure 10.4: uv-coverage for a 10^h synthesis with the full GMRT array at δ of 19^o, 30^o, -30^o and -45^o. The u and v axes are in meters.

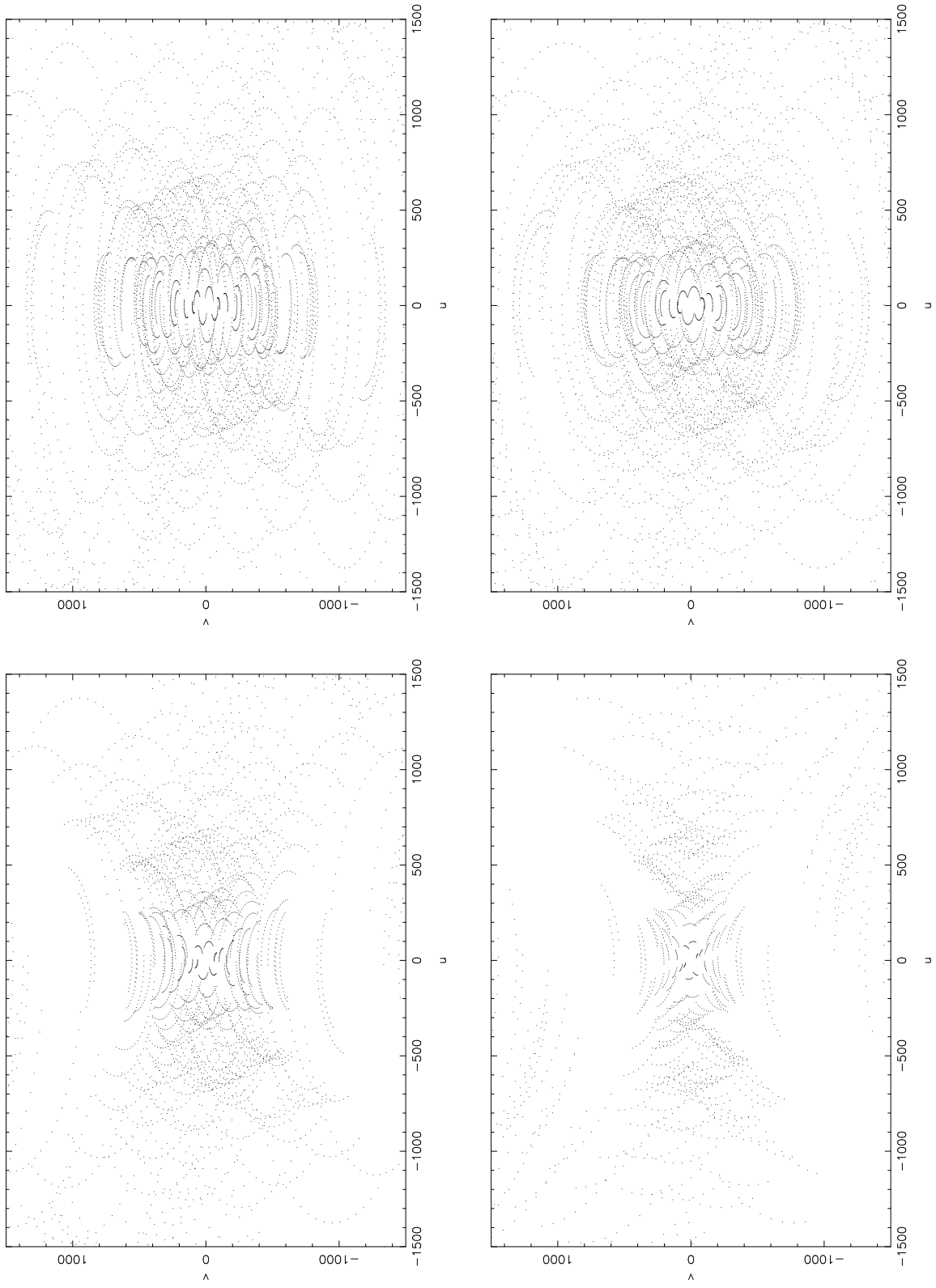


Figure 10.5: uv-coverage for a 10^h synthesis with GMRT Central Square at δ of 19^o, 30^o, -30^o and -45^o. The u and v axes are in meters.

Chapter 11

Mapping II

Sanjay Bhatnagar

11.1 Introduction

An aperture synthesis array measures the visibilities at discrete points in the uv-domain. The visibilities are Fourier transformed to get the *Dirty Map* and the weighted uv-sampling function is Fourier transformed to get the *Dirty Beam* using the efficient FFT algorithm. This lecture describes the entire chain of data processing required to inverted the visibilities recorded as a function of (u, v, w), and the resulting errors/distortions in the final image. In this entire lecture, the '$\star$' operator represents convolution operation, the '.' operator represents point-by-point multiplication and the '$\rightleftharpoons$' operator represents the Fourier transform operator.

As described earlier, the visibility V measured by an aperture synthesis telescope is related to the sky brightness distribution I as

$$V \rightleftharpoons I, \tag{11.1.1}$$

where $\rightleftharpoons$ denotes the Fourier Transform. The above equation is true only for the case of continuous sampling of the uv-plane such that V is measured for all values of (u, v). However since there are finite antennas in an array, uv-plane is sampled at descreet uv points and Eq. 11.1.1 has to be written as

$$V.S \rightleftharpoons I * DB(= I^d), \tag{11.1.2}$$

where I^d is the *Dirty Map*, I is the true brightness distribution, DB is the *Dirty Beam* and S is the uv-sampling function given by

$$S(u, v) = \sum_k \delta(u - u_k, v - v_k), \tag{11.1.3}$$

where u_k and v_k are the actual (u, v) points measured by the telescope. The pattern of all the measured (u, v) points is referred to as the uv-coverage.

This function essentially assigns a weight of unity to all measured points and zero everywhere else in the uv-plane. Fourier transform of S is referred to as the *Dirty Beam*. As written in Eq. 11.1.2, the *Dirty Beam* is the transfer function of the instrument used as an imaging device. The shape of the *Dirty Beam* is a function of the uv-coverage which

in turns is a function of the location of the antennas. *Dirty Beam* for a fully covered uv-plane will be equal to $sin(\pi l\lambda/u_{max})/(\pi l\lambda/u_{max})$ where u_{max} is the largest antenna spacing for which a measurement is available. The width of the main lobe of this function is proportional to λ/u_{max}. The resolution of such a telescope is therefore roughly λ/u_{max} and u_{max} can be interpreted as the size of an equivalent lens. For a real uv-coverage however, S is not flat till $u_{\rm max}$ and has 'holes' in between representing un-sampled (u, v) points. The effect of this missing data is to increase the side-lobes and make the *Dirty Beam* noisy, but in a deterministic manner. Typically, an elliptical gaussian can be fitted to the main lobe of the *Dirty Beam* and is used as the resolution element of the telescope. The fitted gaussian is referred to as the *Synthesized Beam*.

The *Dirty Map* is a convolution of the true brightness distribution and the *Dirty Beam*. I^d is almost never a satisfactory final product since the side-lobes of DB (which are due to missing spacings in the uv-domain) from a strong source in the map will contaminate the entire map at levels higher than the thermal noise in the map. Without removing the effect of DB from the map, the effective RMS noise in the map will be much higher than the thermal noise of the telescope and will result into obscuration of faint sources in the map. This will be then equivalent to reduction in the dynamic range of the map. The process of De-convolving is discussed in a later lecture, which effectively attempts to estimate I from I^d such that $(I - I^d) * DB$ is minimized consistent with the estimated noise in the map.

To use the FFT algorithm for Fourier transforming, the irregularly sampled visibility $V(u, v)$ needs to be gridded onto a regular grid of cells. This operation requires interpolation to the grid points and then re-sampling the interpolated function. To get better control on the shape of the *Dirty Beam* and on the signal-to-noise ratio in the map, the visibility is first re-weighted before being gridded. These operations are described below.

11.2 Weighting, Tapering and Beam Shaping

The shape of the *Dirty Beam* can be controlled by multiplying S with other weighting functions. Note that the measured visibilities already carry a weight which is a measure of the signal-to-noise ratio of each measurement. Since there is no control on this weight while mapping, it is not explicitly written in any of equations here but is implicitly used.

Full weighting function W as used in practice is given by

$$W(u, v) = \sum_k T_k D_k \delta(u - u_k, v - v_k). \tag{11.2.4}$$

The function T_k is the 'uv-tapering' function to control the shape of DB and D_k is the 'density-weighting' function used in all imaging programs. If S was a smooth function, going smoothly to zero beyond the maximum sampled uv-point, DB would also be smooth with no side lobes (e.g. if S was a gaussian). However, S is collection of delta functions with gaps in between (for the missing uv-points not measured by the telescope) and has a sharp cut-off at the limit of uv-coverage. This results into DB being a highly non-smooth function with potentially large side-lobes.

As is evident from the plots of uv-coverage, the density of uv-tracks decreases away from the origin. If one were to use the local average of the uv-points in the uv-plane for mapping as is done in the gridding operation described below, the signal-to-noise ratio of the points would be proportional to the number of uv-points averaged. Since the density of measured uv-points is higher for smaller values of u and v, visibilities for shorter spacings get higher weightage in the visibility data effectively making the array

relatively more sensitive to the broader features in the sky. The function D_k controls the weights resulting from non-uniform density of the points in the uv-plane.

Both T_k and D_k provide some control over the shape of the *Dirty Beam*. T_k is used to weight down the outer edge of the uv-coverage to decrease the side-lobes of DB at the expense decreasing the spatial resolution. D_k is used to counter the preferential weight that the uv-points get closer to the origin at the expense of degrading the signal-to-noise ratio.

T_k is a smoothly varying function of (u, v) and is often used as $T(u_k, v_k) = T(u_k)T(v_k)$. For most imaging applications, $T(u_k, v_k)$ is a circularly symmetric gaussian. However other forms are also occasionally used.

Two forms of D_k are commonly used. When $D_k = 1$ for all values of (u, v), it is referred to as 'natural weighting' were the natural weighting of the uv-coverage is used as it is. This gives best signal-to-nose ratio and is good when imaging weak compact sources but is undesirable for extended sources where both large scale and small scale features are present.

When $D_k = 1/N_k$ where N_k is a measure of the local density of uv-points around (u_k, v_k), it is referred to as 'uniform weighting' where an attempt is made to assign uniform weights to the entire covered uv-plane. In standard data reduction packages available for use currently (*AIPS, SDE, Miriad*), while re-gridding the visibilities (discussed below), N_k is equal the number of uv-points within a given cell in the uv-plane. However it can be shown that this can result into serious errors, referred to as *catastrophic gridding error* in some pathological cases. This problem can be handled to some extend by using better ways of estimating the local density of uv-points (Briggs, 1995).

Eq. 11.1.2, using the weighted sampling function W is written as

$$(V.S.W) \rightleftharpoons (I * DB). \tag{11.2.5}$$

Note that $DB \rightleftharpoons S.W$, i.e. the *Dirty Beam* is the Fourier transform of the weighted sampling function.

11.3 Gridding and Interpolation

The inversion of the visibilities to make the *Dirty Map* is done using FFT algorithm which requires that the function be sampled at regular intervals and the number of samples be power of 2. For the case of mapping the sky using an aperture synthesis telescope, this implies that the visibility data be available on a regular 2D grid in the uv plane. Thus re-gridding of the data onto a regular grid is required by potentially interpolating the visibility to the grid points, since the visibility function $V(u, v)$ is measured at discrete points (u, v) which are not assured to be at regular intervals along the u and v axis.

Interpolation of V is done by multiplying a function and averaging all the measured points which lie within the range of the function with a finite support base, centered at each grid point. The resultant average value is assigned to the corresponding grid point. This operation is equivalent to discrete convolution of V with the above mentioned function and then sampling this convolution at the grid points. The convolving function is referred to as the Gridding Convolution Function (GCF). There are other ways of doing this interpolation. However the interpolation in practice is done by convolution since this results into predictable results in the map plane which are easy to visualize. Also using GCF with finite support base results into each grid point getting the value of the local average of the visibilities.

After gridding Eq. 11.2.5 becomes

$$(V.S.W) * C \rightleftharpoons (I * DB).c, \tag{11.3.6}$$

where C represents the GCF and $c \rightleftharpoons C$.

The effect of gridding the visibilities on the map is to multiply the map with function c and since C has a finite support base (i.e. is of finite extent), c is infinite in extent which result into aliasing in the map plane (the other cause of aliasing could be under-sampling of the uv-plane). The amplitude of the aliased component from a position (l, m) in the map is determined by $c(l, m)$. Ideally therefore, this function should be rectangular function with the width equal to the size of the map and smoothly going to zero immediately outside the map. However from the point of efficiency of the gridding process, this is not possible, and GCF used in practice have a trade-off between the roll-off properties at the edge and flatness within the map.

Since the *Dirty Map* is multiplied by c, if c is well known, then effect of convolution by the GCF can be removed by point-wise division of *Dirty Map* and *Dirty Beam* given by $\bar{I}^d = I^d/c$ and $\bar{DB} = DB/c$ for later processing, particularly in deconvolution of I^d. In practice however, this division is not carried out by evaluating $c(l, m)$ over the map. Instead, for efficiency purposes, this function is kept in the computer memory tabulated with a resolution typically $1/100$ times the size of the cell in the image.

To take the Fourier transform of $(V.S.W) * C$ using the FFT algorithm one needs to sample the right hand side of Eq. 11.3.6 by multiplication with the re-sampling function R given by

$$R(u, v) = \sum_{j=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} \delta(j - u/\Delta u, k - v/\Delta v), \tag{11.3.7}$$

where Δu and Δv are the cell size in the uv-domain. Eq. 11.3.6 then becomes

$$R.((V.S.W) * C) \rightleftharpoons r * ((I * DB).c), \tag{11.3.8}$$

where $R \rightleftharpoons r$. Right hand side of this equation then is the approximation of I^d obtained in practice. As discussed in earlier lecture, FFT generates a periodic function (due to the presence of R in the left hand side of Eq. 11.3.8) and I^d represents one period of such a function. To map an angular region of sky of size $(N_l\Delta l, N_m\Delta m)$, using the Nyquist sampling theorem we get $N_l\Delta u = 1/\Delta l$ and $N_m\Delta v = 1/\Delta m$ where Δl and Δm is the cell size in the map and Δu and Δv are cell sizes in the uv-plane.

C is usually real and even and is assumed to be separable as $C(u, v) = C_1(u)C_2(v)$. Various GCFs used in practice are listed below. Functions listed below are in 1-dimension and are truncated (set to zero) for $|u| \geq m\Delta u/2$ where Δu is the size of the grid and m is the number of such cells used.

1. 'Pillbox' function

$$C(u) = \left\{ \begin{array}{c} 1,\ |u| < m\Delta u/2 \\ 0,\ otherwise \end{array} \right\}. \tag{11.3.9}$$

   This amounts to simple averaging of all the uv-points with in the rectangle defined by Eq. 11.3.9. However since its Fourier transform is $sinc$ with large side lobes, it provides poor alias rejection and is almost never used but is useful for intuitive understanding.

2. Truncated exponential function

$$C(u) = e^{\frac{-|u|^\alpha}{w\Delta u}}. \tag{11.3.10}$$

Typically $m = 6$, $w = 1$ and $\alpha = 2$ is used and c can be expressed in terms of error function.

3. Truncated $sinc$ function

$$C(u) = sinc\left(\frac{u}{w\Delta u}\right). \tag{11.3.11}$$

For $m = 6$ and $w = 1$, this is the normal $sinc$ function expressed in terms of sin function. As m increases, the Fourier transform of this function approaches a step function which is constant over the map and zero outside.

4. Sinc exponential function

$$C(u) = e^{\frac{-|u|^\alpha}{w_1\Delta u}} sinc\left(\frac{u}{w_2\Delta u}\right). \tag{11.3.12}$$

For $m = 6$, $w_1 = 2.52$, $w_2 = 1.55$, $\alpha = 2$, the above equation reduces to multiplication of gaussian with the exponential function. This optimizes between the flat response of exponential within the map and suppression of the side-lobes due the presence of the gaussian.

5. Truncated spheroidal function

$$C(u) = |1 - \eta^2(u)|^\alpha \phi_{\alpha 0}(\pi m/2.\eta(u)), \tag{11.3.13}$$

where $\phi_{\alpha 0}$ is the 0-order spheroidal function, $\eta(u) = 2u/m\Delta u$ and $\alpha > -1$.

Of all the square integrable functions, this is the most optimal in the sense that it has maximum contribution to the normalized area from the part of $c(l)$ which is with in the map. This is referred to as the *energy concentration ratio* expressed as $\frac{\int_{map} |c(l)|^2 dl}{\int_{-\infty}^{\infty} |c(l)|^2 dl}$ is maximized.

11.4 Bandwidth Smearing

The effect of a finite bandwidth of observation as seen by the multiplier in the correlator, is to reduce the amplitude of the visibility by a factor given by $sin(\pi l\Delta\nu/\nu_o\theta)/(\pi l\Delta\nu/\nu_o\theta)$, where θ is angular size of the synthesized beam, ν_o is the center of the observing band, l is location of the point source relative to the field center and $\Delta\nu$ is the bandwidth of the signal being correlated.

The distortion in the map due to the finite bandwidth of observation can be visualized as follows. For continuum observations, the visibility data integrated over the bandwidth $\Delta\nu$ is treated as if the observations were made at a single frequency ν_o, the central frequency of the band. As a result the u and v co-ordinates and the value of visibilities are correct only for ν_o. The true co-ordinate at other frequencies in the band are related to the recorded co-ordinates as

$$(u, v) = \left(\frac{\nu_o u_\nu}{\nu}, \frac{\nu_o v_\nu}{\nu}\right). \tag{11.4.14}$$

Since the total weights W used while mapping does not depend on the frequency, the relation between the brightness distribution and visibility at a frequency ν becomes

$$V(u,v) = V\left(\frac{\nu_o u_\nu}{\nu}, \frac{\nu_o v_\nu}{\nu}\right) \rightleftharpoons \left(\frac{\nu}{\nu_o}\right)^2 I\left(\frac{l\nu}{\nu_0}, \frac{m\nu}{\nu_0}\right). \tag{11.4.15}$$

Hence the contribution of $V(u,v)$ to the brightness distribution get scaled by $(\nu/\nu_o)^2$ and the co-ordinates gets scaled by (ν/ν_o). The effect of the scaling of the co-ordinates, assuming a delta function for the *Dirty Beam*, is to smear a point source at position (l,m) into a line of length $(\Delta\nu/\nu_o)\sqrt{l^2+m^2}$ in the radial direction. This will get convolved with the *Dirty Beam* and the total effect can be found by integrating the brightness distribution over the bandwidth as given in Eq. 11.4.15

$$I^d(l,m) = \left[\frac{\int_0^\infty |H_{RF}(\nu)|^2 \left(\frac{\nu}{\nu_o}\right)^2 I\left(\frac{l\nu}{\nu_0}, \frac{m\nu}{\nu_0}\right) d\nu}{\int_0^\infty |H_{RF}(\nu)|^2 d\nu}\right] * DB_o(l,m), \tag{11.4.16}$$

where $H_{RF}(\nu)$ is the band-shape function of the RF band and DB_o is the *Dirty Beam* at frequency ν_o. If one represents the synthesized beam as a gaussian function of standard deviation $\sigma_b = \theta_b/\sqrt{8ln2}$ and the bandpass represented by a rectangular function of width $\Delta\nu$, the fractional reduction in the strength of a source located at a radial distance $r = \sqrt{l^2+m^2}$ is given by

$$R_b = 1.064\frac{\theta_b\nu_o}{r\Delta\nu} erf\left(0.833\frac{r\Delta\nu}{\theta_b\nu_o}\right). \tag{11.4.17}$$

Eq. 11.4.16 is equivalent to averaging large number of maps made from monochromatic visibilities at ν. Since each of such maps would scale by a different factor, the source away from the center would move along the radial line from one map to another, producing the radial smearing convolved with the *Dirty Beam*. Since the source away from the center is elongated radially, its side-lobes (because of the *Dirty Beam*) will also be elongated in the radial direction. As a result the side-lobes of distant sources will be elongated at the origin but not towards 90^o angle from the vector joining the source and the origin.

The effect of bandwidth smearing can be reduced if the RF band is split into frequency channels with smaller channel widths. This effectively reduces the $\Delta\nu$ as seen by the mapping procedure and while gridding the visibilities then, the u and v can be computed separately for each channel and assigned to the correct uv-cell. The FX correlator used in GMRT provides up to 128 frequency channels over the bandwidth of observation.

11.5 Time Average Smearing

As discussed before, the u and v co-ordinates of an antenna are a function of time and continuously change as earth rotates generating the uv-coverage. To improve the signal-to-noise ratio as well as reduce the data volume, the visibility function $V(u,v)$ is recorded after finite integration in time (typically 10-20s for imaging projects) and the average value of the real and imaginary parts of V are used for average values of u and v over the integration time. Effectively then, the assigned values of u and v for each visibility point is evaluated for a time which is wrong from the correct (instantaneous) time by a maximum of $\tau/2$ where τ is the integration time.

In the map domain, the resulting effect can be visualized by treating the resulting map from the time average visibilities as the average for a number of maps made from the

instantaneous (un-averaged) visibilities. The baseline vectors in the uv-domain follow the loci of the uv-tracks (which are parabolic tracks) and rotate at an angular velocity equal to the that of earth, ω_e. Since a rotation of one domain results into a rotation by an equal amount in the conjugate domain in a Fourier transform relation, the effect in the map domain is that the instantaneous maps also are rotated with respect to each other, at the rate of ω_e. Hence, a point source located at (l, m) away from the center of the map would get smeared in the azimuthal direction. This effect is same as the smearing effect due to finite bandwidth of observations, but in an orthogonal direction.

11.6 Zero-spacing Problem

Since visibility and the brightness distribution are related via a Fourier transform, $V(0,0)$ measures the total flux from the sky. However, since the difference between the antenna positions is always finite, $V(0,0)$ is never measured by an interferometer. For a point source, it is easy to estimate this value by extrapolation from the smallest u and v for which a measurement exist, since V as a function of baseline length is constant. However for an extended source, this value remains unknown and extrapolation is difficult.

For the purpose of understanding the effect of missing zero-spacings, we can multiply the visibility in Eq. 11.3.6 by a rectangular function which is 0 around $(u, v) = (0, 0)$ and 1 elsewhere. In the map domain then, the *Dirty Map* gets convolved with the Fourier transform of this function, which has a central negative lobe. As a result, extended sources will appear to be surrounded by negative brightness in the map which cannot be removed by any processing. This can only be removed by either estimating the zero-spacing flux while restoring I from I^d or V, or by supplying the zero-spacing flux as an external input to the mapping/deconvolution programs. The Maximum Entropy class of image restoration algorithms attempt to estimate the zero-spacing flux, while the CLEAN class of image restoration algorithms needs to be supplied this number externally. Both these will be discussed in the later lectures.

11.7 Further Reading

1. Interferometry and Synthesis in Radio Astronomy;Thompson, A.Richard, Moran, James M., Swenson Jr., George W.; Wiley-Interscience Publication, 1986.

2. Synthesis Imaging In Radio Astronomy; Eds. Perley, Richard A.,Schwab, Frederic R., and Bridle, Alan H.; ASP Conference Series, Vol 6.

3. High Fidelity Deconvolution of Moderately Resolved Sources; Briggs Daniel; Ph.D. Thesis, 1995, The New Maxico Institute of Mining and Technology, Socorro, New Mexico, USA.

Chapter 12

Deconvolution in sythesis imaging–an introduction

Rajaram Nityananda

12.1 Preliminaries

These lectures describe the two main tools used for deconvolution in the context of radio aperture synthesis. The focus is on the basic issues, while other lectures at this school will deal with aspects closer to the actual practice of deconvolution. The practice is dominated by the descendants of a deceptively simple-looking , beautiful idea proposed by J. Högbom (A&A Suppl. 15 417 1974), which goes by the name of CLEAN. About the same time, another, rather different and perhaps less intuitive idea due to the physicist E.T. Jaynes was proposed by J.G. Ables (A&A Suppl 15 383 1974) for use in astronomy. This goes by the name of the Maximum Entropy Method, MEM for short. MEM took a long time to be accepted as a practical tool and even today is probably viewed as an exotic alternative to CLEAN. We will see, however, that there are situations in which it is likely to do better, and even be computationally faster. The goal of these lectures is to give enough background and motivation for new entrants to appreciate both CLEAN and MEM and go deeper into the literature.

12.2 The Deconvolution Problem

12.2.1 Interferometric Measurements

An array like the GMRT measures the visibility function $V(u,v)$ along baselines which move along tracks in the $u-v$ plane as the earth rotates, For simplicity, let us assume that these measurements have been transferred onto a discrete grid and baselines are measured in units of the wavelength. The sky brightness distribution $I(l,m)$ in the field of view is a function of l,m which are direction cosines of a unit vector to a point on the celestial sphere referred to the u and v axes. The basic relationship between the measured visibility function V and the sky brightness I is a Fourier transform.

$$V(u,v) = \int\int I(l,m) \exp(-2\pi i(lu+mv))\ dl\ dm.$$

This expression also justifies the term "spatial frequency" to describe the pair (u, v), since u and v play the same role as frequency plays in representing time varying signals.

Many things have been left out in this expression, such as the proper units, polarisation, the primary beam response of the individual antennas, the non-coplanarity of the baselines, the finite observing bandwidth, etc. But it is certainly necessary to understand this simplified situation first, and the details needed to achieve greater realism can be put in later.

Aperture synthesis, as originally conceived, involved filling in the $u - v$ plane without any gaps upto some maximum baseline b_{max} which would determine the angular resolution. Once one accepts this resolution limit, and writes down zeros for visibility values outside the measured circle, the Fourier transform can be inverted. One is in the happy situation of having as many equations as unknowns. A point source at the field centre.(which has constant visibility) would be reconstructed as the Fourier transform of a uniformly filled circular disk of diameter $2b_{max}$. This is the famous Airy pattern with its first zero at $1.22/(2b_{max})$. The baseline b is already measured in wavelengths, hence the missing λ in the numerator. But even in this ideal situation, there are some problems. Given an array element of diameter D (in wavelengths again!), the region of sky of interest could even be larger than a circle of angular diameter $2/D$. A Fourier component describing a fringe going through one cycle over this angle corresponds to a baseline of $D/2$. But measuring such a short baseline would put two dishes into collision, and even somewhat larger baselines than D run the risk of one dish shadowing the other. In addition, the really lowest Fourier component corresponds to $(u, v) = (0, 0)$, the total flux in the primary beam. This too is not usually measured in synthesis instruments Thus, there is an inevitable "short and zero spacings problem" even when the rest of the $u - v$ plane is well sampled.

12.2.2 Dirty Map and Dirty Beam

But the real situation is much worse. With the advent of the Very Large Array (VLA), the majestic filling in of the $u - v$ plane with samples spaced at $D/2$ went out of style. If one divides the field of view into pixels of size $1/(2b_{max})$, then the total number of such pixels (resolution elements) would be significantly larger than the number of baselines actually measured in most cases. This is clearly seen in plots of $u - v$ coverage which have conspicuous holes in them. The inverse Fourier transform of the measured visibility is now hardly the true map because of the missing data. But it still has a name - the "dirty map" I^D. We define a sampling cum weighting function $W(u, v)$ which is zero where there are no measurements and in the simplest case (called uniform weighting) is just unity wherever there are measurements. So we can get our limited visibility coverage by taking the true visibilities and multiplying by $W(u, v)$. This multiplication becomes a convolution in the sky domain. The "true" map with full visibility coverage is therefore convolved by the inverse Fourier transform of W which goes by the name of the "dirty beam" $B^D(l, m)$.

$$I^D(l, m) = \int\int I(l', m')B^D(l - l', m - m')\ dl'\ dm'$$

where

$$B^D(l, m) \propto \sum W(u, v) \exp(+2\pi i(lu + mv)).$$

For a patchy $u - v$ coverage, which is typical of many synthesis observations, B^D has strong sidelobes and other undesirable features. This makes the dirty map difficult to interpret. What one sees in one pixel has contributions from the sky brightness in neighbouring and even not so neighbouring pixels. For the case of $W = 1$ within a disk of

radius b_{max} we get an Airy pattern as mentioned earlier. This is not such a dirty beam after all, and could be cleaned up further by making the weighting non-uniform, i.e. tapering the function W down to zero near the edge $|(u,v)| = b_{max}$. For example, if this weighting is approximated by a Gaussian, then the sky gets convolved by its transform, another Gaussian. This dirty map is now related to the true one in a reasonable way. But, as Ables remarked, should one go to enormous expense to build and measure the longest baseline and then multiply it by zero?

12.2.3 The Need for Deconvolution

Clearly, there has to be a better way than just reweighting the data to make the dirty beam look better, (and fatter, incidentally, since one is suppressing high spatial frequencies), But this better way has to play the dangerous game of interpolating (for short spacings and for gaps in the $u-v$ plane) and extrapolating (for values beyond the largest baseline) the visibility function which was actually measured. The standard terminology is that the imaging problem is "underdetermined" or "ill-posed" or "ill-conditioned". It has fewer equations than unknowns. However respectable we try to make it sound by this terminology, we are no better than someone solving $x + y = 1$ for both x and y!. Clearly, some additional criterion which selects one (or a few) solutions out of the infinite number possible has to be used. The standard terminology for this criterion is "a priori information". The term "a priori" was used by the philosopher Kant to describe things in the mind that did not seem to need sensory input, and is hence particularly appropriate here.

One general statement can be made. If one finds more than one solution to a given deconvolution problem fitting a given data set, then subtracting any two solutions should give a function whose visibility has to vanish everywhere on the data set. Such a brightness distribution, which contains only unmeasured spatial frequencies, is appropriately called an "invisible distribution". Our extra- /inter- polation problem consists in finding the right invisible distribution to add to the visible one!

One constraint often mentioned is the positivity of the brightness of each pixel. To see how powerful this can be, take a sky with just one point source at the field centre. The total flux and two visibilities on baselines $(D/2, 0), (0, D/2)$ suffice to pin down the map completely. The only possible value for all the remaining visibilities is equal to these numbers, which are themselves equal. One cannot add any invisible distribution to this because it is bound to go negative somewhere in the vast empty spaces around our source. But this is an extreme case. The power of positivity diminishes as the field gets filled with emisssion.

Another interesting case is when the emission is known to be confined to a window in the map plane. Define a function $w(l,m) = 1$ inside the window and zero outside. Let $\tilde{w}(u,v)$ be its Fourier transform. Multiplying the map by w makes no difference. In Fourier space, this condition is quite non-trivial, viz $V(u,v) = V(u,v) * \tilde{w}(u,v)$. Notice how the convolution on the right transfers information from measured to unmeasured parts of the $u-v$ plane, and couples them.

12.3 CLEAN

12.3.1 The Högbom Algorithm

Consider a sky containing only isolated point sources. In the dirty map, each appears as a copy of the dirty beam, centred on the source position and scaled by its strength. However, the maxima in the map do not strictly correspond to the source positions, because

each maximum is corrupted by the sidelobes of the others, which could shift it and alter its strength. The least corrupted, and most corrupting, source is the strongest. Why not take the largest local maximum of the dirty map as a good indicator of its location and strength? And why not subtract a dirty beam of the appropriate strength to remove to a great extent the bad effects of this strongest source on the others? The new maximum after the subtraction now has a similar role. At every stage, one writes down the coordinates and strengths of the point sources one is postulating to explain the dirty map. If all goes well, then at some stage nothing (or rather just the inevitable instrumental noise) would be left behind. We would have a collection of point sources, the so called CLEAN components, which when convolved with the dirty beam give the dirty map.

One could exhibit this collection of point sources as the solution to the deconvolution problem, but this would be arrogant, since one has only finite resolution. As a final gesture of modesty, one replaces each point source by (say) a gaussian, a so called "CLEAN" beam, and asserts that the sky brightness, convolved with this beam, has been found.

This strategy, which seems so reasonable today, was a real breakthrough in 1974 when proposed by J. Högbom. Suddenly, one did not have to live with sidelobes caused by incomplete $u - v$ coverage. In fact, the planning for new telescopes like the VLA must have taken this into account- one was no longer afraid of holes.

12.3.2 The Behaviour of CLEAN

With hindsight, one can say that the initial successes were also due to the simplicity of the sources mapped. It is now clear that one should not be applying this method to an extended source which covered several times the resolution limit (the width of the central peak of the dirty beam). Such a source could have a broad, gentle maximum in the dirty map, and subtracting a narrow dirty beam at this point would generate images of the sidelobes with the opposite sign. This would generate new maxima where new CLEAN components would be placed by the algorithm, and things could go unstable. One precaution which certainly helps is the "gain factor" (actually a loss factor since it is less than one). After finding a maximum, one does not subtract the full value but a fraction g typically 0.2 or less. In simple cases, this would just make the algorithm slower but not change the solution. But this step actually helps when sources are more complex. One is being conservative in not fully believing the sources found initially. This gives the algorithm a chance to change its mind and look for sources elsewhere. If this sounds like a description of animal behaviour, the impression being conveyed is correct. Our understanding of CLEAN is largely a series of empirical observations and thumb rules, with common sense rationalisations after the fact, but no real mathematical theory. One exception is the work of Schwarz (A&A 65 345 1978) which interpreted each CLEAN subtraction as a least squares fit of the current dirty map to a single point source. This is interesting but not enough. CLEAN carries out this subtraction sequentially, and that too with a gain factor. In principle, each value of the gain factor could lead to a different solution, i.e a different collection of CLEAN components, in the realistic case when the number of $u - v$ points is less than the number of resolution elements in the map. So what are we to make of the practical successes of CLEAN? Simply that in those cases, the patch of the sky being imaged had a large enough empty portion that the real number of CLEAN components neeeded was smaller than the number of data points available in the $u-v$ plane. Under such conditions, one could believe that the solution is unique. Current implementations of CLEAN allow the user to define "windows" in the map so that one does not look for CLEAN components outside them. But when a large portion of the field of view has some nonzero brightness, there are indeed problems with CLEAN. The maps show spurious stripes whose separation is related to unmeasured spatial frequencies

(that's how one deduces they are spurious). One should think of this as a wrong choice of invisible distribution which CLEAN has made. Various modifications of CLEAN have been devised to cope with this, but the fairest conclusion is that the algorithm was never meant for extended structure. Given that it began with isolated point sources it has done remarkably well in other circumstances.

12.3.3 Beyond CLEAN

Apart from the difficulties with extended sources, CLEAN as described above is an inherently slow procedure. If N is the number of pixels, subtracting a single source needs of the order of N operations. This seems a waste when this subtraction is a provisional, intermediate step anyway! B.G. Clark had the insight of devising a faster version, which operates with a truncated dirty beam, but only on those maxima in the map strong enough that the far, weak sidelobes make little difference. Once these sources have been identified by this rough CLEAN (called a "minor cycle"), they are subtracted together from the full map using an fast fourier transform (FFT) for the convolution, which takes only $N \log N$ operations. This is called the "major cycle". The new residual map now has a new definition of "strong" and the minor cycle is repeated.

A more daring variant, due to Steer, Dewdney, and Ito, (hence SDI CLEAN) carries out the minor cycle by simply identifying high enough maxima, without even using CLEAN, which is kept for the major cycle. Other efforts to cope with extended sources go under the name of "multiresolution CLEAN". One could start with the inner part of the $u-v$ plane and do a CLEAN with the appropriate, broader dirty beam. The large scale structure thus subtracted will hopefully now not spoil the next stage of CLEAN at a higher resolution, i.e using more of the $u - v$ plane.

12.4 Maximum Entropy

12.4.1 Bayesian Statistical Inference

This method, or class of methods, is easy to describe in the framework of an approach to statistical inference (i.e all of experimental science?) which is more than two hundred years old, dating from 1763! Bayes Theorem about conditional probabilities states that

$$P(A|B)P(B) = P(B|A)P(A) = P(A, B).$$

As a theorem, it is an easy consequence of the definitions of joint probabilities (denoted by $P(A, B)$), conditional probabilities (denoted by $P(A|B)$) and marginal or unconditional probabilities (denoted by $P(A)$). In words, one could say that the fraction of trials A *and* B *both* happen ($P(A, B)$) is the product of (i) the fraction of trials in which A happens ($P(A)$) irrespective of B, and (ii) the further fraction of A-occurences which are also B-occurences ($P(B|A)$). The other form for $P(A|B)$ follows by interchanging the roles of A and B.

The theorem acquires its application to statistical inference when we think of A as a hypothesis which is being tested by measuring some data B. In real life, with noisy and incomplete data, we never have the luxury of measuring A directly, but only something depending on it in a nonunique fashion. If we understand this dependence, i.e understand our experiment, we know $P(B|A)$. If only, (and this is a big IF!), someone gave us $P(A)$, then we would be able to compute the dependence of $P(A|B)$ on A from Bayes theorem.

$$P(A|B) = P(B|A)P(A)/P(B).$$

Going from $P(B|A)$ to $P(A|B)$ may not seem to be a big step for a man, but it is a giant step for mankind. It now tells us the probability of different hypotheses A being true based on the given data B. Remember, this is the real world. More than one hypothesis is consistent with a given set of data, so the best we can do is narrow down the possibilities. (If "hypothesis" seems too abstract, think of it as a set of numbers which occur as parameters in a given model of the real world)

12.4.2 MEM Images

Descending now from the sublime to aperture synthesis, think of A as the true map and B as the dirty map, or equivalently its Fourier transform, the set of measured visiblilities. We usually want a single map, not a probability distribution of A. So we need the further step of maximising $P(A|B)$ with respect to A. All this is possible if $P(A)$ is available for a given true map $I(l,m)$. One choice, advocated by Gull and Daniell in 1978, was to take

$$\log P(\{I(l,m)\}) \propto -\int\int I(l,m)\ln I(l,m)\ dl\ dm.$$

The curly brackets around I on the left side are meant to remind us that the entropy is a single number computed from the entire information about the brightness, i.e the whole set of pixel values. Physicists will note that this expression seems inspired by Boltzmann's formula for entropy in statistical mechanics, and communication engineers will see the influence of Shannon's concept of information. It was E.T. Jaynes writing in the Physical Review of 1957 who saw a vision of a unified scheme into which physics, communication theory, and statistical inference would fall (with the last being the most fundamental!). In any case, the term "entropy" for the logarithm of the prior distribution of pixel values has stuck. One can see that if the only data given was the total flux, then the entropy as defined above is a maximum when the flux is distributed uniformly over the pixels. This is for the same reason that the Boltzmann entropy is maximised when a gas fills a container uniformly. This is the basis for the oft-heard remark that MEM produces the flattest or most featureless map consistent with the data - a statement we will see requires some qualification. But if one does not want this feature, a modified entropy function which is the integral over the map of $-I\ln(I/I^d)$ is defined. $I^d(l,m)$ is called a "default image". One can now check that if only total flux is given the entropy is a maximum for $I \propto I^d$.

The selection of a prior is, in my view, the weakest part of Bayesian inference, so we will sidestep the debate on the correct choice. Rather, let us view the situation as an opportunity, a license to explore the consequences of different priors on the "true" maps which emerge. This is easily done by simulation – take a plausible map, Fourier transform, sample with a function W so that some information is now missing, and use your favourite prior and maximise "entropy" to get a candidate for the true map. It is this kind of study which was responsible for the great initial interest in MEM. Briefly, what MEM seemed to do in simple cases was to eliminate the sidelobes and even resolve pairs of peaks which overlapped in the true map, i.e it was sometimes "better" than the original! This last feature is called superresolution, and we will not discuss this in the same spirit of modesty that prompted us to use a CLEAN beam. Unlike CLEAN, MEM did not seem to have a serious problem with extended structure, unless it had a sharp edge (like the image of a planet). In this last case, it was found that MEM actually enhanced the ripples near the edge which were sitting at high brightness levels; though it controlled the ripples which were close to zero intensity. This is perhaps not surprising if one looks at the graph of the function $= I\ln I$. There is much more to be gained by removing ripples

near $I = 0$ than at higher values of I, since the derivative of the function is higher near $I = 0$.

Fortunately, these empirical studies of the MEM can be backed up by an analytical/graphical argument due to Ramesh Narayan, which is outlined below. The full consequences of this viewpoint were developed in a review article (Annual review of Astronomy and Astrophysics 24 127 1986), so they will not be elaborated here, but the basic reasoning is simple and short enough. Take the expression for the entropy, and differentiate it with respect to the free parameters at our disposal, namely the *un*measured visibilities, and set to zero for maximisation. The derivative of the entropy taken with respect to a visibility $V(u', v')$ is denoted by $M(u', v')$. The understanding is that u', v' have *not* been measured. The condition for a maximum is

$$M(u', v') = \int\int (-1 - \ln(I(l, m)) \exp(+2\pi i(lu' + mv')\ dl\ dm = 0.$$

This can be interpreted as follows. The *logarithm* of the brightness is like a dirty map, i.e it has no power at unmeasured baselines, and hence has sidelobes etc. But the brightness I itself is the exponential of this "band limited function" (i.e one with limited spatial frequency content). Note first of all that the positivity constraint is nicely implemented– exponentials are positive. Since the exponential varies rather slowly at small values of I, the ripples in the "baseline" region between the peaks are suppressed. Conversely, the peaks are sharpened by the steep rise of the exponential function at larger values of I. One could even take the extreme point of view that the MEM stands unmasked as a model fitting procedure with sufficient flexibility to handle the cases usually encountered. Högbom and Subrahmanya independently emphasised very early that the entropy is just a penalty function which encourages desirable behaviour and punishes bad features in the map (IAU Colloq. 49, 1978). Subrahmanya's early work on the deconvolution of lunar occultation records at Ooty (TIFR thesis, 1977) was indeed based on such penalties.

More properties of the MEM solution are given in the references cited earlier. But one can immediately see that taking the exponential of a function with only a limited range of spatial frequencies (those present in the dirty beam) is going to generate all spatial frequencies, i.e., one is extrapolating and interpolating in the $u - v$ plane. It is also clear that the fitting is a nonlinear operation because of the exponential. Adding two data sets and obtaining the MEM solution will not give the same answer as finding the MEM solution for each separately and adding later! A little thought shows that this is equally true of CLEAN.

If one has a default image I^d in the definition of the entropy function, then the same algebra shows that I/I^d is the exponential of a band-limited function. This could be desirable. For example, while imaging a planet, if the sharp edge is put into I^d, then the MEM does not have to do so much work in generating new spatial frequencies in the ratio I/I^d. The spirit is similar to using a window to help CLEAN find sources in the right place.

12.4.3 Noise and Residuals

The discussion so far has made no reference to noise in the interferometric measurements. But this can readily be accomodated in the Bayesian framework. One now treats the measurements not as constraints but as having a Gaussian distribution around the "true" value which the real sky would Fourier transform to. Thus the first factor $P(B|A)$ on the right hand side of Bayes theorem would now read

$$P(B|A) = \prod \exp(-(\int\int I(l, m) exp(-2\pi i(lu + mv))\ dl\ dm - V_m(u, v)|^2/2\sigma^2_{u,v}.$$

The product is over measured values of u, v. A nice feature of the gaussian distribution is that when we take its logarithm, we get the sum of the squares of the residuals between the model predictions (the integral above) and the measurements $V_m(u, v)$ – also known as "chi-squared" or χ^2. The logarithm of the prior is of course the entropy factor. So, in practice, we end up maximising a linear combination of the entropy and χ^2, the latter with a negative coefficient. This is exactly what one would have done, using the method of Lagrange multipliers, if we were maximising entropy subject to the constraint that the residuals should have the right size, predicted by our knowledge of the noise.

All is not well with this recipe for handling the noise. The discrepancy between the measured data and the model predictions can be thought of as a residual vector in a multidimensional data space. We have forced the length to be right, but what about the direction? True residuals should be random, i.e the residual vector should be uniformly distributed on the sphere of constant χ^2. But since we are maximising entropy on this sphere, there will be a bias towards that direction which points along the gradient of the entropy function. This shows in the maps as a systematic deviation tending to lower the peaks and raise the "baseline" i.e the parts of the image near zero I. To lowest order, this can be rectified by adding back the residual vector found by the algorithm. This does not take care of the invisible distribution which the MEM has produced from the residuals, but is the best we can do. Even in the practice of CLEAN, residuals are added back for similar reasons.

The term "bias" is used by statisticians to describe the following phenomenon. We estimate some quantity, and even after taking a large number of trials its average is not the noise-free value. The noise has got "rectified" by the non-linear algorithm and shows itself as a systematic error. There are suggestions for controlling this bias by imposing the right distribution and spatial correlations of residuals. These are likely to be algorithmically complex but deserve exploration. They could still leave one with some subtle bias since one cannot really solve for noise. But to a follower of Bayes, bias is not necesarily a bad thing. What is a prior but an expression of prejudice? Perhaps the only way to avoid bias is to stop with publishing a list of the measured visibility values with their errors. Perhaps the only truly open mind is an empty mind!

12.5 Further Reading

1. R.A. Perley, F.R. Schwab, & A.H. Bridle, eds., 'Synthesis Imaging in Radio Astronomy', ASP Conf. Series, vol. 6.

2. Thompson, R.A., Moran, J.M. & Swenson, G.W. Jr., 'Interferometry & Synthesis in Radio Astronomy', Wiley Interscience, 1986.

3. Steer, D.G., Dewdney, P.E. & Ito, M.R., "Enhancements to the deconvolution algorithm 'CLEAN"', 1984,A&A,137,159.

Chapter 13

Spectral Line Observations

K. S. Dwarakanath

This chapter is intended as an introduction to spectral line observations at radio wavelengths. While an attempt will be made to put together most of the relevant details, it is not intended to be an exhaustive guide to spectral line observations but instead focuses more on the basics of spectral line observations, keeping in mind synthesis arrays like the Giant Meterwave Radio Telescope (GMRT).

13.1 Spectral Lines

Spectral lines originate under a variety of circumstances in Astronomy. The most ubiquitous element in the Universe, the Hydrogen atom, gives rise to the 21-cm-line ($\nu \sim$ 1420.405 MHz) due to a transition between the hyperfine levels of its ground state. If the Hydrogen atom is ionized, subsequent recombinations of electrons and protons lead to a series of recombination lines of the Hydrogen atom. It is easy to see that such transitions between higher Rydberg levels give rise to spectral lines at radio wavelengths. Transitions around Rydberg levels of 280, for e.g., give rise to recombination lines at $\nu \sim 300$ MHz. In cold (kinetic temperature $\sim$ 100 K), and dense ($\sim$ 1000 cm^{-3}) environments Hydrogen atoms form molecules. The CO molecule which has been used as a tracer of molecular Hydrogen has a rotational transition at $\nu \sim 115$ GHz. These are a few illustrative examples.

The widths of spectral lines arise due to different mechanisms. One such is the Doppler effect. The particles in a gas have random motions corresponding to the kinetic temperature of the gas. The observed frequency of the line is thus different from the rest frequency emitted by the particles. In a collision-dominated system, the number density of particles as a function of velocity is expected to be a Maxwellian distribution. The width of this distribution will result in a corresponding broadening of the observed spectral line due to Doppler Effect. This width, arising due to the temperature of the gas, is called thermal broadening. In addition to the thermal motion of the particles, there can also be turbulent velocities associated with macroscopic gas motions. These motions are often accounted for by an effective Doppler width, which includes both thermal and turbulent broadening, assuming a gaussian distribution for the turbulent velocities also. Another mechanism which can contribute to the line width is pressure broadening. This arises due to collisions and is particularly relevant in high density environments and/or for lines arising through transitions between high Rydberg levels. In addition, there is always a natural width to the spectral line imposed by the uncertainty principle, but it is

almost always overwhelmed by that due to the mechanisms mentioned earlier.

An observed spectral feature can be much wider than that expected on the basis of the above mentioned mechanisms. This is usually due to systematic motion of the gas responsible for the spectral feature like, for e.g., rotation of a gas cloud, expansion of a gas cloud, differential rotation of a galaxy, etc..

13.2 Rest Frequency and Observing Frequency

The rest frequency of a spectral line of interest can be calculated if it is not already tabulated. The apparent frequency (or, the observing frequency), however, needs to be calculated for each source since it depends on the relative velocity between the source and the observer. The observed frequency (ν_o) of a given transition is related to the rest frequency of the line (ν_l) and the radial velocity of the source w.r.t the observer (v_r) as ($\nu_l - \nu_o$)=$\nu_o v_r$/c, where, c is the velocity of light. This relation is valid for $v_r \ll$ c, and $\theta \ll \pi/2$, where θ is the angle between the velocity vector and the radiation wave vector. The radial velocity is positive if the motion is away from the observer and the observed frequency is smaller than the rest frequency of the line. In this situation, the line is redshifted. If the velocity (v_r) is known, the observing frequency can be calculated. While dealing with extragalactic systems, one quotes the redshift rather than the radial velocity. The redshift (z) is related to the rest and observed frequencies as z = ($\nu_l - \nu_o$)/ν_o and approximates to v_r/c for $v_r \ll$ c.

It is more useful, and common to define velocities w.r.t. the 'local standard of rest' than w.r.t. an arbitrary frame of reference. This transformation takes into account the radial velocity corrections due to the rotation of the earth about its own axis, the revolution of the earth around the Sun, and the motion of the Sun w.r.t. the local group of stars. The magnitudes of these corrections are within $\sim$ 1 km s^{-1}, 30 km s^{-1}, and 20 km s^{-1} respectively. The actual value of the total correction depends on the equatorial coordinates of the source, the ecliptic coordinates of the source, the longitude of the Sun, the hour angle of the source, and the geocentric latitude of the observer.

In principle, the apparent frequency of a spectral line from a source is always changing due to the change in the radial velocity between the source and the observer. In a given observing session during a day the source can be observed from rise to set. During this period the radial component of the velocity between the source and the earth due to the rotation of the earth can (in an extreme case) change from -0.465 to +0.465 km s^{-1}. Consider observing a narrow spectral line (width $\sim$ 0.5 km s^{-1}) from this source using a spectral resolution $\sim$ 0.1 km s^{-1}. If no extra precautions are taken, the peak of the spectral line will appear to slowly drift across the channels during the course of the day. This drift, if not accounted for, will decrease the signal-to-noise ratio of the line, and increase its observed width in the time-averaged spectrum. Depending on the circumstances, this can completely wash out the spectral line. In order to overcome this, the continuous change in the apparent frequency is to be corrected for during an observing session so that the spectral line does not drift across frequency but stays in the same channels. This process of correction is known as Doppler Tracking. I would like to emphasize that this is important if one is observing narrow lines with high spectral resolution and that there is a significant change in the sight-line component of the earth's rotation during the observing session.

13.3 Setting the Observing Frequency and the Bandwidth

Once the apparent frequency ν_o of the transition of interest is known, the Local Oscillator (LO) frequencies can be tuned to select this frequency for observations. In general, there can be more than one LO that need to be tuned. Consider the situation at the GMRT. The First LO (ν_{ILO}) can be chosen such that ν_{ILO} = $\nu_o \pm \nu_{IF}$, where, ν_{IF} is the Intermediate Frequency (IF). The First LO can be tuned in steps of 5 MHz. The IF is 70 MHz. The IF bandwidth ($\delta\nu_{IF}$) can be chosen from one of 6, 16, and 32 MHz. Thus, the output of the first mixer will be over a frequency range of $\nu_{IF} \pm \delta\nu_{IF}/2$. The baseband LO ($\nu_{BBLO}$) can be tuned in the range of 50 to 90 MHz in steps of 100 Hz to bring the IF down to the baseband. The bandwidth of the baseband filter ($\delta\nu_{BB}$) can be chosen from 62.5 KHz to 16 MHz in steps of 2. The bands from $-\delta\nu_{BB}/2$ to 0, and from 0 to $\delta\nu_{BB}/2$, which are the lower, and the upper side bands respectively, will be processed separately. The FX Correlator at the GMRT will produce 128 spectral channels (0 – 127) covering each of these bands. The 0th channel corresponds to a frequency of $\nu_o + \nu_{BBLO} - \nu_{IF}$ and the frequency increases with channel number in the USB spectrum and decreases with channel number in the LSB spectrum.

While setting the LO frequencies one needs to make sure that (a) the desired LO frequency is in the allowed range and that the oscillator is 'locked' to a stable reference, and, (b) that the required power output is available from the oscillator. The choice of the baseband filter bandwidth depends on the velocity resolution and the velocity coverage required for a given observation. In addition, it is preferable to have as many line-free channels in the band as there are channels with the line in order to be able to obtain a good estimate of the observed baseline (or reference spectrum). One would also like to center the spectral feature within the observed band so that line-free channels on either side can be used to estimate the baseline. The velocity resolution should be at least a factor of two better than the full width at half maximum of the narrowest feature one is expecting to detect.

At present, the FX Correlator at the GMRT produces 128 channels per side band for each of the two polarizations. The two polarizations are identified as the 130 MHz and the 175 MHz channels. In principle it should be possible to drop one of the polarizations to obtain 256 channels for one polarization. This will improve the spectral resolution by a factor of 2 keeping the velocity coverage (the bandwidth) the same. This can be very useful in observing narrow lines over a wider range of velocities.

13.4 Calibration

The observed spectrum has to be corrected for the telescope response as a function of frequency across the band to obtain an estimate of the true spectrum. The telescope response is in general complex with both amplitude and phase variations across the observing band. This overall response across the band can be split into two components : (1) an overall gain (amplitude and phase) of the telescope for a reference radio frequency (RF) within the observing band, and (2) a variation of this gain across channels (the bandshape). The telescope response is thus a combination of RF gain calibration and IF bandshape calibration. This way of looking at the telescope calibration is useful since the requirements for determining these two parts of the telescope response can be different. For e.g., the IF bandshape variation is expected to be slower in time than the RF gain variation and hence need to be estimated less often. The spectral scale for the IF bandshape is however narrower compared to that of the RF gain.

13.4.1 Gain Calibration

This is usually achieved by observing a bright, unresolved source which is called a calibrator. In the case of a synthesis array like, for e.g., the GMRT, the gain calibration amounts to estimating the gains of the individual antennas in the array. The gains of any given pair of antennas reflect in the visibility (or the cross correlation) of the calibrator measured by them. In an array with N antennas, there are N(N-1)/2 independent estimates of the calibrator (an unresolved bright source) visibility at any give instant of time. However, there are only 2N unknowns, viz., N amplitudes and N phases of the N antennas. Hence, the measured visibilities can be used in a set of simultaneous equations to solve for these 2N unknowns. In practice, a calibrator close (in direction) to the source is observed for a suitable length of time using the same setup as that for the spectral line observations towards the source. A suitable number of spectral channels are averaged to improve the signal-to-noise ratio on the calibrator which is then used to estimate the gains of the antennas. Apart from the instrumental part, the gains include atmospheric offsets/contributions also. The proximity of the calibrator to the source ensures that the atmospheric offsets/contributions are similar in both observations and hence get corrected for through the 'calibration' process.

How often does one do the calibration depends on various factors, like for e.g., the observing frequency, the length of the baseline involved, the telescope characteristics, the time scale for variations in the atmospheric offsets/contributions, etc.. The frequency of calibration can vary from once in $\sim$ 10 minutes to once in an hour depending on these factors.

13.4.2 Bandshape Calibration

In this case too, a bright, unresolved source is used as a calibrator but the nearness requirement (as in the gain calibration) is not essential. On the other hand, the calibrator should not have any spectral features in the band of interest. The measured visibilities from the calibrator across the band of interest can once again (like in the earlier gain calibration) be used to estimate the antenna bandshapes. The observed spectrum from the source is divided by the bandshapes to obtain the true spectrum. The bandshape should have a signal-to-noise ratio (snr) significantly greater than that of the observed spectrum so that the snr in the corrected spectrum is not degraded. For e.g., if the bandshape and the observed spectrum have equal snr, then the corrected spectrum will have an snr which is square root of 2 worse (assuming gaussian statistics of noise). Ideally, one wouldn't want the corrected spectrum to degrade in its snr by more than $\sim$ 10%. This can be used as a criterion to judge if a given calibrator is bright enough and to decide the amount of integration time required for the source and for the calibrator.

There are two methods of bandshape calibration.

(1) Position Switching : In this method, the telescope cycles through the source and a bandshape calibrator but observing both at the same frequency and bandwidth. Depending on the accuracy to which the corrected bandshape is required, and the stability of the receiver, the frequency of bandshape calibration can vary from once in $\sim$20 minutes to once in a few hours.

(2) Frequency Switching : There are situations when position switching is not a suitable scheme to do the bandshape calibration. This can happen due to (at least) two reasons : (a) the band of interest covers the Galactic HI. In this situation, all calibrators will also have some spectral feature within this band due to the ubiquitous presence of Galactic HI. No calibrator is suitable for bandshape calibration. (b) The band is outside the Galactic HI but the source of interest is a bright unresolved source. In this case one

might end up observing any other calibrator much longer ($\sim$ 10 times) than the source in order to achieve the desired signal-to-noise ratio on the bandshape. In either of these situations position switching is not desirable. An alternative scheme is employed.

If a spectral feature covers a bandwidth of $\delta\nu$ centered at ν, quite often it is possible to find line-free regions in the bands centered at $\nu \pm \delta\nu$. The bandshapes at these adjacent frequencies can be used to calibrate the observed spectrum. This works well because the bandshape is largely decided by the narrowest band in the signal path through the telescope. This is usually decided by the baseband filter. The bandwidth of this filter is selected to be the same while observing at frequencies $\nu - \delta\nu$, ν, and $\nu + \delta\nu$. It is important to keep in mind that frequency switching works as long as $\delta\nu$ is small compared with the bandwidth of the front-end devices, and feeds. This is usually the case. For e.g., at the GMRT, the 21-cm feeds have a wide-band response, over 500 MHz. This is divided into 4 sub-bands each of 120 MHz width. If the amount by which the frequency is switched is small compared to 120 MHz this technique should work quite satisfactorily. A typical frequency switching observation would thus have an "off1", "on", and an "off2" setting. The "on" setting centers the band at the spectral feature of interest (at ν) with a bandwidth of $\delta\nu$ while the "off1" and "off2" settings will be centered at $\nu - \delta\nu$ and $\nu + \delta\nu$ respectively. The three settings will be cycled through with appropriate integration times. The average of the "off1" and "off2" bandshapes can be the effective bandshape to calibrate the "on" spectrum. In this situation, equal amounts of time are spent "off" the line and "on" the line to achieve the optimum signal-to-noise ratio in the final spectrum. However, the switching frequency itself will depend on the receiver stability, and the flatness of the corrected bandshape required. This could vary from once in $\sim$20 minutes to once in a few hours.

There are situations when one might to do both frequency and position switching. If one is observing Galactic HI absorption towards a weak continuum source, it is advantageous to obtain bandshape calibration by observing a brighter continuum source with frequency switching.

13.5 Smoothing

The cross power spectrum is obtained by measuring the correlation of signals from different antennas as a function of time offset between them. A spectrum with a bandwidth $\delta\nu$ and N channels is produced by cross correlating signals sampled at interval of τ with relative time offset in the range –Nτ to (N-1)τ, where τ = 1/(2$\delta\nu$). Because of this truncation in the offset time range amounting to a rectangular window, the resulting spectrum is equivalent to convolving the true spectrum by a Sinc function. Thus, a delta function in frequency (a narrow spectral line, for e.g.) will result in an appropriately shifted $\sin(N\pi\nu/\delta\nu)/(N\pi\nu/\delta\nu)$ pattern, where $\delta\nu$/N is the channel separation. The full width at half maximum of the Sinc function is 1.2$\delta\nu$/N. This is the effective resolution. Any sharp edge in the spectrum will result in an oscillating function of this form. This is called the Gibbs' phenomenon. There are different smoothing functions that bring down this unwanted ringing, but at the cost of spectral resolution. One of the commonly used smoothing functions in radio astronomy is that due to Hanning weighting of the correlation function. This smoothing reduces the first sidelobe from 22% (for the Sinc function) to 2.7%. The effective resolution will be 2$\delta\nu$/N. After such a smoothing, one retains only the alternate channels. For Nyquist sampled data, the Hanning smoothing is achieved by replacing every sample by the sum of one half of its original value and one quarter the original values at the two adjacent positions.

Apart from Hanning smoothing which is required to reduce the ringing, additional

smoothing of the spectra might be desirable. The basic point being that a spectral line of given width will have the best signal-to-noise ratio when observed with a spectral resolution that matches its width. This is the concept of 'matched-filtering' and is particularly important in detection experiments.

13.6 Continuum Subtraction

Quite often spectral line observations include continuum flux density present in the band. The continuum in the band can arise due to a variety of reasons. Ionized Hydrogen regions, for e.g., give rise to the radio recombination lines of Hydrogen due to bound-bound transitions and the radio continuum due to thermal bremsstrahlung. Galaxies can have strong non-thermal radio continuum as well as 21-cm-line emission and/or absorption. In addition, any absorption spectral line experiment involves a bright continuum background source. In these and similar situations, detecting a weak spectral line in the presence of strong continuum contribution can be very difficult. Depending on the complexity of the angular distribution of the continuum flux density and that of the spectral feature this task might almost become impossible.

The basic problem here is one of spectral dynamic range (SDR). The spectral dynamic range is the ratio of the weakest spectral feature that can be detected to the continuum flux density in the band. This is limited by the residual errors which arise due to a variety of reasons like, for e.g., the instrumental variations, the atmospheric gain changes, the deconvolution errors, etc.. Of these, the multiplicative errors limit the SDR depending on the continuum flux density in the band. Thus, if the multiplicative errors are at 1% level, and , if the continuum flux density in the band is 10 Jy, no spectral line detection is possible below 100 mJy. On the other hand, a continuum subtraction (if successful) will lead to a situation where the SDR is decided by the peak spectral line flux density rather than the continuum flux density. Apart from the continuum flux density any other systematics which have a constant value or a linear variation across frequency will be subtracted out in the continuum subtraction procedure. This can lead to improvements in the SDR by several orders of magnitude.

There are several methods for subtracting the continuum flux density from a spectral line data. It is beyond the scope of this lecture to discuss all of these. A brief mention will be made of one of these simpler methods to illustrate some of the principles involved. In this method, which has been called 'visibility-based subtraction', a linear fit to the visibilities as a function of frequency is performed for every sample in time. This best-fit continuum can then be subtracted from the original visibilities. The resulting data can be Fourier transformed to produce continuum-free images. This method works quite well if the continuum emission is spread over a sufficiently small field of view. This limitation can be understood in the following way. Consider a two-element interferometer separated by d. Let each of the elements of the interferometer be pointing towards θ_0 which is also the fringe tracking (phase tracking) center. The phase difference between θ_0, and an angle θ close to this, is $\phi = 2\pi\nu d(sin(\theta) - sin(\theta_0))/c$, where, ν is the observing frequency, and c is the velocity of light. For the present purpose of illustration, assume that θ is in the plane containing the pointing direction (θ_0) and d. The visibilities from a source at θ will have the form $A_\nu cos(\phi)$ and $A_\nu sin(\phi)$, where, A_ν is the amplitude of the source at ν. Writing $\nu = \nu_0 + \delta\nu$, and $\theta = \theta_0 + \delta\theta$, where, ν_0 is the frequency of the center of the band, it can be shown that the frequency-dependent part of the phase is $\phi_\nu = 2\pi\delta\nu dcos(\theta_0)\delta\theta/\nu_0\lambda_0$, where, $c = \nu_0\lambda_0$. It is easy to see that the variation of visibilities as a function of frequency is linear if $\phi \ll 2\pi$. This implies that $\delta\nu\delta\theta/(\nu_0\theta_{syn}) \ll 1$, where, $\theta_{syn} = \lambda_0/d$. Thus, this method of continuum subtraction works if most of the continuum is within $\nu_0/\delta\nu$ synthesized beams

from the phase tracking center.

13.7 Line Profiles

If the line width is greater than the spectral resolution one can discuss the variation of the intensity of the line as a function of frequency. This description, called the line profile, can be denoted by $\phi(\nu)$. If the reason for the line width is thermal broadening or turbulent broadening, the line profile will have a gaussian profile such that $\phi(\nu) \propto e^{-(\nu-\nu_l)^2/(\delta\nu)^2}$, where ν_l is the frequency at the line center and $\delta\nu$ is the rms value of the gaussian. The width of the line refers to the full-width at half-maximum and is equal to ~ 2.35 $\delta\nu$. The observed width of the line ($\delta\nu_o$) and the true width of the line ($\delta\nu_l$) are related by $\delta\nu_o^2 = \delta\nu_l^2 + \delta\nu_r^2$, where, $\delta\nu_r$ is the width of each channel (spectral resolution). This simple relation is strictly true only when the spectral channels have a gaussian response. In addition, this is relevant if the widths of the spectral line and the spectral channel are comparable.

Pressure broadened lines show Voigt profiles. This will have a Doppler (gaussian) profile in the center of the line whereas the wings are dominated by the Lorentz profile. Obviously an analysis of the line profile is crucial in understanding the physical conditions of the system producing the spectral line.

Acknowledgments: I would like to thank A.A. Deshpande for a critical reading of the manuscript and for useful comments to improve its clarity.

13.8 Further Reading

1. Cornwell, T. C. et al., 1992, A&A, 258, 583.
2. Harris, 1978, Proc. IEEE, 66, 51.
3. Lang, K. R. “Astrophysical formulae : a compendium for the physicist and astrophysicist”, Springer-Verlag, Berlin.
4. R. A. Perley, F. R. Schwab, & A. H. Bridle, eds., ‘Synthesis Imaging in Radio Astronomy’, ASP Conf. Series, vol. 6.
5. Rybicki, G. B. & Lightman, A. P., “Radiative Process in Astrophysics”, John Wiley, New York.
6. Thompson, R. A., Moran, J. M. & Swenson, G. W. Jr., ‘Interferometry & Synthesis in Radio Astronomy’, Wiley Interscience, 1986.

Chapter 14

Wide Field Imaging

Sanjay Bhatnagar

14.1 Introduction

It has been shown in Chapter 2 that the visibility measured by the interferometer, ignoring the phase rotation, is given by

$$V(u,v,w) = \int\int I(l,m)B(l,m)e^{-2\pi\iota(ul+vm+w(\sqrt{1-l^2-m^2}))}\frac{dldm}{\sqrt{1-l^2-m^2}}, \quad (14.1.1)$$

where (u,v,w) defines the co-ordinate system of antenna spacings, (l,m,n) defines the direction cosines in the (u,v,w) co-ordinates system, I is the source brightness distribution (the image) and B is the far field antenna reception pattern. For further analysis we will assume $B=1$, and drop it from all equations (for typing convenience[1]!)

Eq. 14.1.1 is not a Fourier transform relation. For a small field of view ($l^2+m^2 << 1$) the above equation however can be approximated well by a 2D Fourier transform relation. The other case in which this is an exact 2D relation is when the antennas are arranged in a perfect East-West line. However often array configurations are designed to maximize the uv-coverage and the antennas are arranged in a 'Y' shaped configuration. Hence, Eq. 14.1.1 needs to be used to map full primary beam of the antennas, particularly at low frequencies. Eq. 14.1.1 reduces to a 2D relation also for non-EW arrays if the time of observations is sufficiently small (snapshot observations).

In the first part of this chapter we will discuss the implications of approximating Eq. 14.1.1 by a 2D Fourier transform relation and techniques to recover the 2D sky brightness distribution.

The field of view of a telescope is limited by the primary beams of the antennas. To map a region of sky where the emission is at a scale larger than the angular width of the primary beams, mosaicing needs to be done. This is discussed in the second part of this lecture.

[1]The same assumptuin has been made in Chapter 2

14.2 Mapping with Non Co-planar Arrays

14.2.1 Image Volume

Let $n = \sqrt{1-l^2-m^2}$ be treated as an independent variable. Then one can write a 3D Fourier transform of $V(u,v,w)$ with the conjugate variable for (u,v,w) being (l,m,n), as

$$F(l,m,n) = \int\int\int V(u,v,w)e^{2\pi\iota(ul+vm+wn)}dudvdw. \tag{14.2.2}$$

Substituting for $V(u,v,w)$ from Eq. 14.1.1 we get

$$F(l,m,n) = \int\int\left\{\int\int\int \frac{I(l',m')}{\sqrt{1-l^2-m^2}}e^{-2\pi\iota(u(l'-l)+v(m'-m))}e^{-2\pi\iota(w(\sqrt{1-l'^2-m'^2}-n))}dudvdw\right\}dl'dm'. \tag{14.2.3}$$

Using the general result

$$\delta(l'-l) = \int e^{-2\pi\iota u(l'-l)}du, \tag{14.2.4}$$

we get

$$F(l,m,n) = \int\int \frac{I(l',m')}{\sqrt{1-l^2-m^2}}\delta(l'-l)\delta(m'-m)\delta(\sqrt{1-l'^2-m'^2}-n)dl'dm'. \tag{14.2.5}$$

This equation then provides the connection between the 2D sky brightness distribution given by $I(l,m)$ and the result of 3D Fourier inversion of $V(u,v,w)$ given by $F(l,m,n)$ referred to as the *Image volume*.

$$F(l,m,n) = \frac{I(l,m)\delta(\sqrt{1-l^2-m^2}-n)}{\sqrt{1-l^2-m^2}}. \tag{14.2.6}$$

Hereafter, I would use $I(l,m,n)$ to refer to the this *Image volume*.

In Eq.14.1.1, we have ignored the fringe rotation term $2\pi\iota w$ in the exponent. This is done here only for mathematical (and typing!) convenience. The effect of including this term would be a shift of the *Image volume* by one unit in the conjugate axis, namely n. Hence, the effect of fringe stopping is to make the top most plane of $I(l,m,n)$ tangent to the phase center position on the celestial sphere with the rest of the sphere completely contained inside the *Image volume* as shown in Fig. 14.1.

Remember that the third variable n of the *Image volume* is not an independent variable and is constrained to be $n = \sqrt{1-l^2-m^2}$. Eq 14.2.6 then gives the physical interpretation of $I(l,m,n)$. Imagine the celestial sphere defined by (l,m,n) enclosed by the *Image volume* $I(l,m,n)$, with the top most plane being tangent to the celestial sphere as shown in Fig. 14.1. Eq. 14.2.6 then says that only those parts of the *Image volume* correspond to the physical emission which lie on the surface of the celestial sphere. Note that since the visibility is written as a function of all the three variables (u,v,w), the transfer function will also be a volume. A little thought will then reveal that $I(l,m,n)$ will be finite away from the surface of the celestial sphere also, but that would correspond to non-physical emission in the *Image volume* due to the side lobes of the telescope transfer function (referred to by *Point spread function (PSF)* or *Dirty beam* in the literature). A 3D deconvolution using the *Dirty image-* and the *Dirty beam-volumes* will produce a *Clean image-volume*. Therefore, after deconvolution, one must perform an extra operation of projecting all points in the *image volume* along the celestial sphere onto the 2D tangent plane to recover the 2D sky brightness distribution. Fig. 14.2 is the graphical equivalent of the statements in this paragraph.

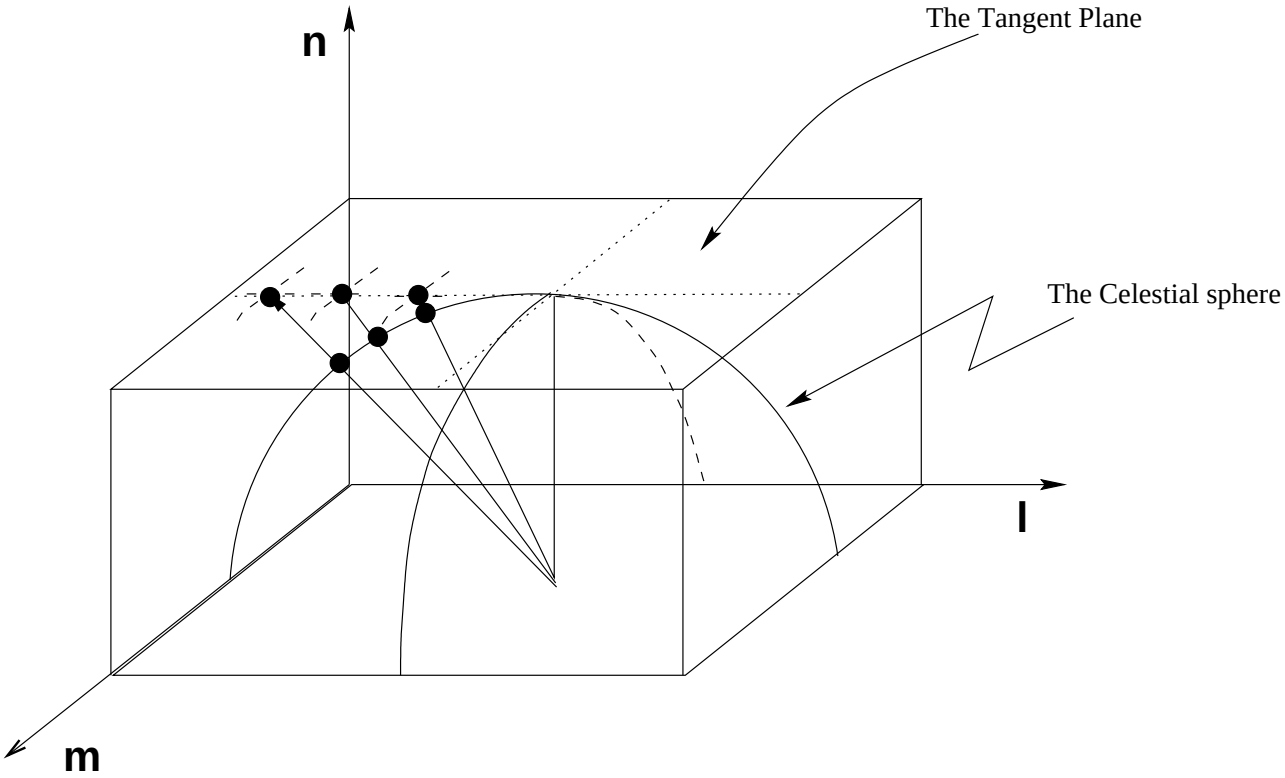


Figure 14.1: Graphical representation of the geometry of the *Image volume* and the celestial sphere. The point at which the celestial sphere touches the first plane of the *Image volume* is the point around which the 2D image inversion approximation is valid. For wider fields, emission at points along the intersection of celestial sphere and the various planes (labeled here as the celestial sphere) needs to be projected to the tangent plane to recover the undistorted 2D image. This is shown for 3 points on the celestial sphere, projected on the tangent plane, along the radial directions.

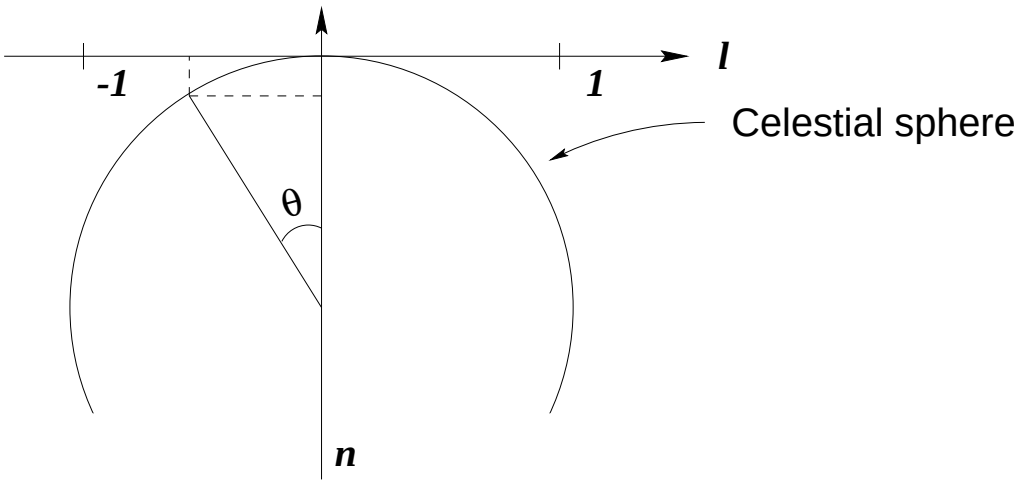


Figure 14.2: Graphical illustration to compute the distance between the tangent plane and a point in the sky at an angle of θ.

14.2.2 Interpretation of the w-term

The term $w\sqrt{1-l^2-m^2}$ is often referred to as the w-term in the literature. The origin of this term is purely geometrical and arises due to the fact that fringe rotation effectively phases the array for a point in the sky referred to as the phase center direction. A wave front originating for this direction will then be received by all antennas and the signals will be multiplied in-phase at the correlator (effectively phasing the array). The locus of all points in 3D space, for which the array will remain phased is a sphere, referred to as the celestial sphere. A wave front from a point away from the phase tracking center but on the surface of such a sphere, will carry an extra phase, not due to the geometry of the array but because of its separation from the phase center. In that sense, the phase of the wavefront measured by a properly phased array in fact carries the information about the source structure and the w-term is the extra phase due to the spherical geometry of the problem. The sky can be approximated by a 2D plane *close* to the phase tracking center and the w-term can be ignored, which is another way of saying that a 2D approximation can be made for a small field of view. However sufficiently far away from the phase center, the phase due to the curvature of the celestial sphere, the w-term, must be take into account, and to continue to approximate the sky as a 2D plane, we will have to rotate the visibility by the w-term. This will be equivalent to shifting the phase centre and corresponds to a shift of the equivalent point in the image plane. Since the w-term is a function of the image co-ordinates, this shift is different for different parts of the image. Shifting the phase centre to any *one* of the points in the sky, will allow a 2D approximation only *around* that direction and *not* for the entire image. Hence the errors arising due to ignoring the w-term cannot be removed by a constant phase rotation of all the visibilities. This is another way of understanding that, in the strict sense, the sky brightness is *not* a Fourier transform of the visibilities.

14.2.3 Inversion Of Visibilities

3D Imaging

The most straight forward method suggested by Eq. 14.2.5 for recovering the sky brightness distribution, is to perform a 3D Fourier transform of $V(u, v, w)$. This requires that the w axis be also sampled at least at Nyquist rate. For most observations it turns out that this is rarely satisfied and doing a FFT on the third axis would result into severe aliasing. Therefore in practice, the transform on third axis is usually done using the direct Fourier transform (DFT), on the un-gridded data.

For performing the 3D FT (FFT on the u and v axis and DT on the w axis) one would still need to know the number of planes needed along the n axis. This can be found using the geometry as shown in Fig. 14.2. The size of the synthesized beam in the n direction is comparable to that in the other two directions and is given by $\approx \lambda/B_{max}$ where B_{max} is the longest projected baseline length. Therefore the separation between the planes along n should be $\leq \lambda/2B_{max}$. The distance between the tangent plane and points separated by θ from the phase center is given by $1 - cos(\theta) \approx \theta^2/2$. For critical sampling then would be

$$N_n = B_{max}\theta^2/\lambda. \tag{14.2.7}$$

At 327 MHz for GMRT, $B_{max} \approx 25$ km. Therefore, for mapping 1° field of view without distortions, one would required 8 planes along the n axis. With central square alone however, one plane should be sufficient. At these frequencies it becomes important to map most of the primary beam since the number and the intensity of the background sources increase and the side lobes of these background sources limit the dynamic range in the maps. Hence, even if the source of interest is small, to get the achievable dynamic range (or close to it!), one will need to do a 3D inversion (and deconvolution).

Another reason why more than one plane would be required for very high dynamic range imaging is as follows. Strictly speaking, the only point which completely lies in the tangent plane is the point at which the tangent plane touches the celestial sphere. All other points in the image, even close to the phase center, lie slightly below the tangent plane. Deconvolution of the tangent plane then results into distortions for the same reason as the distortions arriving from the deconvolution of a point source which lies between two pixels in the 2D case. As in the 2D case, this problem can be minimized by over sampling the image and that, in this case, implies having at least 2 planes in the n axis, even if the Eq. 14.2.7 tells that 1 plane is sufficient.

Polyhedron Imaging

As mentioned above, emission from the phase center and from points close to it lie approximately in the tangent plane. Polyhedron imaging relies on exploiting this fact by approximating the celestial sphere by a number of tangent planes as shown in Fig. 14.3. The visibility data is phase rotated to shift the phase center to the tangent points of the various planes and a small region around the tangent point is then mapped using the 2D approximation. In this case however, one needs to perform a joint deconvolution involving all tangent planes since the sides lobes of a source in one plane would leak into other planes as well.

The number of planes required to map an object of size θ can be found simply by requiring that maximum separation between the tangent plane and the region around each tangent point be less than λ/B_{max}, the size of the synthesized beam. As shown earlier, the separation of a point θ degrees away from the tangent point is $\approx \theta^2$. Hence for critical sampling, the number of planes required is equal to the solid angle subtended by

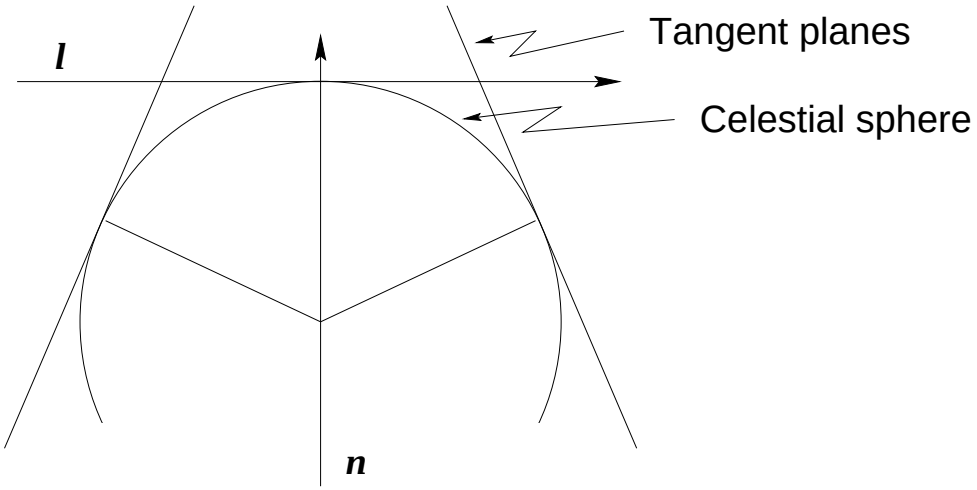


Figure 14.3: Approximation of the celestial sphere by multiple tangent planes (polyhedron imaging).

the sky being mapped (θ_f^2) divided by the solid angle of the synthesized beam (θ^2)

$$N_{poly} \;=\; 2\theta_f^2 B_{max}/\lambda \;=\; 2B_{max}\lambda/D^2 \;\; (\text{for } \theta_f \;=\; \text{full primary beam}). \tag{14.2.8}$$

Notice that the number of planes required is twice as many as the number of planes required for 3D inversion. However since a small portion around the tangent point of each plane is used, the size of each of these planes can be small, offsetting the increase in computations due to the increase in the number of planes required. Another approach which is often taken for very high dynamic range imaging is to do a full 3D imaging on each of the planes. This would effectively increase the size of the field that can be imaged on each tangent plane, thereby reducing the number of planes required.

The polyhedron imaging scheme is available in the current version of AIPS data reduction package and the 3D inversion (and deconvolution) is implemented in the (not any more supported) SDE package written by Tim Cornwell et al. Both these schemes, in their full glory, will be available in the (recently released) AIPS++ package.

14.3 Mosaicing

The problem due to non co-planarity discussed above are for mapping the sky within the primary beam of the antennas (which are assumed to be identical). In this section we discuss the techniques used to handle the problem of mapping fields of interest which are larger than the primary beam of the antennas. The approach used is similar to that used for mapping with a single dish, namely to scan the source to be mapped. The fact that we are using an interferometer to synthesis the "lens" (or the a "single dish") adds some more complications.

These techniques are useful for mapping with interferometers operating in the millimeter range where the size of the primary beams is less than an arcmin and at meter wavelengths where the primary beams are larger but so is the extent of emission. For

example, the primary beam of GMRT antennas at 327 MHz is $\approx 1.3^\circ$ and there are mapping projects which would benefit from mapping regions of the sky larger than this (for example, in the Galactic plane).

14.3.1 Scanning Interferometer

The co-planar approximation of Eq. 14.1.1 for a pointing direction given by (l_o, m_o) can be written as

$$V(u,v,l_o,m_o) = \int\int I(l,m)B(l-l_o,m-m_o)e^{2\pi\iota(ul+vm)}dldm. \tag{14.3.9}$$

Here we also assume that B is independent of the pointing direction and we label V with not just the (u,v) co-ordinates, but also with pointing direction since visibilities for different directions will be used in the analysis that follows. The advantage of writing the visibility as in Eq. 14.3.9 is that the pointing center (given by (l_o, m_o)) and the phase center (given by $(l,m) = (0,0)$) are separated.

$V(0,0,l_o,m_o)$ represents the single dish observation in the direction (l_o, m_o) and is just the convolution of the primary beam with the source brightness distribution, exactly as expected intuitively. Extending the intuition further, as is done in mapping with a single dish, we need to scan the source around (l_o, m_o) with the interferometer, which is equivalent to scanning with a single dish with a primary beam of the size of the synthesized beam of the interferometer. Then Fourier transforming $V(u,v,l_o,m_o)$ with respect to (l_o, m_o), assuming that B is symmetric, one gets, from Eq. 14.3.9

$$\int\int V(u,v,l_o,m_o)e^{2\pi\iota(u_o l_o + v_o m_o)}dl_o dm_o = b(u_o,v_o)i(u+u_o, v+v_o), \tag{14.3.10}$$

where (u_o, v_o) corresponds to the direction (l_o, m_o) and $b \rightleftharpoons B$ and $i \rightleftharpoons I$. This equation essentially tells us the following: Fourier transform of the visibility with respect to the pointing directions, from a scanning interferometer is equal to the visibility of the *entire source* modulated by the Fourier transform of the primary beams for each pointing direction. For a given direction (l_o, m_o) we can recover spatial frequency information spread around a nominal point (u,v) by an amount D/λ where D is the size of the dish. In terms of information, this is exactly same as recovering spatial information smaller than the size of the resolution of a single dish by scanning the source with a single dish. As in the case of a single dish, continuous scanning is not necessary and two points separated by half the primary beam is sufficient. In principle then, by scanning the interferometer, one can improve the short spacings measurements of V, which is crucial for mapping large fields of view.

Image of the sky can now be made using the full visibility data set (made using the Eq. 14.3.10). However, this involves the knowledge of Fourier transform of the sky brightness distribution, which in-turn is approximated after deconvolution. Hence, in practice one uses the MEM based image recovery where one maximizes the entropy given by

$$H = -\sum_k I_k ln\frac{I_k}{M_k}, \tag{14.3.11}$$

with χ^2 evaluated as

$$\chi^2 = \sum_k \frac{|V(u_k,v_k,l_{ok},m_{ok}) - V^M(u_k,v_k,l_{ok},m_{ok})|^2}{\sigma^2_{V(u_k,v_k,l_{ok},m_{ok})}}, \tag{14.3.12}$$

where $V^M(u_k,v_k,l_{ok},m_{ok})$ is the model visibility evaluated using Eq. 14.3.9. For calculation of $\Delta\chi^2$ in each iteration is estimated by the following steps:

- initialize $\Delta\chi^2 = 0$
- For all pointings
  1. Apply the appropriate primary beam correction to the current estimate of the image
  2. FT to generated V^M
  3. Accumulate χ^2
  4. Subtract from the observed visibilities
  5. Make the residual image
  6. Apply the primary beam correction to the residual image
  7. Accumulate $\Delta\chi^2$

The operation of primary beam correction on the residual image is understood by the following argument: For any given pointing, an interferometer gathers radiation within the primary beam. In the image plane then, any feature, outside the range of the primary beam would be due to the side lobes of the synthesized beam and must be suppressed before computation of $\Delta\chi^2$ and this is achieved by primary beam correction, which essentially divides the image by gaussian which represents the main lobe of the antenna radiation pattern.

This approach (rather than joint deconvolution) has several advantages.

1. Data from potentially different interferometers for different pointings can be used
2. Weights on each visibility from each pointing are used in the entire image reconstruction procedure
3. Single-dish imaging emerges as a special case
4. It is fast for extended images

The most important advantage that one gets by MEM reconstruction is that the deconvolution is done simultaneously on all points. That this is an advantage over joint-deconvolution can be seen as follows: If a point source at the edge of the primary beam is sampled by 4 different pointings of the telescope, this procedure would be able to use 4 times the data on the same source as against data from only one pointing in joint-deconvolution (where deconvolution is done separately on each pointing). This, apart from improvement in the signal-to-noise ratio also benefits from a better uv-coverage available.

Flexible software for performing Mosaic-ed observations is one of the primary motivation driving the AIPS++ project in which algorithms to handle mosaic-ed observations would be available in full glory.

14.4 Further Reading

1. Interferometry and Synthesis in Radio Astronomy; Thompson, A. Richard, Moran, James M., Swenson Jr., George W.; Wiley-Interscience Publication, 1986.
2. Synthesis Imaging In Radio Astronomy; Eds. Perley, Richard A., Schwab, Frederic R., and Bridle, Alan H.; ASP Conference Series, Vol 6.

Chapter 15

Polarimetry

Jayaram N. Chengalur

15.1 Introduction

Consider the simplest kind of electromagnetic wave, i.e. a plane monochromatic wave of frequency ν propagating along the +Z axis of a cartesian co-ordinate system. Since electro-magnetic waves are transverse, the electric field $\mathbf{E}$ must lie in the X-Y plane. Further since the wave is mono-chromatic one can write

$$\mathbf{E}(t) = E_x \cos(2\pi\nu t)\mathbf{e}_x + E_y \cos(2\pi\nu t + \delta)\mathbf{e}_y, \tag{15.1.1}$$

i.e. the X and Y components of the electric field differ in phase by a factor which does not depend on time. It can be shown[1] that the implication of this is that over the course of one period of oscillation, the tip of the electric field vector in general traces out an ellipse. There are two special cases of interest. The first is when $\delta = 0$. In this case the tip of the electric field vector traces out a line segment, and the wave is said to be *linearly polarized*. The other special case is when $E_x = E_y$ and $\delta = \pm\pi/2$. In this case the electric field vector traces out a circle in the X-Y plane, and depending on the sense[2] in which this circle is traversed the wave is called either *left circular polarized* or *right circular polarized*.

As you have already seen in chapter 1, signals in radio astronomy are not monochromatic waves, but are better described as quasi-monochromatic plane waves[3]. Further, the quantity that is typically measured in radio astronomy is not related to the field (i.e. a voltage), but rather a quantity that has units of voltage squared, i.e. related to some correlation function of the field (see chapter 4). For these reasons, it is usual to characterize the polarization properties of the incoming radio signals using quantities called *Stokes* parameters. Recall that for a quasi monochromatic wave, the electric field $\mathbf{E}$ could be considered to be the real part of a complex analytical signal $\mathcal{E}(t)$. If the X and Y components of this complex analytical signal are $\mathcal{E}_x(t)$, and $\mathcal{E}_y(t)$, respectively, then the four

[1] See for example, Born & Wolf 'Principles of Optics', Sixth Edition, Section 1.4.2

[2] Note that there is an additional ambiguity here, i.e. are you looking along the direction of propagation of the wave, or against it? To keep things interesting neither convention is universally accepted, although in principle one should follow the convention adopted by the IAU (Transactions of the IAU Vol. 15B, (1973), 166.)

[3] Recall that as all astrophysically interesting sources are distant, the plane wave approximation is a good one

Stokes parameters are defined as:

$$\begin{array}{rcl} I &=& <\mathcal{E}_x\mathcal{E}_x^*> + <\mathcal{E}_y\mathcal{E}_y^*> \\ Q &=& <\mathcal{E}_x\mathcal{E}_x^*> - <\mathcal{E}_y\mathcal{E}_y^*> \\ U &=& <\mathcal{E}_x\mathcal{E}_y^*> + <\mathcal{E}_x^*\mathcal{E}_y> \\ V &=& \frac{1}{i}(<\mathcal{E}_x\mathcal{E}_y^*> - <\mathcal{E}_x^*\mathcal{E}_y>) \end{array} \quad \text{or} \quad \begin{array}{rcl} <\mathcal{E}_x\mathcal{E}_x^*> &=& (I+Q)/2 \\ <\mathcal{E}_y\mathcal{E}_y^*> &=& (I-Q)/2 \\ <\mathcal{E}_x\mathcal{E}_y^*> &=& (U+iV)/2 \\ <\mathcal{E}_x^*\mathcal{E}_y> &=& (U-iV)/2. \end{array} \tag{15.1.2}$$

where the angle brackets indicate taking the average value[4]. The Stokes parameters as defined in equation (15.1.2) clearly depend on the orientation of the co-ordinate system. In radio astronomy it is conventional (see chapter 10) to take the +X axis to point north and the +Y axis to point east. It is important to realize that the Stokes parameters are descriptors of the intrinsic polarization state of the electro-magnetic wave, i.e. the Stokes vector $(I\ Q\ U\ V)^T$ is a true vector. The equations (15.1.2) simply give its components in a particular co-ordinate system, the linear polarization co-ordinate system[5]. One would instead work in a circularly polarized reference frame, i.e. where the electric field is decomposed into two circularly polarized components, $\mathcal{E}_r(t)$, and $\mathcal{E}_l(t)$. The relation between these components and the Stokes parameters are:

$$\begin{array}{rcl} I &=& <\mathcal{E}_r\mathcal{E}_r^*> + <\mathcal{E}_l\mathcal{E}_l^*> \\ Q &=& <\mathcal{E}_r\mathcal{E}_l^*> + <\mathcal{E}_r^*\mathcal{E}_l> \\ U &=& \frac{1}{i}(<\mathcal{E}_r\mathcal{E}_l^*> - <\mathcal{E}_r^*\mathcal{E}_l>) \\ V &=& <\mathcal{E}_r\mathcal{E}_r^*> - <\mathcal{E}_l\mathcal{E}_l^*> \end{array} \quad \begin{array}{rcl} <\mathcal{E}_r\mathcal{E}_r^*> &=& (I+V)/2 \\ <\mathcal{E}_l\mathcal{E}_l^*> &=& (I-V)/2 \\ <\mathcal{E}_r\mathcal{E}_l^*> &=& (Q+iU)/2 \\ <\mathcal{E}_r^*\mathcal{E}_l> &=& (Q-iU)/2. \end{array} \tag{15.1.3}$$

Interestingly, equations (15.1.3) are formally identical to equations (15.1.2) apart from the following transformations viz. $Q^+ \rightarrow V^\odot$, $U^+ \rightarrow Q^\odot$, $V^+ \rightarrow U^\odot$, where the superscript $+$ indicates linear polarized co-ordinates and $\odot$ circular polarized co-ordinates. Although these two co-ordinate systems are the ones most frequently used, the Stokes vector could in principle be written in any co-ordinate system based on two linearly independent (but not necessarily orthogonal) polarization states. In fact, as we shall see, such non orthogonal co-ordinate systems will arise naturally when trying to describe measurements made with non ideal radio telescopes.

The degree of polarization of the wave is defined as

$$P = \frac{\sqrt{Q^2+U^2+V^2}}{I}. \tag{15.1.4}$$

From equation (15.1.2) we have

$$I^2 - Q^2 - U^2 - V^2 = 2\left(\langle\mathcal{E}_x^2\rangle\langle\mathcal{E}_y^2\rangle - \langle\mathcal{E}_x\mathcal{E}_y\rangle^2\right) \tag{15.1.5}$$

and hence from the Schwarz inequality it follows that $0 \leq P \leq 1$ and that $P = 1$ iff $\mathcal{E}_x = c\mathcal{E}_y$, where c is some complex constant. For a mono-chromatic plane wave (equation (15.1.1)) therefore, $P = 1$ or equivalently $I^2 = Q^2 + U^2 + V^2$, i.e. there are only three independent Stokes parameters. For a general quasi mono-chromatic wave, $P < 1$, and the wave is said to be *partially polarized*.

It is also instructive to examine the Stokes parameters separately for the special case of a monochromatic plane wave. We have (see equations (15.1.1) and (15.1.2)):

$$\begin{array}{rcl} I &=& E_x^2 + E_y^2 \\ Q &=& E_x^2 - E_y^2 \end{array} \qquad \begin{array}{rcl} U &=& 2E_xE_y\cos(\delta) \\ V &=& 2E_xE_y\sin(\delta), \end{array}$$

[4]Strictly speaking this is the ensemble average. However, as always, we will assume that the signals are ergodic, i.e. the ensemble average can be replaced with the time average.

[5]These polaraization co-ordinate systems are of course in some abstract polarization space and not real space

i.e. for a linearly polarized wave ($\delta = 0$) we have V = 0, and for a circularly polarized wave ($E_x = E_y, \delta = \pm\pi/2$) we have $Q = U = 0$. So Q and U measure linear polarization, and V measures circular polarization. This intepretation continues to be true in the case of partially polarized waves.

15.2 Polarization in Radio Astronomy

Emission mechanisms which are dominant in low frequency radio astronomy, produce linearly polarized emission. Thus extra-galactic radio sources and pulsars are predominantly linearly polarized, with polarization fractions of typically a few percent. These sources usually have no circular polarization, i.e. $V \sim 0$. Maser sources however, in particular OH masers from galactic star forming regions often have significant circular polarization. This is believed to arise because of Zeeman splitting. Interstellar maser sources also often have some linear polarization, i.e. all the components of the Stokes vector are non zero. In radio astronomy the polarization is fundamentally related to the presence of magnetic fields, and polarization studies of sources are aimed at understanding their magnetic fields.

The raw polarization measured by a radio telescope could differ from the true polarization of the source because of a number of effects, some due to propagation of the wave through the medium between the source and the telescope, (see chapter 16) and the other because of various instrumental non-idealities. Since we are eventually interested in the true source polarization our ultimate aim will be to correct for these various effects, and we will therefore find it important to distinguish between depolarizing and non-depolarizing systems. A system for which the outgoing wave is fully polarized if the incoming wave is fully polarized is called non-depolarizing. The polarization state of the output wave need not be identical to that of the incoming wave, it is only necessary that $P_{out} = 1$ if $P_{in} = 1$.

The most important propagation effect is *Faraday rotation*, which is covered in some detail in chapter 16. Here we restrict ourselves to stating that the plane of polarization of a linearly polarized wave is rotated on passing through a magnetized plasma. Faraday rotation can occur both in the ISM as well as in the earth's ionosphere. If the Faraday rotating medium is mixed up with the emitting region, then radiation emitted from different depths along the line of sight are rotated by different amounts, thus reducing the net polarization. This is called *Faraday depolarization*. If the medium is located between the source and the observer, then the only effect is a net rotation of the plane of polarization, i.e.

$$\mathcal{E}_x^{'} = \mathcal{E}_x \cos\chi + \mathcal{E}_y \sin\chi, \qquad \mathcal{E}_y^{'} = -\mathcal{E}_x \sin\chi + \mathcal{E}_y \cos\chi, \tag{15.2.6}$$

where $\mathcal{E}_x$, $\mathcal{E}_x^{'}$ are the X components of the incident and emergent field respectively and similarly for $\mathcal{E}_y$, $\mathcal{E}_y^{'}$. In terms of the Stokes paramters, the transformation on passing through a Faraday rotating medium is

$$\begin{array}{rclrcl} I^{'} & = & I & Q^{'} & = & Q\cos 2\chi + U\sin 2\chi \\ V^{'} & = & V & U^{'} & = & -Q\sin 2\chi + U\cos 2\chi. \end{array} \tag{15.2.7}$$

i.e. a rotation of the Stokes vector in the (U,V) plane. The fractional polarization is hence preserved[6]. Equation (15.2.7) can also be easily obtained from equation (15.1.3)

[6]Note that non-depolarizing only means that $P_{out} = 1$ if $P_{in} = 1$, and this does not necessarily translate into conservation of the fractional polarization when $P < 1$. Pure faraday rotation is hence not only non-depolarizing, it also preserves the fractional polarization.

by noting that in a circularly polarized co-ordinate system, the effect of faraday rotation is to introduce a phase difference of 2χ between $\mathcal{E}_r$ and $\mathcal{E}_l$.

Consider looking at an extended source which is not uniformly polarized with a radio telescope whose resolution is poorer than the angular scale over which the source polarization is coherent. In any given resolution element then there are regions with different polarization characteristics. The beam thus smoothes out the polarization of the source, and the measured polarization will be less than the true source polarization. This is called *beam depolarization*. Beam depolarization cannot in principle be corrected for, the only way to obtain the true source polarization is to observe with sufficiently high angular resolution.

A dual polarized radio telescope has two voltage beam patterns, one for each polarization. These two patterns are often not symmetrical, i.e. in certain directions the telescope response is greater for one polarization than for the other. The difference in gain between these two polarizations usually varies in a systematic way over the primary beam. Because of this asymmetry, an unpolarized source could appear to be polarized, and further its apparent Stokes parameters in general depend on its location with respect to the center of the primary beam. The polarization properties of an antenna are also sharply modulated by the presence of feed legs, etc. and are hence difficult to determine with sufficient accuracy. For this reason determining the polarization across sources with dimensions comparable to the primary beam is a non trivial problem. Given the complexity of dealing with extended sources, most analysis to date have been restricted to small sources, ideally point sources located at the beam center.

Most radio telescopes measure non-orthogonal polarizations, i.e. a channel p which is supposed to be matched to some particular polarization p also picks up a small quantity of the orthogonal polarization q. Further, this leakage of the orthogonal polarization in general changes with position in the beam. However, for reflector antennas, there is often a leakage term that is independent of the location in the beam, which is traditionally ascribed to non idealities in the feed. For example, for dipole feeds, if the two dipoles are not mounted exactly at right angles to one another, the result is a real leakage term, and if the dipole is actually matched to a slightly elliptical (and not purely linear) polarization the result is an imaginary leakage term. For this reason, the real part of the leakage is sometimes called an *orientation* error, and the imaginary part of the leakage is referred to as an *ellipticity* error[7]. However, one should appreciate that the actual measurable quantity is only the antenna voltage beam, (i.e. the combined response of the feed and reflector) and this decomposition into 'feed' related terms is not fundamental and need not in general be physically meaningful.

The final effect that has to be taken into account has to do with the orientation of the antenna beam with respect to the source. For equitorially mounted telescopes this is a constant, however for alt-az mounted telescopes, the telescope beam rotates on the sky as the telescope tracks the source. This rotation is characterized by an angle called the parallactic angle, ψ_p, which is given by:

$$\tan\psi_p = \frac{\cos\mathcal{L}\sin\mathcal{H}}{\sin\mathcal{L}\cos\delta - \cos\mathcal{L}\sin\delta\sin\mathcal{H}}, \tag{15.2.8}$$

where $\mathcal{L}$ is the latitude of the telescope, $\mathcal{H}$ is the hour-angle of the source, and δ is the apparent declination of the source. So if one observes a source at a parallactic angle ψ_p with a telescope that is linearly polarized, the voltages that will be obtained at the

[7]Several telescopes, such as for example the GMRT, use feeds which are sensitive to linear polarization, but by using appropriate circuitry (viz a $\pi/2$ phase lag along one signal path before the first RF amplifier) convert the signals into circular polarization. Non idealities in this linear to circular conversion circuit could also produce complex leakage terms even if the feed dipoles themselves are error free.

terminals of the X and Y receivers will be

$$V_x = G_x(\mathcal{E}_x \cos\psi_p + \mathcal{E}_y \sin\psi_p), \quad V_y = G_y(-\mathcal{E}_x \sin\psi_p + \mathcal{E}_y \cos\psi_p), \tag{15.2.9}$$

where G_x and G_y are the complex gains (i.e. the product of the antenna voltage gains and the receiver gains) of the X and Y channels.

15.3 The Measurement Equation

In this section we will develop a mathematical formulation useful for polarimetric interferometry. The theoretical framework is the van Cittert-Zernike theorem, which was discussed in chapter 2 in the context of the reconstruction of the Stokes I parameter of the source. However, as can be trivially verified, the theorem holds good for any of the Stokes parameters. So, apart from the issues of spurious polarization produced by propagation or instrumental effects, making maps of the Q, U, and V Stokes parameters is in principle[8] identical to making a Stokes I map.

Not surprisingly, matrix notation leads to an elegant formulation for polarimetric interferometry[9]. Let us begin by defining a coherency vector,

$$\begin{pmatrix} < \mathcal{E}_{ap}\mathcal{E}^*_{bp} > \\ < \mathcal{E}_{ap}\mathcal{E}^*_{bq} > \\ < \mathcal{E}_{aq}\mathcal{E}^*_{bp} > \\ < \mathcal{E}_{aq}\mathcal{E}^*_{bq} > \end{pmatrix},$$

where a, b refer to the two antennas which compose any given baseline, and p, q are the two polarizations measured by the antenna. The coherency vector can be expressed as an outer product of the electric field, viz:

$$\begin{pmatrix} < \mathcal{E}_{ap}\mathcal{E}^*_{bp} > \\ < \mathcal{E}_{ap}\mathcal{E}^*_{bq} > \\ < \mathcal{E}_{aq}\mathcal{E}^*_{bp} > \\ < \mathcal{E}_{aq}\mathcal{E}^*_{bq} > \end{pmatrix} = \left\langle \begin{pmatrix} \mathcal{E}_{ap} \\ \mathcal{E}_{aq} \end{pmatrix} \otimes \begin{pmatrix} \mathcal{E}^*_{bp} \\ \mathcal{E}^*_{bq} \end{pmatrix} \right\rangle. \tag{15.3.10}$$

The Stokes vector can be obtained by multiplying the coherency vector with the Stokes matrix, (**S**). In a linear polarized co-ordinate system the components are:

$$\begin{pmatrix} I \\ Q \\ U \\ V \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 1 & 0 \\ 0 & -i & i & 0 \end{pmatrix} \begin{pmatrix} < \mathcal{E}_{ax}\mathcal{E}^*_{bx} > \\ < \mathcal{E}_{ax}\mathcal{E}^*_{by} > \\ < \mathcal{E}_{ay}\mathcal{E}^*_{bx} > \\ < \mathcal{E}_{ay}\mathcal{E}^*_{by} > \end{pmatrix}. \tag{15.3.11}$$

The component form could also be written down in the circular polarized co-ordinate system, in which case the matrix **S** would be:

$$\begin{pmatrix} I \\ Q \\ U \\ V \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & -i & i & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} < \mathcal{E}_{ar}\mathcal{E}^*_{br} > \\ < \mathcal{E}_{ar}\mathcal{E}^*_{bl} > \\ < \mathcal{E}_{al}\mathcal{E}^*_{br} > \\ < \mathcal{E}_{al}\mathcal{E}^*_{bl} > \end{pmatrix}. \tag{15.3.12}$$

[8] apart from the fact that one has to record four correlation functions, $< \mathcal{E}_{ap}\mathcal{E}^*_{bp} >$, $< \mathcal{E}_{ap}\mathcal{E}^*_{bq} >$, $< \mathcal{E}_{aq}\mathcal{E}^*_{bp} >$, $< \mathcal{E}_{aq}\mathcal{E}^*_{bq} >$, where a, b refer to the two antennas which compose any given baseline, and p, q are the two polarizations measured by the antenna. Since Stokes I maps are often all that is required, many observatories, including the GMRT, make a trade off such that fewer spectral channels are available if you record all four correlation products, than if you recorded only the two correlation products which are required for Stokes I.

[9] Although this formulation has been in use in the field of optical polarimetry for decades, it was not appreciated until recently (Hamaker *et al.* 1996, and Sault *et al.* 1996) that it is also extendable to radio interferometric arrays.

The matrix in equation (15.3.12) is related to that in equation (15.3.11) by a simple permutation of rows, as expected.

The outer product has the following associative property, viz. for matrices, **A**,**B**, **C**, and **D**,

$$(\mathbf{AB}) \otimes (\mathbf{CD}) = (\mathbf{A} \otimes C)(\mathbf{B} \otimes D).$$

For any one antenna a, putting in all the various effects discussed in section(15.2) we can write the voltage at the antenna terminals as:

$$\begin{aligned} \mathcal{V}_a &= \mathbf{G}_a\mathbf{B}_a\mathbf{P}_a\mathbf{F}_a\mathcal{E}_a \\ &= \mathbf{J}_a\mathcal{E}_a. \end{aligned} \tag{15.3.13}$$

where,

$\mathcal{V}_a$	=	the voltage vector at the terminals of antenna a
$\mathbf{G}_a$	=	the complex gain of the receivers of antenna a
$\mathbf{B}_a$	=	the voltage beam matrix for antenna a
$\mathbf{P}_a$	=	the parallactic angle matrix for antenna a
$\mathbf{F}_a$	=	the Faraday rotation matrix for antenna a
$\mathcal{E}_a$	=	the electric field vector at antenna a
$\mathbf{J}_a$	=	the Jones matrix for antenna a

The Jones matrix has been so called because of its analogy with the Jones matrix in optical polarimetry. All of these matrices are 2×2. In the linear polarized co-ordinate system. For example, we have:

$$\begin{aligned} \mathbf{F} &= \begin{pmatrix} \cos\chi & \sin\chi \\ -\sin\chi & \cos\chi \end{pmatrix} \qquad \mathbf{P} = \begin{pmatrix} \cos\psi_p & \sin\psi_p \\ -\sin\psi_p & \cos\psi_p \end{pmatrix} \\ \mathbf{B} &= \begin{pmatrix} b_{pp}(l,m) & b_{pq}(l,m) \\ b_{qp}(l,m) & b_{qq}(l,m) \end{pmatrix} \quad \mathbf{G} = \begin{pmatrix} g_p & 0 \\ 0 & g_q \end{pmatrix}. \end{aligned} \tag{15.3.14}$$

The Jones matrix in polarimetric interferometry plays the same role as the complex gain does in scalar interferometry. Consequently one could conceive of schemes for self-calibration, since for an array with a large enough number of antennas sufficient number of closure constraints are available. However, since astrophysical sources are usually only weakly polarized, the signal to noise ratio in the cross-hand correlation products is often too low to make use of these closure constraints.

In scalar interferometry, phase fluctations caused by the atmosphere and/or ionosphere were lumped together with the instrumental gain fluctuations. In the vector formulation however, this is strictly speaking not possible, since these corrections occur at different points along the signal path, (see equations (15.3.13)) and matrices in equations (15.3.14) do not in general commute. However, for most existing radio telescopes, and for sources small compared to the primary beam, the matrices in equations (15.3.14) (apart from the Faraday rotation and Parallactic angle matrices) differ from the identity matrix only to first order (i.e. the off diagonal terms are small compared to the diagonal terms, and the diagonal terms are equal to one another to zeroth order), and consequently these matrices commute to first order. To first order hence, it is correct to lump the phase differences accumulated at different points along the signal path into the receiver gain. Alternatively, if we make the (reasonable) assumption that the complex atennuation (i.e. any absorption and phase fluctuation) produced by the atmosphere is identical for both polarizations, then it can be modeled as a constant times the identity matrix. Since the identity matrix commutes with all the other matrices, this factor can be absorbed in the receiver gain matrix, exactly as was done when dealing with interferometry of scalar

fields. This is the reason why no separate matrix was introduced in equation (15.3.13) to account for atmospheric phase and amplitude fluctuations.

The matrix **B** in this formulation also deserves some attention. It simply contains the information on the relation between the electric field falling on the source and the voltage generated at the antenna terminals. It is an extension of the voltage beam in scalar field theory, and each element in the matrix depends on the sky co-ordinates (l, m). As described above in section(15.2), it is traditional to decompose it into a part which does not depend on (l, m), which is called the leakage (or in the matrix formulation, the leakage matrix "**D**"), and a part which depends on (l, m). Provided that the leakage terms are small compared to the parallel hand antenna voltage gain, it can be shown that this decomposition is unique to first order.

In terms of the Jones matrix, the measured visibility on a single baseline for a point at the phase center can be written as:

$$\begin{pmatrix} \mathcal{V}_I \\ \mathcal{V}_Q \\ \mathcal{V}_U \\ \mathcal{V}_V \end{pmatrix} = \mathbf{S}\mathbf{J}_a \otimes \mathbf{J}_b^* \mathbf{S}^{-1} \begin{pmatrix} I \\ Q \\ U \\ V \end{pmatrix}. \tag{15.3.15}$$

Note that this is a matrix equation, valid in all co-ordinate frames, i.e. it holds regardless of whether the antennas are linear polarized or circular polarized. In fact it holds even if some of the antennas are linear polarized, and the others are circular polarized.

If the point source were not at the phase center, then the visibility phase is not zero, and in equation (15.3.15), one would have to pre-multipy the Jones matrices with a matrix containing the Fourier kernel, viz. $\mathbf{K}_a(l, m)$, and $\mathbf{K}_b(l, m)$defined as:

$$\begin{aligned} \mathbf{K}_a(l,m) &= \begin{pmatrix} e^{-2\pi(u_a l + v_a m)} & 0 \\ 0 & e^{-2\pi(u_a l + v_a m)} \end{pmatrix}, \\ \mathbf{K}_b(l,m) &= \begin{pmatrix} e^{-2\pi(u_b l + v_b m)} & 0 \\ 0 & e^{-2\pi(u_b l + v_b m)} \end{pmatrix}. \end{aligned} \tag{15.3.16}$$

To get the visibility for an extended incoherent source, one would have to integrate over all (l, m), thus recovering the vector formulation of the van Cittert-Zernike theorem. In order to invert this equation, it is necessary not only to do the inverse fourier transform, but also to correct for the various corruptions introduced, i.e. the data has to be calibrated. The rest of this chapter discusses ways in which this polarization calibration can be done.

15.4 Polarization Calibration

We restrict our attention to a point source at the phase center[10]. The visibility that we measure, averaged over all baselines is

$$\mathcal{V} = \frac{1}{N(N-1)} \mathbf{S} \sum_{a \neq b} (\mathbf{J}_a \otimes \mathbf{J}_b^*) \mathbf{S}^{-1}. \tag{15.4.17}$$

Any system describable by a Jones matrix is non-depolarizing[11] In the general case however, the summation in equation (15.4.17) cannot be represented by a single Jones

[10]For VLBI observations this is a very good approximation, since the source being imaged is very small compared to the primary beams of any of the antennas in the VLBI array.

[11]This follows trivially from the fact that for 100% polarization we must have $\mathcal{E}_p = c\mathcal{E}_q$, where p, q are any two orthogonal polarizations, and c is some complex constant. Multiplication by the Jones matrix will preserve this relationship (only changing the value of the constant c) thus producing another 100% polarized wave.

matrix, and an interferometer is not therefore a non-depolarizing system. However, ideally, after calibration, the effective Jones matrices are all the unit matrix, and the interferometer would then be non-depolarizing.

Intuitively, it is clear that if one looks at an unpolarized calibrator source, one should be able to solve for the leakage terms, (which will produce apparent polarization) but that some degrees of freedom would remain unconstrained. Further it is also intuitive that the degrees of freedom which remain unconstrained are the following: (1) The absolute orientation of the feeds, (2) The intrinsic polarization of the feeds (i.e. for example, are they linear polarized or circular polarized?) and (3) The phase difference between the two polarizations. While one would imagine that the situation may be improved by observation of a polarized source, it turns out that this too is not sufficient to determine all the free parameters. What is required is observations of at least three differently polarized sources. For alt-az mounted dishes, the rotation of the beam with respect to the sky changes the apparent polarization of the source. For such telescopes hence, it is sufficient to observe a single source at several, sufficiently different hour angles. This is the polarization strategy that is commonly used at most telescopes. Faraday rotation due to the earth's ionosphere is more difficult to correct for. In principle models of the ionosphere coupled with a measure of the total electron content at the time of the observation can be used to apply a first order correction to the data.

We end this chapter with a brief description of the effect of calibration errors on the derived Stokes parameters. When observing with linearly polarized feeds, from equation (15.1.2) it is clear that if one observes a linearly polarized calibrator, the parallel-hand correlations will contain a contribution due to the Q component of the calibrator flux. Consequently, if one assumes (erroneously) that the calibrator was unpolarized the gain of the X channel will be overestimated and that of the Y channel underestimated. For this reason, for observations which require only measurement of Stokes I, circular feeds are preferable, since the Stokes V component of most calibrators is negligible, and consequently, measurements of the parallel hand correlations[12] are sufficient to measure the correct Stokes I flux.

It is easy to show, that (to first order) if one observes a polarized calibrator with an error free linearly polarized interferometer and solves for the instrumental parameters under the assumption that the calibrator is unpolarized, the derived instrumental parameters of all the antennas will be in error by[13]:

$$\begin{array}{lllll} \Delta g_x & = & +Q/2I & \Delta g_y & = & -Q/2I \\ d_x & = & (Q+iU)/2I & d_y & = & -(U-iQ)/2I. \end{array}$$

where:

Δg_x	is the gain error of the X channel.
Δg_y	is the gain error of the Y channel.
d_x	is the leakage from the Y channel to the X channel.
d_y	is the leakage from the X channel to the Y channel.

If these calibration solutions are then applied to an unpolarized target source, then the source will appear to be polarized, with the same polarization percentage as the calibrator, but opposite sense. This again is simply the extension from scalar interferometry that if the calibrator flux is in error by some amount, the derived target source flux will be in error by the same fractional amount, but with opposite sense.

[12] recall from equations (15.1.3) that when $V = 0$, $< \mathcal{E}_r \mathcal{E}_r^* > + < \mathcal{E}_l \mathcal{E}_l^* >= I$.

[13] A similar result can of course be derived for the case of circularly polarized antennas, the only difference will be the usual transpositions of Q, U, and V.

15.5 Further Reading

1. Born, M. & Wolf, E., *'Principles of Optics'*, Cambridge University Press.

2. Hamaker, J. P., Bregman, J. D. & Sault, R. J., *'Understanding Polarimetric Interferometry I. Mathematical Foundations'*, A&A Supp. Ser., **117**,137, 1996.

3. C. Heiles, *A heuristic introduction to Radio Astronomic Polarisation*, astro-ph/0107372.

4. Sault, R. J., Hamaker, J. P. & Bregman, J. D. & *'Understanding Polarimetric Interferometry II. Instrumental Calibration of an Interferometric Array'*, A&A Supp. Ser., **117**,149, 1996.

5. Thompson, R. A., Moran, J. M. & Swenson, G. W. Jr., 'Interferometry & Synthesis in Radio Astronomy', Wiley Interscience, 1986.

Chapter 16

Ionospheric effects in Radio Astronomy

A. P. Rao

16.1 Introduction

At the low densities encountered in the further reaches of the earth's atmosphere and in outer space, collisions between particles are very rare. Hence, unlike in a terrestrial laboratory, it is possible for gas to remain in an ionized state for long periods of time. Such plasmas are ubiquitous in astrophysics, and have been extensively studied for their own sake. In this chapter however, we focus on the effects of this plasma on radio waves propagating through them, and will find astrophysical plasmas to be largely of nuisance value.

The refractive index of a cold neutral plasma is given by

$$\mu(\nu) = \sqrt{1 - \frac{\nu_p^2}{\nu^2}}, \tag{16.1.1}$$

where ν_p the "plasma frequency is given by

$$\nu_p = \sqrt{\frac{n_e e^2}{\pi m_e}} \simeq 9\sqrt{n_e} \ \ \mathrm{kHz} \tag{16.1.2}$$

where e is the charge on the electron, m_e is the mass of the electron and n_e is the electron number density (in cm^{-3}). At frequencies below the plasma frequency ν_p the refractive index becomes imaginary, i.e. the wave is exponentially attenuated and does not propagate through the medium. The earth's ionosphere has electron densities $\sim 10^4 - 10^5$ cm^{-3}, which means that the plasma frequency is $\sim 1 - 10$ MHz. Radio waves with such low frequencies do not reach the earth's surface and can be studied only by space based telescopes. The plasma between the planets is called the Interplanetary Medium (IPM) and has electron densities ~ 1 cm^{-3} (at the earth's location); the corresponding cut off frequency is ~ 9 kHz. The typical density in the Interstellar Medium (ISM) is ~ 0.03 cm^{-3} for which the cut off frequency is ~ 1 kHz. Waves of such low frequency from extra solar system objects cannot be observed even by spacecraft since the IPM and ISM will attenuate them severely.

The dispersion relationship in a cold plasma is given by $c^2k^2 = \omega^2 - \omega_p^2$. Since this is a non linear relation there are two characteristic velocities of propagation, the phase velocity given by

$$v_p = \frac{\omega}{k} = \frac{c}{\mu} \simeq c\ (1 + \frac{1}{2}\frac{\nu_p^2}{\nu^2}) \tag{16.1.3}$$

and the group velocity which is given by

$$v_g = \frac{d\omega}{dk} = c\mu \simeq c\ (1 - \frac{1}{2}\frac{\nu_p^2}{\nu^2}). \tag{16.1.4}$$

Where for the last expression we have assumed that $\nu >> \nu_p$ (which is usually the regime of interest).

16.2 Propagation Through a Homogeneous Plasma

Even above the cutoff frequency there are various propagation effects that are important for a radio wave passing through a plasma. Let us start with the most straightforward ones. Consider a radio signal passing through a homogeneous slab of plasma of length L. The signal is delayed (with respect to the propagation time in the absence of the plasma) by the amount

$$\Delta T = \frac{L}{v_g} - \frac{L}{c} = \frac{L}{c}(1/\mu - 1) \simeq \frac{L}{c}\frac{1}{2}\frac{\nu_p^2}{\nu_2}.$$

The magnitude of the propagation delay can hence be written as

$$|\Delta T| = \frac{L}{c} \times \frac{4 \times 10^6}{\nu_{\rm Hz}^2} n_e.$$

The propagation delay can also be considered as an "excess path length" ΔL = c ΔT. Further since $(v_g/c - 1)$ and $(v_p/c - 1)$ differ only in sign[1], the magnitude of the "excess phase" (viz. $2\pi\nu(L/v_p - L/c)$) is given by $\Delta\Phi = 2\pi\nu\Delta T$. Note that since the propagation delay is a function of frequency ν, waves of different frequencies get delayed by different amounts. A pulse of radiation incident at the far end of the slab will hence get smeared out on propagation through the slab; this is called "dispersion". If the plasma also has a magnetic field running through it then it becomes birefringent – the refractive index is different for right and left circularly polarized waves. A linearly polarized wave can be considered a superposition of left and right circularly polarized waves. On propagation through a magnetized plasma the right and left circularly polarized components are phase shifted by different amounts, or equivalently the plane of polarization of the linearly polarized component is rotated. This rotation of the plane of polarization on passage through a magnetized plasma is called "Faraday rotation". The angle through which the plane of polarization is rotated is given by

$$\Theta = RM\lambda^2 = 0.81\lambda^2 \int n_e B_{||} dl.$$

and RM is called the rotation measure. For the second equality λ is in meters, n_e is in cm^{-3}, $B_{||}$ is in μG and the length is in parsecs.

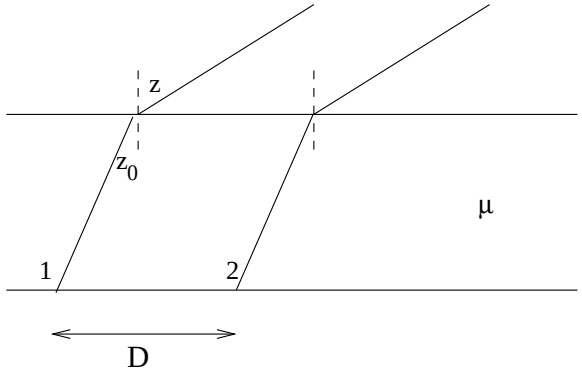


Figure 16.1: Propagation through a plane parallel ionosphere

16.3 Propagation Through a Smooth Ionosphere

For an interferometer, there are two quantities of interest (i) the delay difference between the signals reaching the two arms of the interferometer ($\delta T = \Delta T_1 - \Delta T_2$), where ΔT_1 and ΔT_2 are the propagation delays for the two arms of the interferometer, and (ii) the phase difference between the signals reaching the two arms of the interferometer ($\delta\phi = 2\pi/\lambda(\Delta L_1 - \Delta L_2)$, where ΔL_1 and ΔL_2 are the excess path lengths for the two arms of the interferometer. Generally δT is small compared to the coherence bandwidth of the signal and can be ignored to first order, however $\delta\phi$ could be substantial.

In a homogeneous plane parallel ionosphere with refractive index μ (see Figure 16.1), we have from Snell's law $\mu\sin(z_0) = \sin(z)$. The observed geometric delay is $\tau_g = \mu D\sin(z_0)/c$, since the group velocity is c/μ. From Snell's law therefore, $\tau_g = D\sin(z)/c$, the same as would have been observed in the absence of the ionosphere. A homogeneous plane parallel ionosphere hence produces no net effect on the visibilities, even though the apparent position of the source has changed. In the case where the interferometer is located outside the slab, there is neither a change in the apparent position nor a change in the phase, as is obvious from the geometry. This entire analysis holds for a stratified plane parallel ionosphere (since it is true for every individual plane parallel layer). However, in the real case of a curved ionosphere, with a radial variation of electron density, then neither the change in the apparent position nor $\delta\phi$ are zero even outside the ionosphere. Effectively, the direction of arrival of the rays from the distant source appears to be different from the true direction of arrival (as illustrated in Figure 16.2) and unlike in the plane parallel case this is not exactly canceled out by the change in the refractive index. If $\Delta\theta$ is the difference between the true direction and apparent directions of arrival, then one can compute that

$$\Delta\theta = \frac{A\sin(z_0)}{r_0}\int_0^\infty \frac{\alpha^2\mu(h)dh}{(1-\alpha^2\sin^2(z_0))}. \tag{16.3.5}$$

where z_0 is the observed zenith angle, r_0 is the radius of the earth, h is the height above the earth's surface and, $\mu(h)$ is the refractive index at height h, and A is a constant. For baseline lengths typical of the GMRT, this value is the same for both arms of the baseline. If the baseline has UV co-ordinates (u,v), then the phase difference due to the apparent change in the source position is given by

$$\Delta\phi = 2\pi(u\Delta\theta_{EW} + v\Delta\theta_{NS}).$$

[1] to first order for $\nu >> \nu_p$, as can be easily verified.

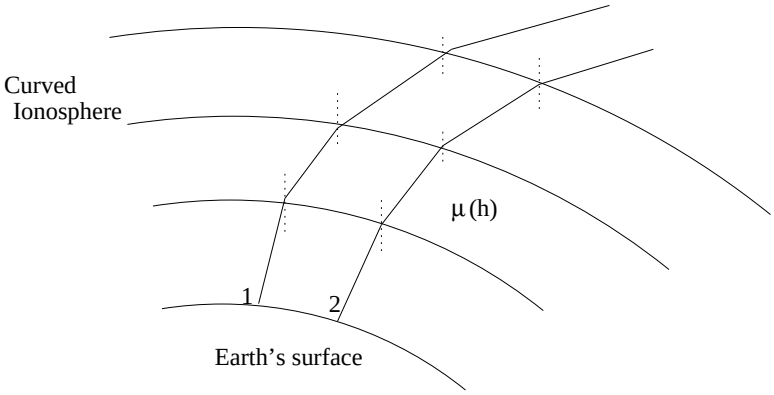


Figure 16.2: Propagation through a curved ionosphere

	Max. Val (Day)	Min Val (Night)	Freq. Dependence
TEC	5×10^{13} cm^{-2}	5×10^{12} cm^{-2}	-
Group Delay	12 μsec	1.2 μsec	ν^{-2}
Excess Path	3500 m	350 m	ν^{-2}
Phase Change	7500 rad	750 m	ν^{-2}
Phase Fluctuation	± 150 rad	± 15 rad	ν^{-2}
Mean Refraction	$6^{'}$	$0.6^{'}$	ν^{-2}
Faraday Rotation	15 cycles	1.5 cycles	ν^{-2}

Table 16.1: Typical numerical values of various ionospheric effects

Typical values for some of the ionospheric prorogation effects that we have been discussing are given in Table 16.1.

16.4 Propagation Through an Inhomogeneous Ionosphere

So far we have been dealing with an ionosphere, which, while not homogeneous, is still fairly simple in that the density fluctuations are smooth, slowly varying functions. Further, the ionospheric density was assumed to not vary with time. In reality, the earth's ionosphere shows density fluctuations on a large range of length and time scales. A density fluctuation of length scale l at a height h above the earth's surface corresponds to a fluctuation on an angular scale of l/h. For a typical length scale l of 10 km, at a height of 200 km, the corresponding angular scale is $\sim 3^o$. This means that the phase difference introduced by the ionosphere changes on an angular scale of 3^o. If this phase is to be calibrated out, then one would need to pick a calibrator that is within 3^o of the target source — for most sources it turns out that there is no suitable calibrator this close by. This problem gets increasingly worse as one goes to lower frequencies since the excess ionospheric phase increases as ν^{-2}. As discussed in Chapter 5 therefore, as long as the excess ionospheric phase is constant over the field of view, this phase can be lumped in with the electronic phase of receiver chain, and can be solved for using self-calibration.

However, for a given antenna, as one observes at lower and lower frequencies, the field

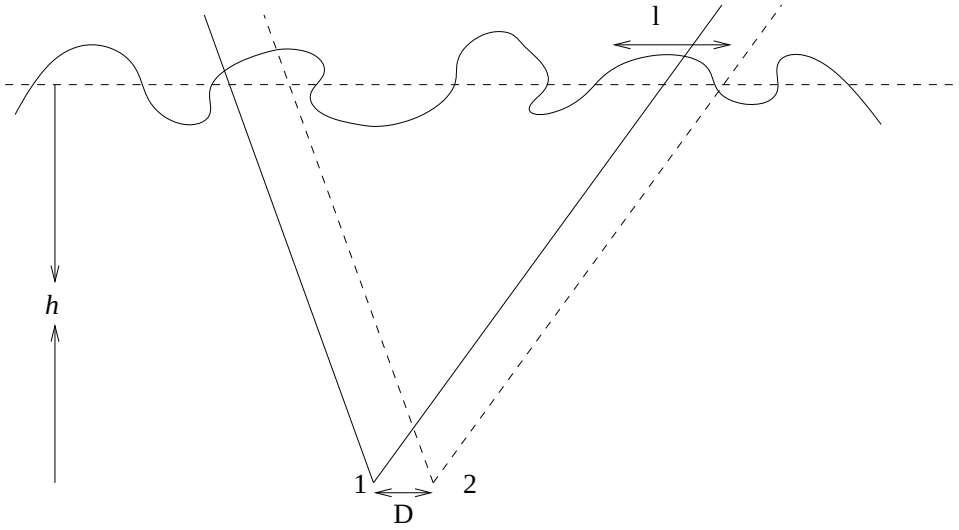


Figure 16.3: For short enough baselines, the isoplantic assumption holds even if the field of view is larger than the typical coherence length of the ionospheric irregularities. This is because both arms of the interferometer get essentially the same excess phase.

of view increases as ν^{-1}. Since the excess ionospheric phase is also increasing rapidly with decreasing frequency, one will soon hit a point where the assumption that the excess phase is constant over the field of view is a poor one. At this point the self-calibration algorithm is no longer applicable. Variations of the ionospheric phase over the field of view are referred to as "non isoplanaticity". As illustrated in Figure 16.3, when the baseline length is small compared to the typical length scale of ionospheric density fluctuations, even though the ionospheric phase is different for different sources in the field of view, the excess phase is nearly identical at both ends of the baseline. Since interferometers are sensitive only to phase differences between the two antennas, the isoplanatic assumption still holds. The non isoplanaticity problem hence arises only when the baselines as well as the field of view are sufficiently large. For the GMRT, isoplanaticity is often a poor assumption at frequencies of 325 MHz and lower.

16.5 Angular Broadening

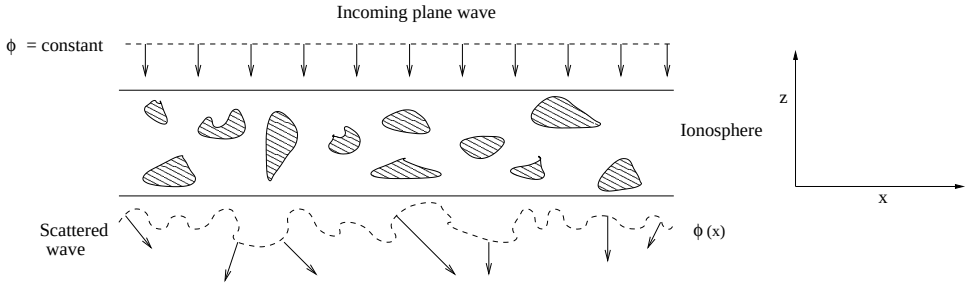


Figure 16.4: Angular broadening.

As discussed in the previous sections, the small scale fluctuations of electron density in

the ionosphere lead to an excess phase for a radio wave passing through it. This excess phase is given by

$$\begin{aligned}\phi(x) &= \frac{2\pi}{\lambda}\int \Delta\mu dz,\\ \phi(x) &= C\lambda\int \Delta n(x,z)dz,\end{aligned}$$

where $\Delta\mu$ is the change in refractive index due to the electron density fluctuation, C is a constant and $\Delta n(x,z)$ is the fluctuation in electron density at the point (x,z) and the integral is over the entire path traversed by the ray (see Figure 16.4).

If we assume that $\phi(x)$ is a zero mean Gaussian random process, with auto-correlation function given by $\phi_0^2\rho(r)$, where $\rho(r) = e^{-r^2/2a_\phi^2}$, then from the relation above for $\phi(x)$ we can determine that $\phi_0^2 \propto \lambda^2\Delta n^2 L$, where L is the total path length through the ionosphere[2]. Let us assume that a plane wavefront from an extremely distant point source is incident on the top of such an ionosphere. In the absence of the ionosphere the wave reaching the surface of the earth would also be a plane wave. For a plane wave the correlation function of the electric field (i.e. the visibility) is given by $\langle E_i(x)E\_i^*(x+r)\rangle = {E_i}^2$, i.e. a constant independent of r. On passage through the ionosphere however, different parts of the wave front acquire different phases, and hence the emergent wavefront is not plane. If $E(x)$ is the electric f ield at some point on the emergent wave, then we have $E(x) = E_i e^{-i\phi(x)}$. Since E_i is just a constant, the correlation function of the emergent electric field is

$$\langle E(x)E^*(x+r)\rangle = E_i^2\langle e^{-i(\phi(x)-\phi(x+r))}\rangle.$$

From our assumptions about the statistics of $\phi(x)$ this can be evaluated to give

$$\langle E(x)E^*(x+r)\rangle = E_i^2 e^{-2\phi_0^2(1-\rho(r))}. \tag{16.5.6}$$

If ϕ_0^2 is very large, then the exponent is falls rapidly to zero as $(1-\rho(r))$ increases (or equivalently when r increases). It is therefore adequate to evaluate it for small values of r, for which $\rho(r)$ can be Taylor expanded to give $\rho(r) \simeq 1 - 1/2r^2/a_\phi^2$. and we get

$$\langle E(x)E^*(x+r)\rangle = E_i^2 e^{-\phi_0^2\frac{r^2}{a_\phi^2}}.$$

The emergent electric field hence has a finite coherence length (while the coherence length of the incident plane wave was infinite). From the van Cittert-Zernike theorem this is equivalent to saying that the original unresolved point source has got blurred out to a source of finite size. This blurring out of point sources is called "angular broadening" or "scatter broadening". If we define $a = a_\phi/\phi_0$ then the visibilities have a Gaussian distribution given by e^{-ir^2/a^2}, meaning that the characteristic angular size θ_{scat} of the scatter broadened source is $\sim \lambda/a \propto \lambda^2\sqrt{\Delta n^2 L}$. θ_{scat} is called the "scattering angle".

On the other hand if ϕ_0^2 is small then the exponent in eqn 16.5.6 can be Taylor expanded to give

$$\begin{aligned}\langle E(x)E^*(x+r)\rangle &= E_i^2\big[1-2\phi_0^2(1-\rho(r))\big],\\ &= E_i^2\big[(1-2\phi_0^2)+2\phi_0^2 e^{\frac{-r^2}{2a_\phi^2}}\big].\end{aligned}$$

This corresponds to the visibilities from an unresolved core (of flux density $E_i^2\ (1-2\phi_0^2)$) surrounded by a weak halo.

[2]This follows from the equation for $\phi(x)$ if you also assume that $<\Delta n(x,z)\Delta n(x,z')> = \Delta n^2\delta(z,z')$.

16.6 Scintillation

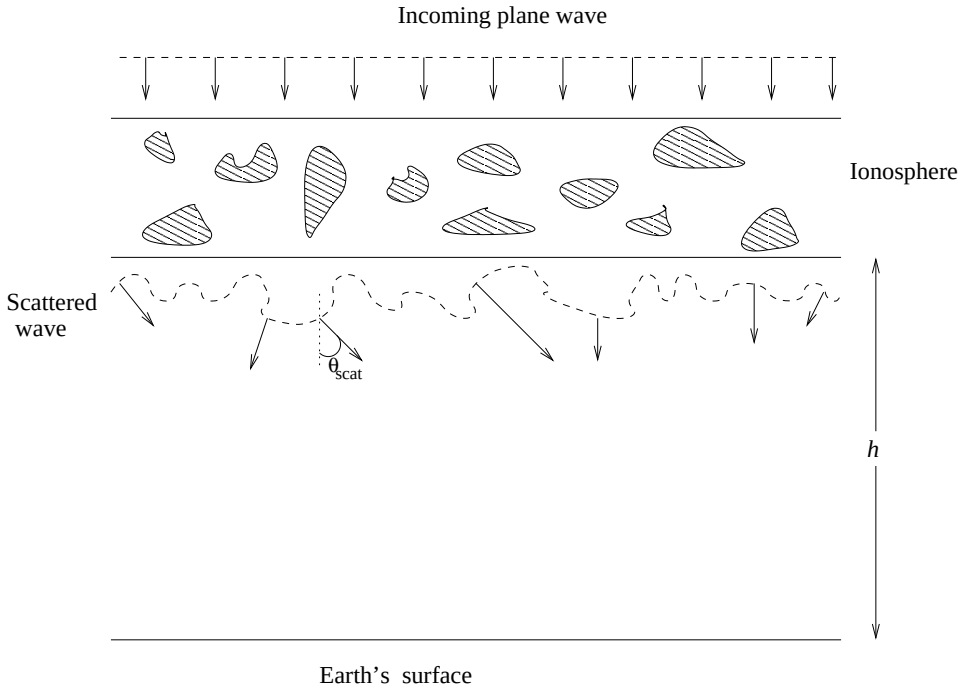


Figure 16.5: Scintillation due to the ionosphere

In the last section we dealt with an ionosphere which had random density fluctuations in it. In the model we assumed the density was assumed to vary randomly with position, but not with time. In the earth's ionosphere however, the density does vary with both position and time. Temporal variations arise both because of intrinsic variation as well as because of traveling disturbances in the ionosphere, because of which a given pattern of density fluctuations could travel across the line of sight.

This temporal variation of the density fluctuations means that the coherence function (even at some fixed separation on the surface of the earth) will vary with time. This phenomena is generically referred to as "scintillation". Depending on the typical scattering angle as well as the typical height of the scattering layer from the surface of the earth, the scintillation could be either "weak" or "strong".

As discussed in the previous section, rays on passing through an irregular ionosphere get scattered by a typical angle θ_{scat}. If the scattering occurs at a height h above the antennas, then as shown in Figure 16.5 these scattered rays have to traverse a further distance h before being detected. The transverse distance traveled by a scattered ray is $\sim h\theta_{scat}$. If this length is much less than the coherence length a, then the rays scattered by different irregularities in the scattering medium do not intersect before reaching the ground. The corresponding condition is that $h\theta_{scat} < a$, i.e. $h\theta_{scat} < \lambda/\theta_{scat}$ or $h\theta_{scat}^2 < \lambda$.

If this condition holds, then, at any instant of time, (as discussed in the previous section), what the observer sees is an undistorted image of the source, which is shifted in position due to refraction. As time passes, the density fluctuations change[3] and so

[3]but we assume that their statistics remain exactly the same, i.e. that they continue to be realization of a Gaussian random process with variance ϕ_0 and auto-correlation $\rho(r)$

the image appears to wander in the sky and in a long exposure image which averages many such wanderings, the source appears to have a scattered broadened size θ_{scat}. Provided that one can do self calibration on a time scale that is small compared to the time scale of the "image wander", this effect can be corrected for completely. On the other hand, when the $h\theta_{scat}^2 > \lambda$ the rays from different density fluctuations will intersect and interfere with one another. The observer sees more than one image, and because of the interference, the amplitude of the received signal fluctuates with time. This is called "amplitude" scintillation. Amplitude scintillation at low frequencies, particularly over the Indian subcontinent can be quite strong. The source flux could change be factors of 2 or more on very short timescales. This effect cannot be reliably modeled and removed from the data, and hence observations are effectively precluded during periods of strong amplitude scintillation.

16.7 Further Reading

1. Interferometry and Synthesis in Radio Astronomy; Thompson, A. Richard, Moran, James M., Swenson Jr., George W.; Wiley-Interscience Publication, 1986.

2. Synthesis Imaging In Radio Astronomy; Eds. Perley, Richard A., Schwab, Frederic R., and Bridle, Alan H.; ASP Conference Series, Vol 6.

Chapter 17

Pulsar Observations

Yashwant Gupta

17.1 Introduction

Amongst the various kinds of sources observed in Radio Astronomy, pulsars are perhaps the most unique kind, from many points of view. A pulsar is a neutron star – the ultra-dense core that remains after a massive star undergoes a supernova explosion – spinning at very rapid rates ranging from once in a few seconds to as much as $\sim$ 1000 times per second. A pulsar has a magnetosphere with a very high value of the magnetic field ($\sim 10^6 - 10^9$ Gauss). The emission mechanism (which is not understood yet) produces radio frequency radiation that comes out in two beams, one from each pole of the magnetosphere. These rotating beams of radiation are seen by us whenever they intersect our line of sight to the pulsar, much like a lighthouse on the sea-shore. Each rotation of the pulsar thus produces a narrow pulse of radiation that can be picked up by a radio telescope. Several properties of pulsars – such as their ultra-compact size, the occurrence of narrow duty cylce pulses with highly stable periods, intensity fluctuations on very short time scales and high degree of polarisation of the radiation – make for a set of observation and data analysis techniques that are very different from those used in radio interferometry. Here we take a look at these special techniques in some detail.

17.2 Requirements for Pulsar Observations

17.2.1 Phased Array Requirements

Like all radio sources, the sensitivity of pulsar observations benefits from the availability of a large collecting area. However, because of the compact nature of the source of radiation (typically a few hundred kilometers across), a pulsar is effectively a point source for the largest interferometer baselines on the Earth. Hence, there is not much to be learnt from making a map of a pulsar! This means that single dish observations are enough for pulsar work. However, since pulsars are relatively weaker sources (typical average flux densities ≤ 100 mJy, large collecting areas are very useful and hence array telescopes are used for this advantage. These array telescopes are not used in the interferometer mode, but in the phased array mode (see chapter 6). This means that much of the complicated hardware of the correlator required for measuring the visibilities on all baselines is not needed. In phased array mode, pulsar observations can be carried out in two different

ways : (i) incoherent phased array observations and (ii) coherent phased array observations. In the incoherent phased array mode, the signal from each antenna is put through a detector and the output from these is added to obtain the net signal. In coherent phased array mode, the voltage signal from each antenna is added and the summed output is put through a detector to obtain the final power signal. For an array of N antennas, the incoherently phased array gives a sensitivity of $\sqrt{N}$ times that of a single antenna, while the coherent array gives a sensitivity of N times that of a single antenna. The incoherent array has an effective beam that is same as that of a single antenna of the array, whereas the coherent array has a beam width that is much narrower than that of a single antenna, being $\sim \lambda/D$, where D is the largest spacing between antennas in the array. The coherent phased array mode is ideally suited for observations of known pulsars. The incoherent phased array mode is most useful for large scale pulsar search observations, where the aim is to cover a maximum area of the sky in a given time, at a given level of sensitivity. For a sparsely filled aperture array, incoherent phased array observations will certain be faster for such applications.

17.2.2 Spectral Resolution Requirements

Again like all radio sources, pulsar observations also benefit from large bandwidths of observation. However, unlike any other kind of continuum radio source, pulsar observations can not often combine the data from across a large bandwidth in a single detector. This is mainly because of the smearing of the pulses produced by differential dispersion delay of frequencies across the band, due to propagation of the pulsar signal through the interstellar medium. This is explained in some detail in section 4 below. In the simplest technique for reducing the effect of dispersion delay smearing, the pulsar signal is processed in a multichannel receiver where the observing band is broken up into narrower frequency channels. The signal in each channel is detected and acquired separately. This requirement of narrower frequency channels across the observing band makes a pulsar receiver similar to a spectral line receiver, though for entirely different reasons.

17.2.3 Requirements for Time Resolution and Accurate Time Keeping

Unlike other radio sources which are taken to be statistically constant in their strength as a function of time, pulsar signals are intrinsically periodic signals. The pulses have periods ranging from a few seconds for the slowest pulsars to about a millisecond for the fastest pulsars known. Further, the pulses have a very small duty cycle, with typical pulse widths of the order of $5 - 10\%$ of the period. Thus typical pulse widths range from a few tens of milliseconds down to a fraction of a millisecond. Study of such pulsar signals clearly requires the final data to have time resolutions ranging from $\sim$ milliseconds to $\sim$ microseconds. Pulsar observations thus require very fast sampling times for the data. This leads to a substantial increase in the speed (and therefore complexity) of the back-end designed for pulsar observations and also in the speed of the data acquisition system and off-line computing capabilities. Also, the value of the sampling interval needs to be known quite accurately in order to preserve the pulse phase coherence over a long stretch of pulsar data spanning many periods.

The other property of the time variation of pulsar signals is that the rotation rate of pulsars is very accurate. This means that if the time of arrival of the Nth and (N+1)th pulses is known, the arrival time for the (N+M)th pulse can be predicted very accurately. Further, slow variations of the pulsar period (for example due to rotational slow down

of the pulsar) can be studied if the absolute time of arrival of the pulses can be measured sufficiently accurately. This requires the availability of a very precise clock at the observatory, such as that provided by a GPS receiver (see section 17.7 for more details).

17.2.4 Requirements for Polarimetry

Radiation from pulsars has been shown to be highly polarised. The linear polarisation can at times reach close to 100%. Significant amounts of circular polarisation is also seen frequently. The study of these polarisation characteristics is very important for understanding the emission mechanism of pulsars. Hence pulsar studies often require that the telescope support full polarisation observations that finally yield the four Stokes parameters, as a function of time and frequency. Remember that each of these polarisation parameters needs to satisfy all the time and frequency resolution criteria outlined above, leading to a four fold increase in hardware complexity and data flow rate over simple total power observations.

17.2.5 Flux Calibration Requirements

The intensity of individual pulses varies randomly over various time scales. On the shortest time scale, pulse to pulse intensity fluctuations are thought to be due to intrinsic processes in the pulsar magnetosphere. Longer time scale fluctuations in the mean pulsar flux are produced by propagation processes in the ionised plasma of the interstellar medium (ISM). Furthermore, some of these intensity fluctuations can be uncorrelated over large frequency intervals. Thus for purposes of estimating the pulsar flux (including estimates of the spectral index) and for studying the variations in the pulsar flux to understand properties of the ISM, pulsar observations need to be calibrated with known sources of power. This can be done by using either calibrated noise sources that can be switched into the signal path or known calibration sources in the sky.

17.3 Basic Block Diagram of a Pulsar Receiver

Incorporating the above requirements into a realistic set-up for pulsar observations leads to the following block level diagram for pulsar observations (see Fig 17.1). In a modern radio telescope, most of the processing of the signals is carried out in the digital domain, after down conversion to a baseband signal (of bandwidth B). Hence the first block is an analog to digital convertor (ADC), which is run on an accurate and controlled sampling clock. For multi-element or array telescopes, the signals from the different elements need to be phased. This involves proper adjustments of amplitude, delay and phase of the signals (see chapter 6). The output of this block is the phased array signal which goes to the 'Spectral Resolution Block'. For a single dish telescope, the signal comes directly from the sampler to this block. This block produces the multiple narrow-band channels from the single broad-band data. This can be achieved using a filter bank or a FFT spectrometer or an auto-correlation spectrometer. The output is a baseband voltage signal for each of N_{ch} frequency channels, sampled at the Nyquist rate. For a multi-element telescope, the location of this block and the Phased Array block can be interchanged, in part or in whole. For example, at the GMRT, the integer sample delay correction is done before the FFT; the fractional sample delay correction and the phase correction is done in the last stage of the FFT and the addition of the signals is done in a separate block located after the FFT. Note that for incoherent phased array operation to be possible, the addition of the signals MUST be after the spectral resolution block,

BLOCK LEVEL DIAGRAM FOR A PULSAR RECEIVER

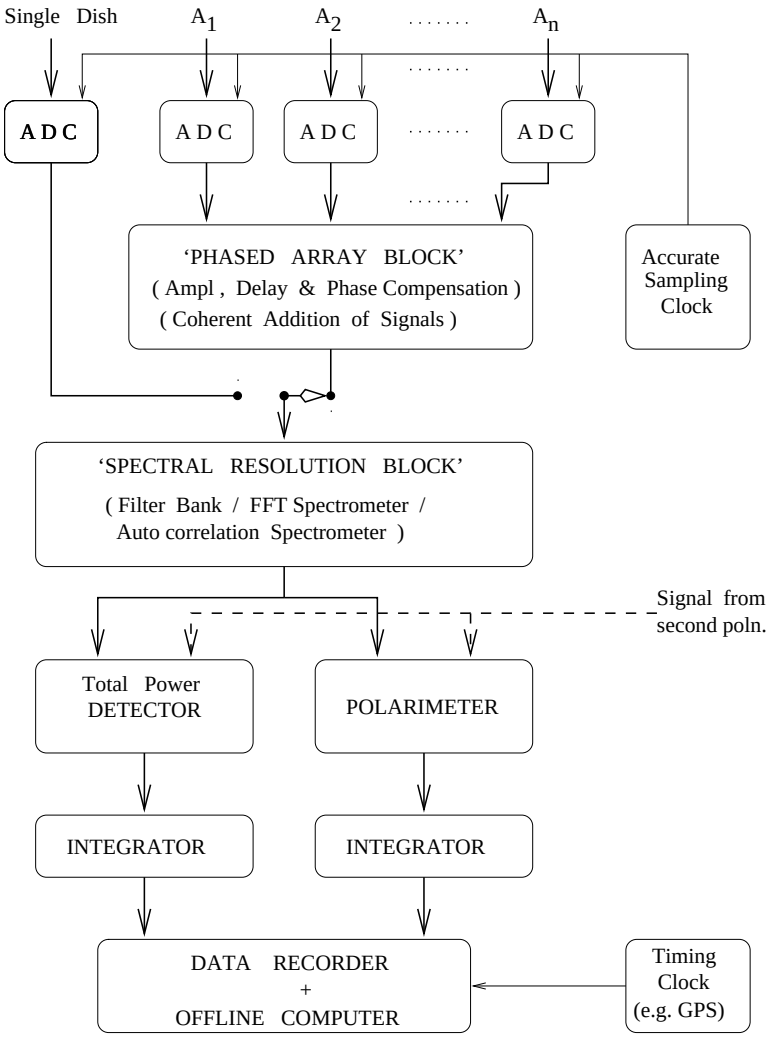


Figure 17.1: Block diagram of a typical pulsar receiver

because the square law detection has to be carried out before the incoherent addition can be done.

The second orthogonal polarisation from the telescope is also processed similarly till the output from the spectral resolution block. These outputs can then be given to two different kinds of processors. The first is a total power adder that simply adds the powers of the signals in the two polarisations to give a measure of the total intensity from the telescope as a function of time and freqency. The second is a polarimeter that takes the voltage signals from the two polarisations and produces the four Stokes parameters, as a function of time and frequency. The data from the incoherent phased array, for example, can only be put through the total power path. The outputs from these two processors are then put through an adder that integrates the data to the required time constant, τ_s. The final output going to the recorder then is either one (total intensity mode) or four

(polarimetry mode) signals each containing N_{ch} frequency channels coming at the rate of $1/\tau_s$ samples per second. The net data rate into the recorder is then N_{ch}/τ_s samples per second for the total intensity mode and four times as much for the polarimetry mode. As an example, if data from 256 spectral channels is being acquires with a time constant of 0.25 millisec, the data rate is 1 mega samples per second for the total intensity mode. If one sample is stored as a two byte word, we can see that a storage space of 1 gigabyte would get filled with about 2 minutes of data! In cases where the data rate going into the recorder in the above set-up is difficult to handle for storage or off-line processing, special purpose hardware to do some of the processing on-line can also be used. Typical examples of such processing would be on-line dedispersion, on-line folding at the pulsar period and on-line gating of the data (to pass on only some region of each pulsar period that is around the on-pulse region). Each of these techniques reduces the net data rate so that it can be comfortably acquired and further processed off-line. The choice of the processing technique depends on the scientific goals of the observations.

17.4 Dispersion and Techniques for its Correction

As mentioned earlier, propagation of pulsar signals through the tenuous plasma of the ISM produces dispersion of the pulses. This is because the speed of propagation through a plasma varies with the frequency of the wave (see chapter 16). Low frequency waves travel progressively slowly, with a cut-off in propagation at the plasma frequency. At high frequencies, the velocity reaches the velocity of light asymptotically. The difference in travel time between two radio frequencies f_1 and f_2 is given by

$$t_d \;=\; K\,DM\,\left(\frac{1}{f_1^2}\;-\;\frac{1}{f_2^2}\right) \quad , \tag{17.4.1}$$

where $DM \;=\; \int n_e\,dl$ is the dispersion measure of the pulsar, usually measured in the somewhat unusual units of $pc\,cm^{-3}$, and $K \;=\; 4.149 \times 10^6$ is a constant. In this equation, t_d is in units of millisec and f_1, f_2 are in units of MHz. For the typical ISM, a path length of 1 kiloparsec amounts to a DM of about $30\,pc\,cm^{-3}$. Equation (1) can be used to derive the following approximate relationship for the dispersion smear time for a bandwidth B centred at a frequency of observation f_0, for the case when $B \ll f_0$

$$\tau_{disp} \;\simeq\; \left(\frac{202}{f_0}\right)^3 DM\,B \quad , \tag{17.4.2}$$

where τ_{disp} is in millisec, f_0 and B are in MHz and DM is in the units given earlier. Interstellar dispersion degrades the effective time resolution of pulsar data due to smearing, and this effect becomes worse with decreasing frequency of observation. For example, the dispersion smear time is about 0.25 millisec per MHz of bandwidth per unit DM at an observing frequency of 325 MHz. This means that a pulse of 25 millisec width would be broadened to twice its true width when observed with a bandwidth of 10 MHz, for a DM of $10\,pc\,cm^{-3}$. Even worse, signal from a pulsar of period 25 millisec would be completely smeared out and not be visible as individual pulses. Thus it is important to reduce the effect of interstellar dispersion in pulsar data. This is called dedispersion.

There are two main methods used for dedispersion – incoherent dedispersion and coherent dedispersion. In incoherent dedisperion, which is a post-detection technique, the total observing band (of bandwidth B) is split into N_{ch} channels and the pulsar signal is acquired and detected in each of these. The dispersion smearing in each channel is less than the total smearing across the whole band, by a factor of N_{ch}. The detected signal

from each channel is delayed by the appropriate amount so that the dispersion delay between the centers of the channels is compensated. These differentially delayed data trains from the N_{ch} channels are added to obtain a final signal that has the dispersion smearing time commensurate with a bandwidth of B/N_{ch}, thereby reducing the effect of dispersion. In practical realisations of this scheme, the splitting of the band into narrow channels is usually carried out on-line in dedicated hardware (as described in section 17.3) while the process of delaying and adding the detected signals from the channels can be done on-line using special purpose hardware or can be carried out off-line on the recorded, multi-channel data. In this scheme, the final time resolution obtained for a given pulsar observation is limited by the number of frequency channels that the band is split into.

In coherent dedispersion, one attempts to correct for interstellar dispersion in a pulsar signal of bandwidth B before the signal goes through a detector, i.e. when it is still a voltage signal. It is based on the fact that the effect of interstellar scintillation on the electromagnetic signal from the pulsar can be modelled as a linear filtering operation. This means that, if the response of the filter is known, the original signal can be deconvolved from the received voltage signal by an inverse filtering operation. The time resolution achievable in this scheme is $1/B$ – the maximum possible for a signal of bandwidth B. Thus coherent dedispersion gives a better time resolution than incoherent dedispersion, for the same bandwidth of observation. It is the preferred scheme when very high time resolution studies are required – as in studies of profiles of millisecond pulsars and microstructure studies of slow pulsars. The main drawback of coherent dedispersion is that practical realisations of this scheme are not easy as it is a highly compute intensive operation. This is because the duration of the impulse response of the dedispersion filter (equal to the dispersion smear time across the bandwidth) can be quite long. To reduce the computational load, the deconvolution operation of the filtering is carried out in the Fourier domain, rather than in the time domain. Nevertheless, real time realisations of this scheme are limited in their bandwidth handling capability. Most coherent dedispersion schemes are implemented as off-line schemes where the final baseband signal from the telescope is recorded on high speed recorders and analysed using fast computers.

17.5 Pulse Studies

Pulsar pulse studies encompass a broad set of topics ranging from the study of the average properties of pulsar profiles to the study of microscopic pheonomena in individual pulses. Though individual pulses from a pulsar show tremendous variations in properties such as shape, width, amplitude and polarisation, it is found that when a few thousand pulses (typically) are accumulated synchronously with the pulsar period, the resulting average profile shows a steady and constant form which can be considered to be a signature of that pulsar. Such an average profile typically exhibits one or more well defined regions of emission within the profile window. These are usually referred to as emission components and they can be partially or completed separated in pulse longitude. Similarly, the average polarisation properties also show a well defined signature in terms of the variations (across the profile window) of the amplitudes of linear and circular polarisation, as well as the angle of the linear polarisation vector. The average profile however does change with observing frequency for a given pulsar, with the typical signature being that profiles become wider at lower frequencies. Average pulse profile studies are important for characterising the overall properties of a pulsar.

To obtain accurate average pulse profiles, one needs to observe the pulsar for a long enough stretch so that (i) the profile converges to a stable form and (ii) there is enough signal to noise. The time resolution should be enough to resolve the features of interest in

the profile (typically 1% to 0.1% of the pulse period). Since the average profile is obtained by synchronous accumulation at the pulsar period (this is called 'folding' in pulsar jargon), the period and the sampling interval need to be known with sufficient accuracy to avoid any distortions due to smearing effects. It is easy to show that the fractional error in the period and the resultant fractional error in phase are related by

$$\frac{\Delta P}{P} = \frac{1}{N_p}\frac{\Delta\phi}{\phi} \quad , \tag{17.5.3}$$

where N_p is the number of pulses used in the folding. As an example, if the distortions due to phase error are to be kept under one part in a thousand and $N_p = 1000$, then the period needs to be known to better than 1 part in a million.

Let us know look at the signal to noise ratio (SNR) for an average profile observation. For a pulsar of period P and pulse width W having a time average flux S_{av}, observed with a telescope of effective aperture A_{eff} and system noise temperature T_{sys}, using a bandwidth B and time constant τ_s, the signal to noise ratio at a point on a profile obtained from N_p pulses is given by

$$SNR_{avg} = \frac{S_{av}\,A_{eff}}{k\,T_{sys}}\,\frac{P}{W}\,\sqrt{B\tau_s}\,\sqrt{N_p} \quad . \tag{17.5.4}$$

Here the P/W term is to convert the time average flux to on-pulse flux and the $\sqrt{N_p}$ term accounts for the SNR improvement due to addition of N_p pulses. The other terms are as for normal SNR calculations for continuum sources.

When single pulses from a pulsar are examined in detail, it is seen that the radiation in each pulse does not always occur all over the average profile profile window. Usually, the signal is found located sporadically at different longitudes in the profile window. These intensity variations are called sub-pulses and they have a typcial width that is less than the width of the average profile. For some pulsars, sub-pulses in successive pulses don't always occur randomly in the profile window; they are found to move systematically in longitude from one pulse to the next. These are called drifting sub-pulses and are thought be one of the intriguing features of the emission mechanism. For some pulsars, there are times when there is practically no radiation seen in the entire profile window for one or more successive pulses. This phenomenon is called nulling and is another of the unexplained mysteries of pulsar radiation. Polarisation properties of sub-pulses also show significant deviations from the overall polarisation properties of the average profile. Studies of sub-pulses require time resolutions that are 0.1% of the pulse period, or better.

When single pulses are observed with still further time resolution, it is found that narrow bursts of emission are also seen with time scales much shorter than sub-pulse widths. This is called microstructure and the time scales go down to microseconds and less. Seeing pulsar microstructure almost always requires the use of coherent dedispersion techniques to achieve the desired time resolution. It is clear from the above that pulsar intensities show fluctuations at various time scales within a pulse period. A useful analysis technique that separates out the various time scales is the intensity correlation function.

It is worth pointing out that single pulse observations are the worst affected among all kinds of pulsar studies, from the point of view of signal to noise ratio. This is simply because the $\sqrt{N_p}$ advantage in equation (3) is not available. Also, as τ_s is reduced for higher time resolution studies, the SNR decreases further. Hence such studies need the largest collecting area telescopes and can often be done on only the strongest pulsars.

17.6 Interstellar Scintillation Studies

The propagation of pulsar signals through the interstellar medium of the Galaxy modifies the properties of the received radiation in several ways. A study of these effects can give useful information about the interstellar medium. One of these effects that has already been looked at is interstellar dispersion. It gives us information about the mean electron density of the interstellar plasma.

Another effect that is significant in pulsar observations is interstellar scintillations. It is caused by scattering of the radiation due to random fluctuations of electron density in the interstellar plasma. It produces the following effects (not all of which are easily observable!): (i) angular broadening of the source, as scattered radiation now arrives from a range of angles around the direction to the pulsar; (ii) temporal pulse broadening due to the delayed arrival of scattered radiation; (iii) random fluctuations of pulsar intensity as a function of time and frequency due to interference effects between radiation arriving from different directions. All these effects increase in strength with decreasing frequency and with increasing length of plasma between source and observer. A detailed study of interstellar scintillation effects in pulsar signals can be used to obtain valuable constraints on the extent and location of scattering plasma in the interstellar medium, as well as on the spatial power spectrum of electron density fluctuations in the medium.

Of the three effects of scintillations described above, the random fluctuations of intensity are the most easily observable and form the best probes of the phenomenon. They are readily seen in pulsar dynamic spectra which are records of the on-pulse intensity as a function of time and frequency. A single time sample in the dynamic spectra is obtained by accumulating the total energy under the pulse window for a given number of pulses, for each of N_{ch} channels. These random intensity fluctuations have typical decorrelation scales in time and frequency, which are estimated by performing a two dimensional autocorrelation of the dynamic spectra data. These decorrelation widths are of the order of a few minutes and hundreds of kHz, respectively, at metre wavelengths. This means that typical observations have to be carried out with time and frequency resolutions of the order of tens of seconds and tens of kHz in order to observe the scintillations. This requirement becomes more stringent at lower frequencies and for more distant pulsars (which are more strongly scattered). Also, the observations need to span enough number of these random scintillations in order to obtain statistically relaible values for the two decorrelation widths. This usually requires observing durations of an hour or so with bandwidths of a few MHz.

Due to the effect of large scale electron density fluctuations in the interstellar medium, the values of the decorrelation widths and the mean pulsar flux, fluctuate with time. A study of this phenomenon (called refractive scintillations) requires regular monitoring of pulsar dynamic spectra at different epochs, typically a few days apart and spanning several weeks to months. Such data can also be used to estimate the mean transverse speeds of pulsars.

17.7 Pulsar Timing Studies

Pulsar timing studies involve accurate measurements of the time of arrival of the pulses, followed by appropriate modelling of the observed arrival times to study and understand various phenomena that can effect the arrival times.

The first step of accurate estimation of arrival times is achieved as follows. First, at each epoch of observation, data from the pulsar is acquired with sufficient resolution in time and frequency and over a long enough stretch so that a reliable estimate of

the average profile can be obtained. The effective time resolution should be about one-thousandth of the period. Second, the absolute time for at least one well defined point in the observation interval is measured with the best possible accuracy. Traditionally, atomic clocks have been used for this purpose. With the advent of the Global Positioning System (GPS), absolute time (UTC) tagging with an accuracy of $\sim$ 100 nanosec is possible using commercially available GPS receivers. Third, the fractional phase offset with respect to a reference epoch is calculated for the data at each epoch. This is generally best achieved by cross-correlating the average profile at the epoch with a template profile and estimating the shift of the peak of the cross-correlation function. This shift, in units of time, is added to the arrival time measurement to reference the arrival times to the same phase of the pulse. Fourth, the arrival times measured at the observatory on the Earth are referred to a standard inertial point, which is taken as the barycenter of the solar system. These corrections include effects due to the rotation and revolution of the Earth, the effect of the Earth-Moon system on the position of the Earth and the effect of all the planets in the Solar System. Relativistic corrections for the clock on the Earth are also included, as are corrections for dispersion delay at the doppler corrected frequency of observation. Last, a pulse number, relative to the pulse at the reference epoch, is attached to the arrival time for each epoch. This can be a tricky affair, since to start with the pulsar period may not be known accurately, and it is possible to err in integer number of pulses when computing the pulse number. To avoid this danger, a boot-strapping technique is used where the initial epochs of observations are close enough so that, given the accuracy of the period, the phase error can not exceed one cycle between two successive epochs. As the period gets determined with better accuracy by modelling the initial epochs, the spacing between successive epochs can be increased. The net result of the above exercise is a series of data pairs containing time of arrival and pulse number, both relative to the same starting point.

The second step in the analysis is the modelling of the above data points. This is usually done by expressing the pulse phase at any given time in terms of the pulsar rotation frequency and its derivatives as follows

$$\phi_i \;=\; \phi_0 + \nu_0 t_i + \dot{\nu_0} {t_i}^2/2 + \ldots \quad , \tag{17.7.5}$$

where $\nu_0 = 1/P$. Least squared fits for ϕ_0, ν_0, $\dot{\nu_0}$ etc., can be obtained from such a model. In addition, by examining the residuals between the model and the data, other parameters that effect the pulsar timing can be estimated. These include errors in the positional estimate of the pulsar, its proper motion, perturbations to the pulsar's motion due to the presence of companions, sudden changes in the pulsar's rotation rate etc. In fact, good quality timing observations can be used to extract a wealth of information, including stability of pulsars vis-a-vis the best terrestial clocks!

17.8 Pulsar Search

At the end, we come to the observation and analysis techniques used for discovering new pulsars. Pulsar searches fall into one of two broad categories : targeted and untargeted searches. In an untargeted search (or survey) for pulsars, the idea is to uniformly cover a large area of the sky with a desired sensitivity in flux level. In targeted searches, one is searching a limited area of the sky where there is a higher than normal possibility of finding a pulsar (for example, the region in and around a supernova remnant or a steep spectrum point source identified in mapping studies). Here some of the parameters of the search can be tailored to suit the a priori knowledge about the search region.

For a pulsar survey, the choice of (i) the range of directions to search in, (ii) the frequency of observations, (iii) the bandwidth and number of spectral channels, (iv) the sampling interval and (v) the duration of the observations are some of the critical items that need to be chosen carefully. The choice of these parameters is interlinked in many cases.

Analysis of pulsar search data is an extremely compute intensive task. For each position in the sky for which data is recorded, the analysis technique needs to search for the presence of a periodic signal in the presence of system noise. However, from the discussion in section 3, it is clear that if appropriate dispersion correction is not done, the sensitivity to the presence of a periodic signal can be reduced significantly. Since a pulsar can be located at any distance (and hence DM) along a given direction in the sky, the search has to be carried out in (at least) two dimensions : DM and period. For this, the data is dedispersed for different trial dispersion measures. For each choice of DM, the dedispersed data is search for a periodic signal.

To reduce the computational load for search data analysis, several optimised algorithms are used. For example, when dedispersing for a range of DM values, it is possible to use the results from the computations for some DM values to compute part of the results for some other DM values. This saves a lot of redundant calculations. This method, known as Taylor's Dedispersion Algorithm, is used quite often. Similarly, there are optimised techniques for searching for periodic signals in the presence of noise. The simplest method is to fold the dedispersed data for each choice of possible period and examine the resulting profile for the presence of a significant peak that is well above the noise level. Once again, computations done for folding at a given period can be used for folding at other periods. This redundancy is exploited by the Fast Folding Algorithm. A signal containing a periodic train of pulses gives a well defined signature in the Fourier domain – its spectrum consists of peaks at the frequency corresponding to the periodicity, and harmonics thereof. It can be shown that it is possible to detect the periodic signal by searching for harmonically related peaks in the spectral domain. It turns out that it is more economical to implement the FFT followed by harmonic search technique compared to the folding search techniques.

Additional complications are introduced in the search algorithm when one allows the parameter space to cover pulsars in binary orbits as the period can actually change during the interval of observation. Special processing techniques are needed to handle such requirements.

17.9 Further Reading

- Hankin, T.H. & Rickett, B.J. "Pulsar Signal Processing", McGraw-Hill Book Company, New York, USA, 1988
- Lyne A.G. & Smith, F.G., "Pulsar Astronomy", Cambridge University Press, Cambridge, UK, 1998
- Manchester, R.N. & Taylor, J.H. "Pulsars", W.H. Freeman & Co, San Fransisco, USA, 1977

Chapter 18

An Overview of the GMRT

Jayaram N. Chengalur

18.1 Introduction

The Giant Metrewave Radio Telescope (GMRT) consists of an array of 30 antennas. Each antenna is 45 m in diameter, and has been designed to operate at a range of frequencies from 50 MHz to 1450 MHz. The antennas have been constructed using a novel technique (nicknamed SMART) and their reflecting surface consists of panels of wire mesh. These panels are attached to rope trusses, and by appropriate tensioning of the wires used for attachment the desired parabolic shape is achieved. This design has very low wind loading, as well as a very low total weight for each antenna. Consequently it was possible to build the entire array very economically. In this chapter I give a very brief overview of the GMRT. Subsequent chapters discuss in detail each of the major subsystems of the GMRT.

18.2 Array Configuration

The GMRT has a hybrid configuration, (see Figure 18.1) with 14 of its antennas randomly distributed in a central region (~ 1 km across), called the central square. The distribution of antennas in the central square was deliberately "randomized" to avoid grating lobes. The antennas in the central square are labeled as Cnn, with nn going from 00 to 14 (i.e. C00, C01,....C14)[1]. The remaining antennas are distributed in a roughly Y shaped configuration, with the length of each arm of the Y being ~ 14 km. The maximum baseline length between the extreme arm antennas is ~ 25 km. The arms are called the "East" "West" and "South" arms and the antennas in these arms are labeled E01..E06, W01..W06 and S01..S06 for the east, west and south arm respectively.

The central square antennas provide a large number of relatively short baselines. This is very useful for imaging large extended sources, whose visibilities are concentrated near the origin of the UV plane. The arm antennas on the other hand are useful in imaging small sources, where high angular resolution is essential. A single GMRT observation hence yields information on a variety of angular scales.

[1]The array was originally meant to have 34 antennas, but because of escalating costs, was finally constructed with 30. Consequently some antenna stations do not actually have any antennas in them, resulting in "missing" numbers (C07, E01, S05) in the numbering sequence.

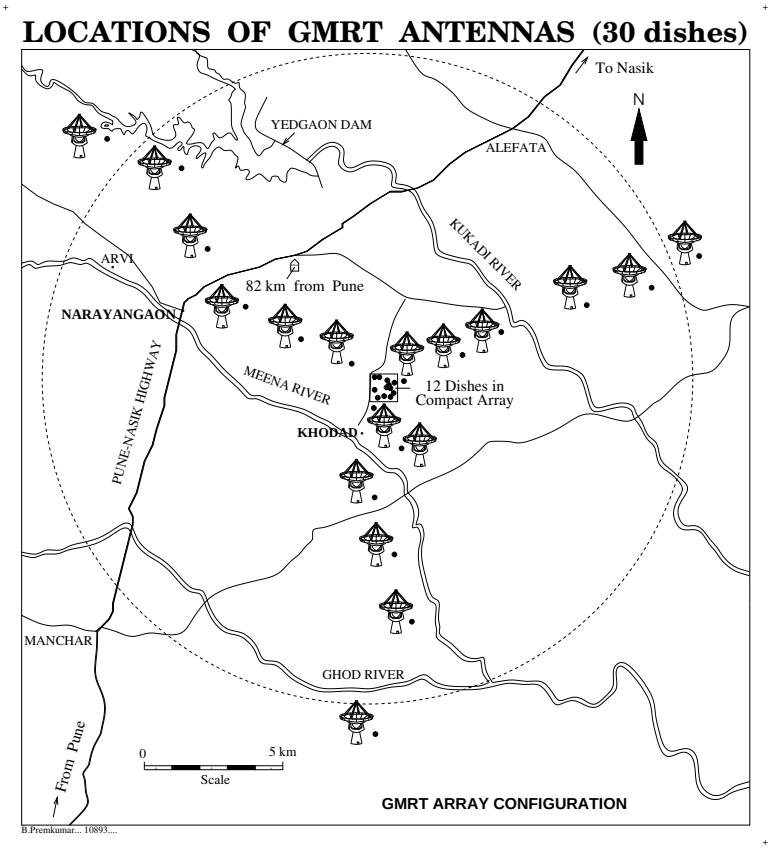


Figure 18.1: GMRT array configuration.

18.3 Receiver System

The GMRT currently operates at 5 different frequencies ranging from 150 MHz to 1420 MHz. Some antennas have been equipped with receivers which work up to 1750 MHz. Above this frequency range however, the antenna performance degrades rapidly both because the reflectivity of the mesh falls and also because of the rapidly increasing aperture phase errors because of the deviations of the plane mesh facets from a true parabola. A 50 MHz receiver system is also planned. Table 18.1 gives the relevant system parameters at the nominal center frequency of the different operating frequencies of the GMRT.

The GMRT feeds, (except for the 1420 feed), are circularly polarized. The circular polarization is achieved by means of a polarization hybrid inserted between the feeds and the RF amplifiers. No polarization hybrid was inserted for the 1420 MHz feed, in order to keep the system temperature low. None of the receivers are cooled, i.e. they all operate at the ambient temperature. The feeds are mounted on four faces of a feed turret placed at the focus of the antenna. The feed turret can be rotated to make any given feed point to the vertex of the antenna. The feed on one face of the turret is a dual frequency feed, i.e. it works at both 233 MHz as well as 610 MHz.

After the first RF amplifier, the signals from all the feeds are fed to a common second stage amplifier (this amplifier has an input select switch allowing the user to choose which RF amplifier's signal is to be selected), and then converted to IF. Each polarization is converted to a different IF frequency, and then fed to a laser-diode. The optical sig-

System Properties	in MHz					
	50	153	233	327	610	1420
Primary beam (degree)		3.8	2.5	1.8	0.9	0.4×(1400/f)
Synthesized beam						
Full array (arcsec)		20	13	09	05	02
Central array (arcmin)		7.0	4.5	3.2	1.7	0.7
System temperature (K)						
(1) $T_{\rm receiver}$		144	55	50	60	40
(including cable losses)						
(2) $T_{\rm ground}$ = $T_{\rm mesh}$ + $T_{\rm spillover}$		30	23	18	22	32
(3) $T_{\rm sky}$		308	99	40	10	4
Total $T_{\rm sys}$		482	177	108	92	76
= $T_{\rm sky}$ + $T_{\rm receiver}$ + $T_{\rm ground}$						
Gain of an antenna (K/Jy)		0.33	0.33	0.32	0.32	0.22
RMS noise in image* (μJy)		46	17	10	09	13

*For assumed bandwidth of 16 MHz, integration of 10 hours and natural weighting (theoretical).

Table 18.1: System parameters of the GMRT

nals generated by the laser-diode are transmitted to a central electronics building (CEB) by fiber optic cables. At the central electronic building, they are converted back into electrical signals by a photo-diode, converted to baseband frequency by another set of mixers, and then fed into a suitable digital backend. Control and telemetry signals are also transported to and from the antenna by on the fiber-optic communication system. Each antenna has two separate fibers for the uplink and downlink.

18.4 Digital Backends

There are a variety of digital backends available at the GMRT. The principle backend used for interferometric observations is a 32 MHz wide FX correlator. The FX correlator produces a maximum of 256 spectral channels for each of two polarizations for each baseline. The integration time can be as short as 128 ms, although in practice 2 sec is generally the shortest integration time that is used. The FX correlator itself consists of two 16 MHz wide blocks, which are run in parallel to provide a total instantaneous observing bandwidth of 32 MHz. For spectral line observations, where fine resolution may be necessary, the total bandwidth can be selected to be less than 32 MHz. The available bandwidths range from 32 MHz to 64 kHz in steps of 2. The maximum number of spectral channels however remains fixed at 256, regardless of the total observing bandwidth. The GMRT correlator can measure all four Stokes parameters, however this mode has not yet been enabled. In the full polar mode, the maximum number of spectral channels available is 128. Dual frequency observations are also possible at 233 and 610 MHz, however in this case, only one polarization can be measured at each frequency. The array can be split into sub-arrays, each of which can have its own frequency settings and target source. The correlator is controlled using a distributed control system, and the data acquisition is also distributed. The correlator output, i.e. the raw visibilities are recorded in a GMRT specific format, called the “LTA” format. Programmes are available for the inspection, display and calibration of LTA files, as well as for the conversion of

LTA files to FITS.

The first block of the GMRT pulsar receiver is the GMRT Array Combiner (GAC) which can combine the signals from the user-selected antennas (up to a maximum of 30) for both incoherent and coherent array operations. The input signals to the GAC are the outputs of the Fourier Transform stage of the GMRT correlator, consisting of 256 spectral channels across the bandwidth being used, for each of the two polarization from each antenna. The GAC gives independent outputs for the incoherent and coherent array summed signals, for each of two polarizations. For nominal, full bandwidth mode of operation, the sampling interval at the output of the GAC is 16μsec.

Different back-end systems are attached to the GAC for processing the incoherent and coherent array outputs. The incoherent array DSP processor takes the corresponding GAC output signals and can integrate the data to a desired sampling rate (in powers of 2 times 16 microsec). It gives the option of acquiring either one of the polarizations or the sum of both. It can also collapse adjacent frequency channels, giving a slower net data rate at the cost of reduced spectral resolution. The data is recorded on the disk of the main computer system.

The coherent array DSP processor takes the dual polarization, coherent (voltage sum) output of the GAC and can produce an output which gives 4 terms – the intensities for each polarization and the real and imaginary parts of the cross product – from which the complete Stokes parameters can be reconstructed. This hardware can be programmed to give a sub-set of the total intensity terms for each polarization or the sum of these two. The minimum sampling interval for this data is 32 microsec, as two adjacent time samples are added in the hardware. Further preintegration (in powers of 2) can be programmed for this receiver. The final data is recorded on the disk of the main computer system.

There is another independent full polarimetric back-end system that is attached to the GAC. This receiver produces the final Stokes parameters, I,Q,U & V. However, due to a limitation of the final output data rate from this system, it it can not dump full spectral resolution data at fast sampling rates. Hence, for pulsar mode observations the user needs to opt for online dedispersion or gating or folding before recording the data (there is also a online spectral averaging facility for non-pulsar mode observations).

In addition, there is a search preprocessor back-end attached to the incoherent array output of the GAC. This unit gives 1-bit data, after subtracting the running mean, for each of the 256 spectral channels. Either one of the polarizations or the sum of both can be obtained.

Most sub-systems of the pulsar receiver can be configured and controlled with an easy to use graphical user interface that runs on the main computer system. For pulsar observations, since it is advisable to switch off the automatic level controllers at the IF and baseband systems, the power levels from each antenna are individually adjusted to ensure proper operating levels at the input to the correlator. The format for the binary output data is peculiar to the GMRT pulsar receiver. Simple programs to read the data files and display the raw data - including facilities for dedispersion and folding - are available at the observatory and can be used for first order data quality checks, both for the incoherent mode and coherent mode systems.

Chapter 19

GMRT Antennas and Feeds

G. Sankar

19.1 Introduction

A radio telescope in its simplest form consists of three components (see also Chapter 3), (i) an antenna that selectively receives radiation from a small region of the sky, (ii) a receiver that amplifies a restricted frequency band from the output of the antenna and (iii) a recorder for registering the receiver output. In this chapter we focus on the antenna, and in particular the antennas used for the GMRT.

The GMRT antennas are parabolic reflector antennas. The first reflector antenna was invented by Heinrich Hertz in 1888 to demonstrate the existence of electromagnetic waves which had been theoretically predicted by J.C.Maxwell. Hertz's antenna was a cylindrical parabola of $f/D = 0.1$ and operated at a wavelength of 66 cm.(450 MHz). The next known reflector antenna was that constructed in 1930 by Marconi for investigating microwave propagation. After that, in 1937, Grote Reber constructed the prototype of the modern dish antenna - a prime-focus parabolic reflector antenna of 9.1 m. diameter, which he used to make the first radio maps of the sky. During and after World War II, radar and satellite communication requirements caused great advances in antenna technology.

19.2 Types of Antennas

A diverse variety of antennas have been used for radio astronomy (see eg. Chapter 3) the principal reason for this diversity being the wide range of observing wavelengths: from ~ 100 m to ~ 1 mm, a range of 10^5. However the most common antenna used for radio astronomy is the paraboloid reflector with either prime-focus feeds or cassegrain type feed arrangement.

Prime-focus parabolic antennas although mechanically simple have certain disadvantages, viz. (i) the image-forming quality is poor due to lower f/D ratios in prime-focus antennas, and (ii) the feed antenna pattern extends beyond the edge of the parabolic reflector and the feed hence picks up some thermal radiation from ground. The cassegrain system which uses a secondary hyperboloid reflector and has the feed located at the second focus of the secondary solves these problems. For cassegrain systems the f/D ratio is higher and further the feed "looks" upwards and hence pick up from the ground is minimized. This is a great advantage at higher frequencies, where the ground brightness

temperature (~ 300 K) is much higher than the brightness temperature of the sky. However this is achieved at the price of increased aperture blockage caused by the secondary reflector.

A primary advantage of paraboloid antennas (prime focus or cassegrain) is the ease with which receivers can be coupled to it. The input terminals are at the feed horn or dipole. A few other advantages are: (i) high gain, a gain of $\simeq 25$ dB for aperture diameters as small as 10λ is easily achievable, (ii) full steerability, generally either by polar or azimuth-elevation mounting. Further the antenna characteristics are to first order independent of pointing, (iii) operation over a wide range of wavelengths simply by changing the feed at the focus.

Compared to optical reflectors paraboloid reflectors used for radio astronomy generally have a short f/D ratio. Highly curved reflectors required for higher f/D ratios result in increased costs and reduced collecting areas. Although the reflecting antennas are to first order frequency independent, there is nonetheless a finite range of frequencies over which a given reflector can operate. The shortest operating wavelength is determined by the surface smoothness of the parabolic reflector. If λ_{mn} is the shortest wavelength,

$$\lambda_{mn} \approx \sigma/20 \tag{19.2.1}$$

where, σ is the rms deviation of the reflector surface from a perfect paraboloid. Below λ_{mn} the antenna performance degrades rapidly with decreasing wavelength. The longest operating wavelength λ_{mx}, is governed by diffraction effects. As a rule of thumb the largest operating wavelength λ_{mx} is given by

$$\lambda_{mx} < 2\bar{L} \tag{19.2.2}$$

where, $\bar{L}$ is the mean spacing between feed–support legs. At $\lambda = \bar{L}$ the feed support structure would completely shadow the reflector.

19.3 Characterizing Reflector Antennas

One important property of any antenna is that its radiation characteristics when it is used as a transmitter are the same as when it is in the receiving-mode. This is a consequence of the well-known electromagnetic fields *principle of reciprocity*. Even though radio telescope antennas are generally used only for receiving signals, it is often simpler to characterize it by considering the antenna to be in the transmitting mode. Antenna terminology is also influenced by the reciprocity principle, for example we have been calling the dipole or horn placed at the focus of the reflector to receive the signal from distant sources as the "**feed**", i.e. as though it were coupled to a transmitter rather than a receiver.

All antennas can be described by the following characteristics (see also Chapter 3)

1. Radiation pattern The field strength that the antenna radiates as a function of direction. The simplest type of antenna normally radiates most of its energy in one direction called the 'primary beam' or 'main lobe'. The angular width of the main lobe is determined by the size and design of the antenna. It is usually parametrized by its full width at half maximum, also called its 3dB beamwidth. Weaker secondary maxima in other directions are called *side lobes*. Although the pattern is a function of both elevation and azimuth angle, it is often only specified as a function of elevation angle in two special orthogonal planes, called the E-plane and the H-plane.

2. Directivity The radiated power in the direction of the main lobe relative to what would be radiated by an isotropic antenna with the same input power. A related quantity

called the Gain also takes into account any electrical losses of the antenna. For reflector antennas, one can also define an aperture efficiency which is the ratio of the effecting collecting area of the telescope to its geometric area. For the relation between the gain and the effective collecting area see Chapter 3.

3. Polarization The sense of polarization that the antenna radiates or receives as a function of direction. This may be linear, circular, or elliptical. Note that when describing the polarization of a wave, it is sufficient to specify the polarization of the electric-field vector.

4. Impedance From the point of view of the microwave circuit behind the antenna, the antenna can be represented as a complex load impedance. The characteristics of this load depend on the radiation patterns of the antenna and hence the design of the antenna. The goal of a good design is to match the impedance of the antenna to the impedance of the transmission line connecting the antenna to the receiver. The impedance match can be characterized by any one of the following parameters:

   - the voltage reflection coefficient, ρ_v.
   - the return loss (in dB), $R_L = -20\log|\rho_v|$.
   - the voltage standing-wave ratio, $VSWR = \frac{1+|\rho_v|}{1-|\rho_v|}$.

5. Phase Center All horns and feeds have a *phase center*. This is the theoretical point along the axis of the feed which is the center of curvature of the phase fronts of the emerging spherical waves.

19.4 Computing Reflector Antenna Radiation Patterns

Reflector antenna radiation patterns are determined by a number of factors, but the most important ones are the radiation pattern of the feed antenna and the shape of the reflector. Parabolic reflectors have the unique feature that all path lengths from the focal point to the reflector and on to the aperture plane are the same. As shown in Figure 19.1,

$$\begin{aligned} FP + PA &= \rho + \rho\cos\theta' \\ &= \rho(1+\cos\theta') \\ &= 2f, \end{aligned} \tag{19.4.3}$$

since the parabola is described in polar form by, $\rho(1+\cos\theta') = 2f$

When the reflector dimensions are large compared to the wavelength, geometrical optics principles can be used to determine the power distribution in the aperture plane. If the feed pattern is azimuthally symmetric, then the normalized far-field radiation pattern of reflector depends on

1. $\pi u = k\ a\ \sin\theta$, where a is the radius of the aperture, $k = 2\pi/\lambda$, and θ is the angle subtended by the far-field point with respect to the parabola's focal axis

2. The feed taper,C [4],[5], which is defined as the amplitude of the feed radiation pattern at the rim of the parabolic reflector relative to the maximum value (assumed to be along the parabola axis). (Note that in standard power plots of radiation patterns (in dB), the edge taper T_E is related to C by $T_E = 20\log C$).

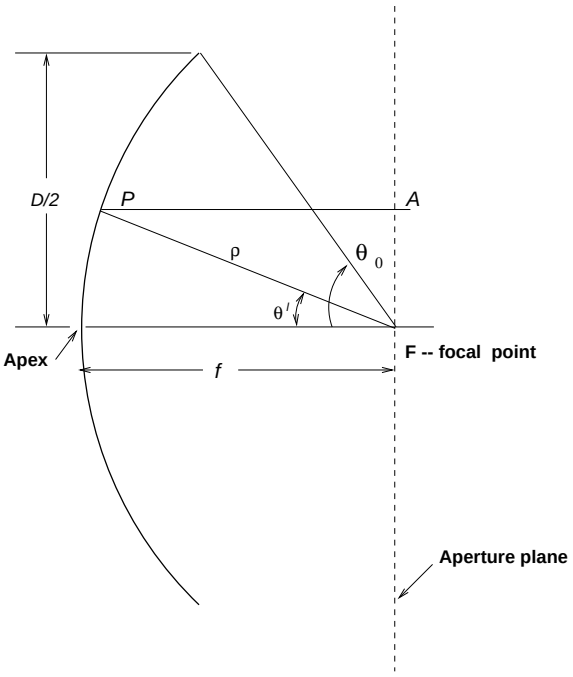


Figure 19.1: Geometry for determining the aperture field distribution for a prime focus parabolic antenna.

3. The focal length f which determines how the power from the feed is spread over the aperture plane. If $\tilde{g}(\theta')$ is the radiation pattern of the feed, r is distance in the aperture plane, and $g(r)$ is the power density in the aperture plane, then we have

$$g(r)\ dr = \tilde{g}(\theta')\ d\theta',\ \text{i.e. } g(r) = \tilde{g}(\theta')\ \frac{d\theta'}{dr} \tag{19.4.4}$$

and from Figure 19.1 we have

$$\frac{d\theta'}{dr} = \frac{2f}{1 + cos(\theta')} \tag{19.4.5}$$

In Chapter 3 we saw that the far field is in general the Fourier transform of the aperture plane distribution. In the case of azimuthally symmetric distributions, this can be written as

$$F(u) \quad = \quad \int_0^\pi g(q) J_0(qu) q\ dq$$

where $F(u)$ is the far field pattern, q is a normalized distance in the aperture plane, $q = \pi(r/a)$, $g(q)$ is the feed's pattern projected onto the aperture plane as discussed above. A convenient parameterization of the feed pattern in terms of the taper, C is

$$g\Big(\frac{r}{a}\Big) \quad = \quad C + (1 - C)\Big[1 - \Big(\frac{r}{a}\Big)^2\Big]^n \tag{19.4.6}$$

$$\tag{19.4.7}$$

The aperture illuminations corresponding to different values of the parameter n are shown in Figure 19.2. The case $n = 0$ corresponds to a uniform aperture distribution.

For uniform illumination the far field pattern is given by

$$F(u) = 2 \cdot \frac{J_1(\pi u)}{(\pi u)} \tag{19.4.8}$$

Simple closed-form expressions are available for integer values of n. If the above expression $F(u)$ is denoted as $F_0(u)$, (since $n = 0$) the general form for any integer n is given by

$$F_n(u) \quad = \quad \frac{n+1}{Cn+1} \cdot \Big[C F_0(u) + \frac{1-C}{n+1} f_n(u)\Big] \tag{19.4.9}$$

where,

$$f_n(u) \quad = \quad 2^{n+1}(n+1)! \frac{J_{n+1}(\pi u)}{(\pi u)^{n+1}} \tag{19.4.10}$$

Table 19.1 gives the halfpower beamwidth (HPBW), the first sidelobe level and the taper efficiency (see Section 19.4.1) for various edge tapers C and shape parameter n.

From Table 19.1 (see also the discussion in Chapter 3) we find that as the edge-taper parameter C decreases, the HPBW increases, the first sidelobe level falls and the taper-efficiency also decreases. Note that C has to be less than unity since we have assumed a radiation pattern which decreases monotonically with increasing angle from the symmetry-axis (Eqn 19.4.6, Fig 19.2).

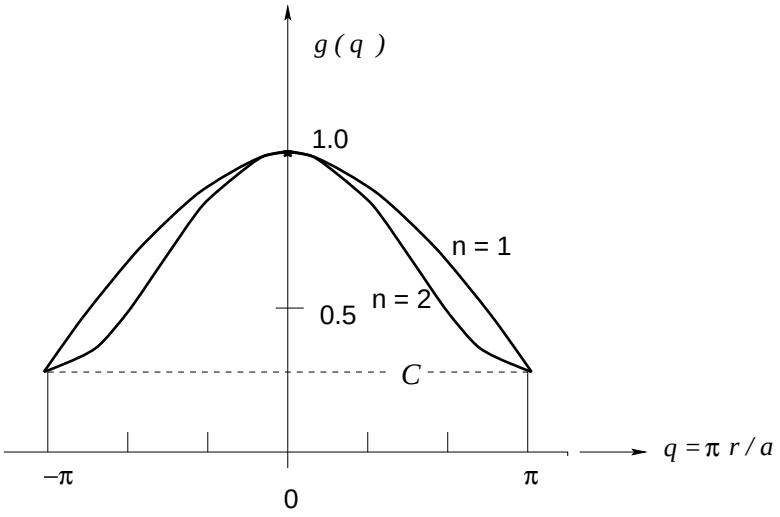


Figure 19.2: The shape of the aperture illumination as given by eqn 19.4.6 for different values of the parameter n.

Table 19.1: Radiation characteristics of circular aperture

Edge Taper		$n = 1$			$n = 2$		
T_E (dB)	C	HPBW (rad.)	Sidelobe level (dB)	η_t	HPBW (rad.)	Sidelobe level (dB)	η_t
-8	0.398	$1.12\lambda/2a$	-21.5	0.942	$1.14\lambda/2a$	-24.7	0.918
-10	0.316	$1.14\lambda/2a$	-22.3	0.917	$1.17\lambda/2a$	-27.0	0.877
-12	0.251	$1.16\lambda/2a$	-22.9	0.893	$1.20\lambda/2a$	-29.5	0.834
-14	0.200	$1.17\lambda/2a$	-23.4	0.871	$1.23\lambda/2a$	-31.7	0.792
-16	0.158	$1.19\lambda/2a$	-23.8	0.850	$1.26\lambda/2a$	-33.5	0.754
-18	0.126	$1.20\lambda/2a$	-24.1	0.833	$1.29\lambda/2a$	-34.5	0.719

19.4.1 Aperture Efficiency

The "aperture efficiency" of an antenna was earlier defined (Sec 19.3) to be the ratio of the effective radiating (or collecting) area of an antenna to the physical area of the antenna. The aperture efficiency of a feed-and-reflector combination can be decomposed into five separate components: (i) the illumination efficiency or "taper efficiency", η_t, (ii) the spillover efficiency, η_S, (iii) the phase efficiency, η_p, (iv) the crosspolar efficiency, η_x and (v) the surface error efficiency η_r.

$$\eta_a = \eta_t \, \eta_S \, \eta_p \, \eta_x \, \eta_r. \tag{19.4.11}$$

The illumination efficiency (see also Chapter 3, where it was called simply "aperture efficiency") is a measure of the nonuniformity of the field across the aperture caused by the tapered radiation pattern (refer Figure 19.2). Essentially because the illumination is less towards the edges, the effective area being used is less than the geometric area of the reflector. It is given by

$$\eta_t = \frac{|\int_0^R g(r)dr|^2}{\int_0^R |g(r)|^2 dr}, \tag{19.4.12}$$

where $g(r)$ is the aperture field. Note that this has a maximum value of 1 when the aperture illumination is uniform, i.e. $g(r) = 1$. The illumination efficiency can also be written in terms of the electric field pattern of the feed $E(\theta)$, viz.

$$\eta_t = 2\cot^2 \frac{\theta_0}{2} \cdot \frac{|\int_0^{\theta_0} E(\theta)\tan(\theta/2)d\theta|^2}{\int_0^{\theta_0} |E(\theta)|^2 \sin(\theta)d\theta}, \tag{19.4.13}$$

where θ_0 is angle subtended by the edge of the reflector at the focus (Figure 19.1).

When a feed illuminates the reflector, only a proportion of the power from the feed will intercept the reflector, the remainder being the spillover power. This loss of power is quantified by the spillover efficiency, i.e.

$$\eta_S = \frac{\int_0^{\theta_0} |E(\theta)|^2 \sin(\theta)d\theta}{\int_0^{\pi} |E(\theta)|^2 \sin(\theta)d\theta}. \tag{19.4.14}$$

Note that the illumination efficiency and the spillover efficiency are complementary; as the edge taper increases, the spillover will decrease (and thus η_S increases), while the illumination or taper efficiency η_t decreases[1] The tradeoff between η_S and η_t has an optimum solution, as indicated by the product η_S * η_t in Figure 19.3. The maximum of $\eta_S \eta_t$ occurs for an edge taper of about -11 dB and has a value of about 80 %. In practice, a value of -10 dB edge taper is frequently quoted as being optimum.

The surface-error efficiency is independent of the feed's illumination. It is associated with far-field cancellations arising from phase errors in the aperture field caused by errors in the reflector's surface. If δ is the rms error in the surface of the reflector, the surface-error efficiency is given by

$$\eta_r = \exp -(4\pi\delta_p/\lambda)^2 \tag{19.4.15}$$

The remaining two efficiencies, the phase efficiency and the cross polarization efficiency, are very close to unity; the former measures the uniformity of the phase across

[1]Recall also from Chapter 3 that as the illumination is made more and more uniform the sidelobe level increases.

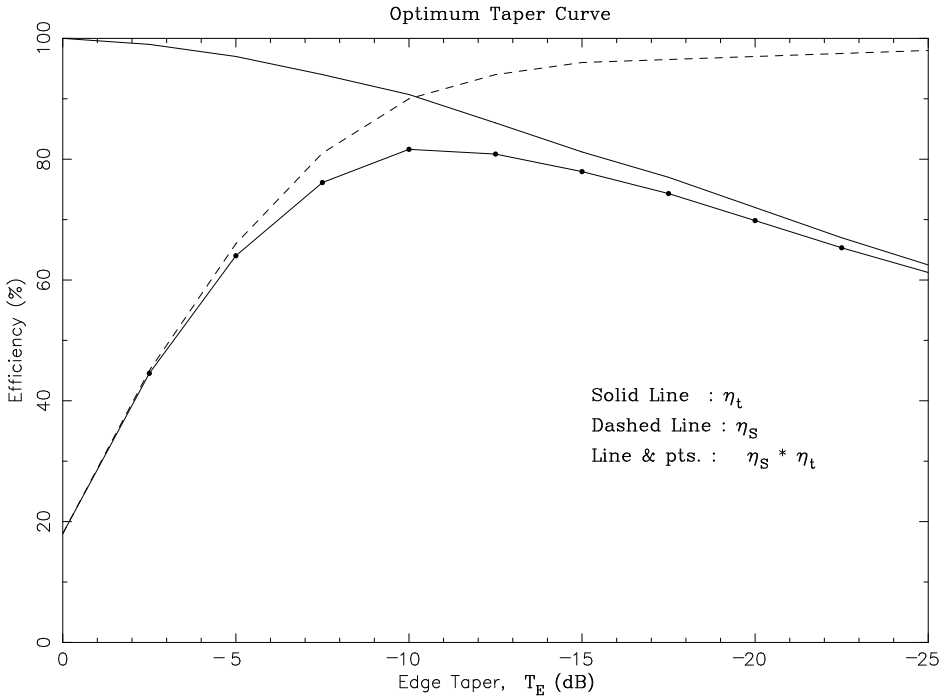


Figure 19.3: Illumination efficiency and spillover efficiency as a function of edge taper. The optimum taper is at ~ -11 dB.

the aperture and the latter measures the amount of power lost in the cross-polar radiation pattern. For symmetric feed patterns[6], η_x is defined thorough the copolar,$C_p(\theta)$ and cross-polar patterns,$X_p(\theta)$:

$$\eta_x = \frac{\int_0^{\theta_0} |X_p(\theta)|^2 \sin(\theta) d\theta}{\int_0^{\theta_0} \left(|C_p(\theta)|^2 + |X_p(\theta)|^2\right) \sin(\theta) d\theta} \tag{19.4.16}$$

where,

$$\begin{aligned} C_p(\theta) &= 1/2[E(\theta) + H(\theta)] \\ X_p(\theta) &= 1/2[E(\theta) - H(\theta)] \end{aligned} \tag{19.4.17}$$

It can be seen that if one can design an antenna,having identical $E(\theta), H(\theta)$ patterns the cross-polar pattern will vanish. Taking the cue from this, *the feed for antenna could also designed with a goal to match E and H patterns at least up to the subtended angle of the dish edge*, θ_0.

With this background we now proceed to take a detailed look at the GMRT antennas.

19.5 Design Specifications for the GMRT Antennas

The f/D ratio for the GMRT antennas was fixed at the value 0.412 based both on structural design issues as well as preliminary studies of various feeds radiation patterns. Since the antennas are to work at meter wavelengths prime focus feeds were preferred. Cassegrain feeds at meter wavelengths would result in impractically large secondary mirrors (the mirror has to be several λ across) and concomitant large aperture blockage.

Six bands of frequencies had been identified [1] for the GMRT observations. It was deemed essential to be able to change the observing frequency rapidly, and consequently the feeds had to mounted on a rotating turret placed at the prime focus. If one were to mount all the six feeds on a rotating hexagon at the focus, the adjacent feeds will be separated by 60°. If one wants to illuminate the entire aperture, then one has to have a feed pattern that extends at least up to the subtended angle of the parabola edge, which is $\theta_0 = 62.5^\circ$ (Note that $\cot(\theta_0/2) = 4f/D$, Figure 19.1). Hence this arrangement of feeds would cause the one feed to "see" the feeds on the adjacent faces. It was decided therefore to mount the feeds in orthogonal faces of a rotating cube. Since one needs six frequency bands, this leads to the constraint that at least two faces of the turret should support dual frequency capability. For astronomical reasons also dual frequency capability was highly desirable.

One specific aspect of GMRT design is the use of mesh panels to make the reflector surface[1]. Since the mesh is not perfectly reflective, transmission losses thorough the mesh have to be taken into account. Further, the expected surface errors of the mesh panels was ~ 5 mm. This implies that the maximum usable frequency is (see Section 19.2) ~ 3000 MHz, independent of the transmission losses of the mesh. (Incidentally, since the mean-spacing of feed-support legs, $\bar{L} = 23.6$ m, the lowest usable frequency is around 6 MHz).

Several analytical methods exist in literature to compute the transmission loss through a mesh as a function of the cell size, the wire diameter and the wavelength of the incident radiation. The one chosen for our application is has good experimental support [2,3]. At the GMRT, the mesh size is 10×10 mm for the central 1/3 of the dish, 15×15 mm of the

Mesh size	$\lambda = 21$ cm.	$\lambda = 50$ cm.
10 mm.	-15.8 dB	-23.3 dB
15 mm.	-11.4 dB	-18.4 dB
20 mm.	-8.1 dB	-14.6 dB

Table 19.2: Transmission losses through the GMRT wire mesh

middle 1/3 of the dish and 20×20 mm for the outer 1/3 of the dish. The wire diameter is 0.55 mm. The transmission loss for at two fiducial wavelengths for these various mesh sizes is given in Table 19.2.

Each section of the dish not only has a separate mesh size but also a separate surface rms error. If we call these rms surface errors $\sigma_1, \sigma_2, \sigma_3$ and the respective transmission losses (at some given wavelength) τ_1, τ_2, τ_3, then the surface rms efficiency given by Eqn 19.4.15 has to be altered to a weighted rms efficiency:

$$\eta_r = \frac{A_1 + A_2 + A_3}{\int_0^{\theta_0} |E(\theta)|^2 \sin(\theta) d\theta}$$

where,

$$A_1 = \exp\left[-\left(\frac{4\pi\sigma_1}{\lambda}\right)^2\right] \int_0^{\theta_2} |E(\theta)|^2 \sin(\theta) d\theta \quad (19.5.18)$$

$$A_2 = \exp\left[-\left(\frac{4\pi\sigma_2}{\lambda}\right)^2\right] \int_{\theta_2}^{\theta_1} |E(\theta)|^2 \sin(\theta) d\theta \quad (19.5.19)$$

$$A_3 = \exp\left[-\left(\frac{4\pi\sigma_3}{\lambda}\right)^2\right] \int_{\theta_1}^{\theta_0} |E(\theta)|^2 \sin(\theta) d\theta \quad (19.5.20)$$

and $\theta_2,\ \theta_1$ are the subtended angles of the first and second points of mesh–transition–zones, as illustrated in Figure 19.4

The transmission loss gives a corresponding mesh–leakage or *mesh–transmission* efficiency,η_{mt}, which is given by

$$\eta_{mt} = \frac{B_1 + B_2 + B_3}{\int_0^{\theta_0} |E(\theta)|^2 \sin(\theta) d\theta} \quad (19.5.21)$$

where,

$$B_1 = (1 - \tau_1) \int_0^{\theta_2} |E(\theta)|^2 \sin(\theta) d\theta \quad (19.5.22)$$

$$B_2 = (1 - \tau_2) \int_{\theta_2}^{\theta_1} |E(\theta)|^2 \sin(\theta) d\theta \quad (19.5.23)$$

$$B_3 = (1 - \tau_3) \int_{\theta_1}^{\theta_0} |E(\theta)|^2 \sin(\theta) d\theta \quad (19.5.24)$$

Efficiencies computed for the different GMRT feeds (using their measured pattern, being the input) are given in Table 19.4.

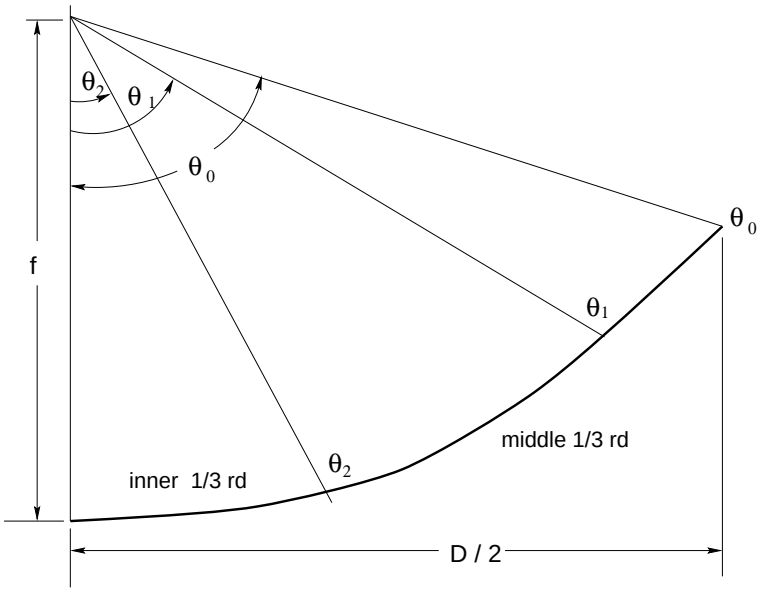


Figure 19.4: Schematic of the sub division of the GMRT antenna surface into 3 zones. The mesh size as well as the rms surface error is different in the different zones.

19.5.1 Secondary Patterns

The antenna pattern at 327 MHz as computed using geometric optics is shown in Figure 19.5. More rigorous analytical models (the *Uniform Theory of Diffraction* [7]) gives the pattern shown in Figure 19.6.

There is a pronounced difference seen at the side–lobe structures between these two models, while the primary beam shows near–identical shapes and the HPBW value matches to a second decimal accuracy. The computed HPBW also agrees to within measurement errors with the observed HPBW of the actual GMRT antennas.

19.6 GMRT Feeds

19.6.1 Feed Placement

Recall that from the constraints outlined in Sec 19.5 it had been decided that the feed turret should be cubical in shape. Fig 19.7 shows the placement of feeds on the turret. The phase–centers of all the feeds are coincident with the paraboloid focus. The space between the turret and the feed is utilized for mounting the front-end electronics. There are six bands altogether, $1000 - 1450$ MHz[2], 610 MHz, 327 MHz, 233 MHz, 150 MHz and 50 MHz. The 50 MHz feed[3] is affixed onto the feed support legs and not onto the turret. As such it is in focus at all times. The 610 MHz and 233 MHz feeds are mounted on the same turret face.

Each type of feed - its design and performance are briefly outlined in the following sections. More information can be found in [8].

[2]Note that some of the antennas have feeds that extend to 1750 MHz.
[3]Which is not yet operational

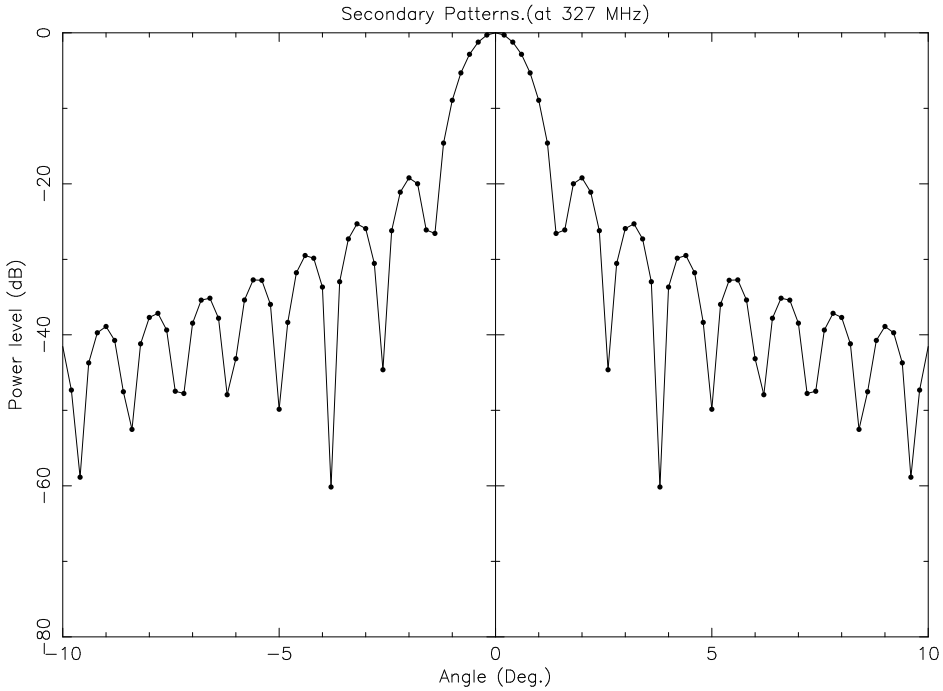


Figure 19.5: Computed pattern (using geometric optics) of a GMRT antenna at 327 MHz.

19.6.2 150 MHz Feed

This feed employs four dipoles in a "boxing ring" configuration, placed above a plane reflector. The unique feature of the dipole is that it is wide-band i.e. has an octave bandwidth. It is a folded dipole with each arm being a "*thick*" dipole. A dipole is called '*thin*' when its diameter, $d > 0.05\lambda$. For such dipoles a sinusoidal current distribution can be assumed for the computation of input impedance and related radiation parameters.

Thin dipoles have narrowband radiation characteristics. One method by which its acceptable operational bandwidth can be increased is to decrease the l/d ratio. For example, an antenna with a $l/d \approx 5000$ has an acceptable bandwidth of about 3%, while an antenna of the same length but with a $l/d \approx 260$ has a bandwidth of about 30%. By folding the dipole, one gets a four–fold increase in input impedance compared to a simple dipole. The 150 MHz feed also has a transmission line impedance transformer coupled to the excitation point [9].

Traditionally crossed–dipoles are used to give sensitivity to both polarizations. However since a crossed–dipole configuration in this design would be extremely cumbersome, a "boxing ring" design was instead chosen. Here one pair of dipoles at $\lambda/2$ spacing provides sensitivity to one linear polarization. Another pair orthogonally oriented with respect to the first pair gives sensitivity to the orthogonal polarization. The overall dimensions of the feed are:

- Folded dipole length : $0.39\ \lambda$
- Dipole height above reflector : $0.29\ \lambda$

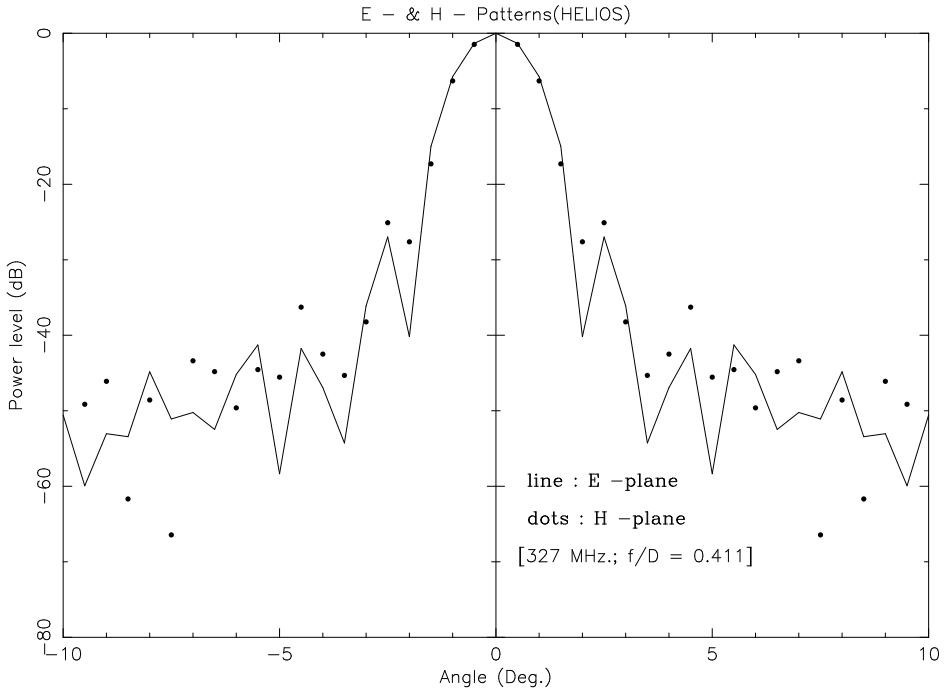


Figure 19.6: Computed pattern (using uniform theory of diffraction) of a GMRT antenna at 327 MHz.

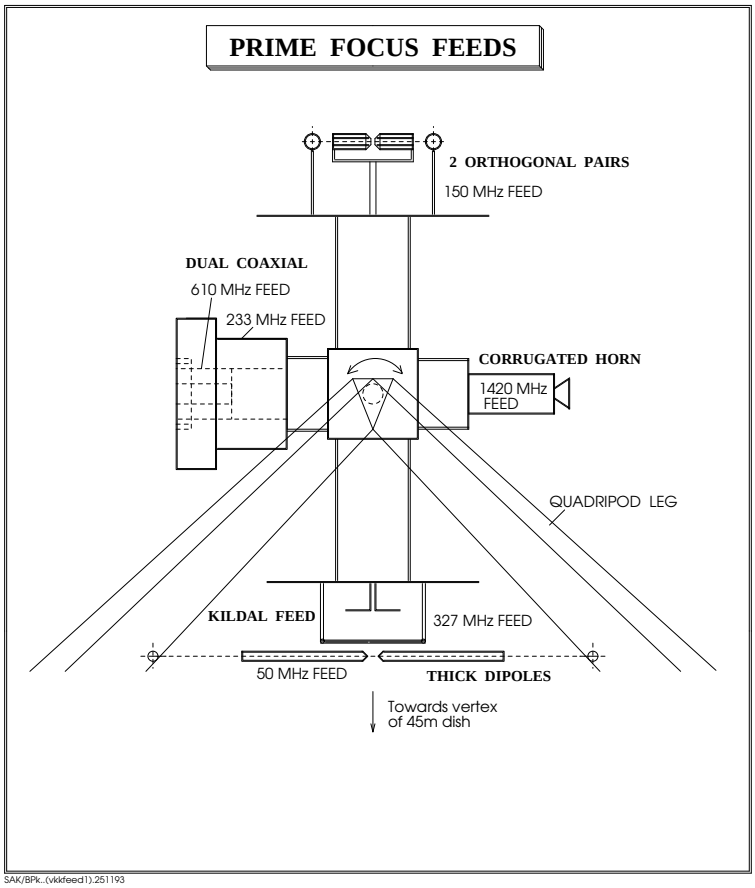


Figure 19.7: Schematic diagram showing the arrangement of the different feeds on the feed turret.

- Reflector (diagonal of octagon) : $1.2\,\lambda$

The dipoles have an l/d ratio of 6.48, and the phase center was determined to be at a height of 100 mm above the reflector. The feed's impedance bandwidth can be seen on the VSWR plot of Figure 19.8

The usable bandwidth for a feed is given approximately by the range for which SWR ≤ 2.0. By this criteria, the frequency coverage of the 150 MHz feed is from 117 MHz to 247 MHz, i.e. a bandwidth of 130 MHz, or 86% bandwidth. The radiation pattern gives an edge taper, $T_E = -9$ dB.

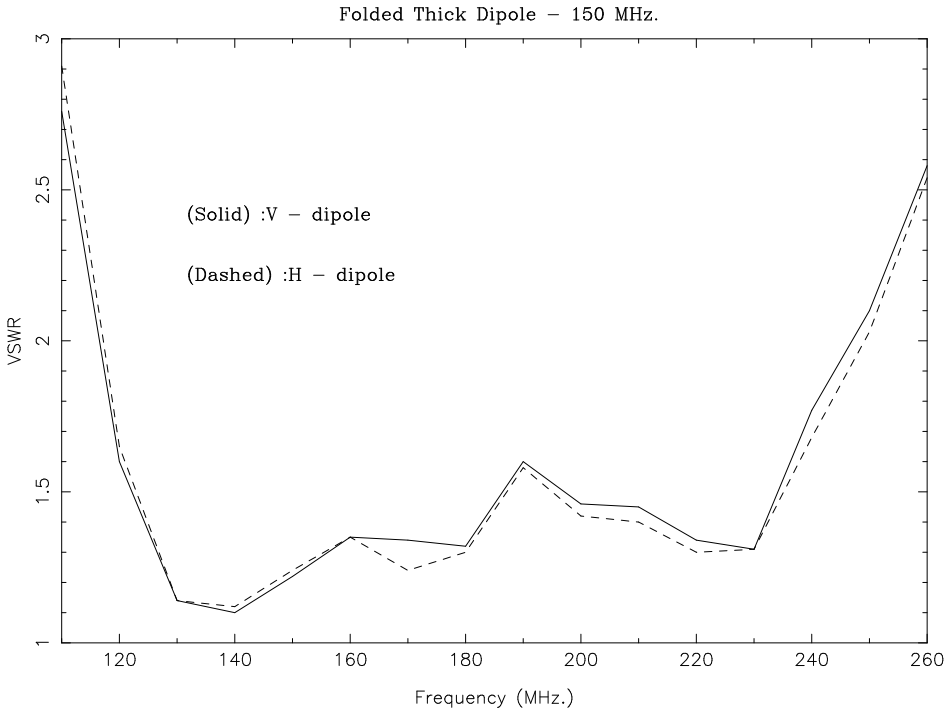


Figure 19.8: VSWR for the 150 MHz feed.

One undesirable feature of this feed is the high value of cross–polarization, as compared to that at other frequencies (see Figure 19.9)[4]. The cross–polar peak for 150 MHz. is -17 dB and the on–axis cross polarization is also at about the same level.

One–pair of outputs from the dipoles which are parallel to each other are connected to a power–combiner, whose output goes to one port of the quadrature hybrid (which adds two linear polarized signals to yield one circular polarized signal). Similarly the orthogonal pair of dipoles are connected to the other port of the hybrid. Both the power combiners and the quadrature hybrid are mounted inside one of the front-end chassis, placed behind the feed.

19.6.3 327 MHz Feed

Generally a dipole has a broader H pattern than its E pattern (the E pattern being in the plane containing the dipole). Recall from the discussion in section 19.4.1 that for good cross–polarization properties it was essential to have matched E and H plane patterns. An elegant method for achieving this pattern matching was given by P.S.Kildal [10], and involves placing a *beam forming ring* (BFR) above the dipole[5]. The conducting ring is placed above the dipole in a plane parallel to the reflector and is supported by dielectric rods. The beam forming ring compresses the H-plane pattern while it has no significant effect on the E-plane.

[4]Note that the cross polar pattern was measured using the standard technique outlined in [4 ; pp.177–79]. The cross–polar levels are measured with respect to a co–polar maximum of 0 dB.

[5]This design has been christened '*Kildal Feed*' in the local jargon.

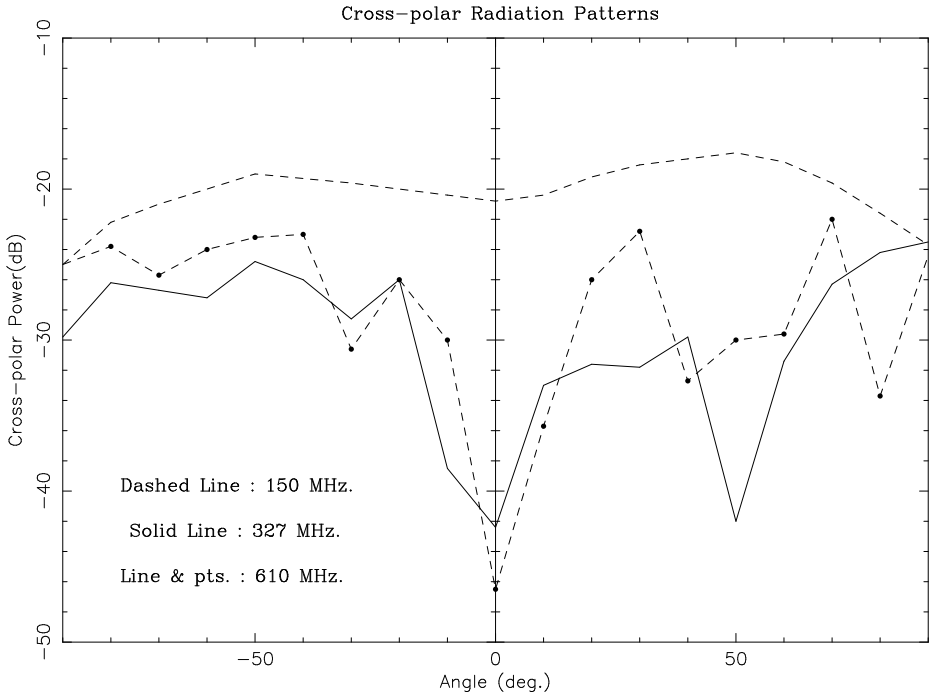


Figure 19.9: The cross polarization of different GMRT feeds. The 150 MHz feed has relatively larger cross-polarization.

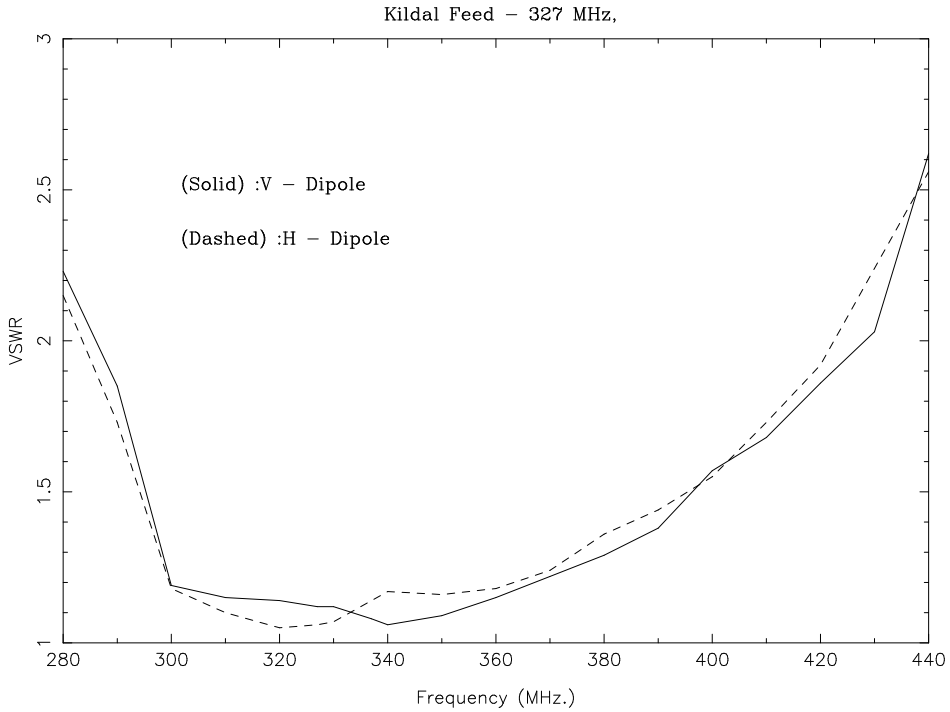


Figure 19.10: The VSWR as a function of frequency for the 327 MHz feed.

The optimum dimensions of the dipole, BFR and reflector were arrived at by careful measurements done on a scaled-up version (i.e. at 610 MHz) and a follow–up measurements on a prototype 327 MHz model. The values arrived at were :

- Reflector diameter : 2.2λ.
- Height of dipole above reflector : 0.26λ.
- BFR diameter : 1.22λ.
- BFR height above reflector : 0.51λ.

The measured phase center is at 26 mm above the reflector for both E and H- planes. Crossed dipoles are employed for dual polarization. The 327 MHz feed actually deviates slightly from the original Kildal's design – there are sleeves over the dipoles. These sleeves increase the bandwidth of the feed [5]. The VSWR plot for the 327 MHz feed is given in Figure 19.10.

For SWR ≤ 2.0, the bandwidth is 138 MHz.(286 to 424 MHz.) The measured antenna pattern is given in Fig 19.11. The edge taper, T_E is -12.2 dB. Fig 19.9 shows the cross-polar pattern. It is seen that a cross-polar maximum of -27.5 dB (mean value) has been achieved.

The linear polarized outputs of the dipoles are mixed in a quadrature hybrid at one of the front-end chassis to produce two circular polarized (both left and right) signals, which go further into the amplifying, signal conditioning circuits of front-end Electronics.

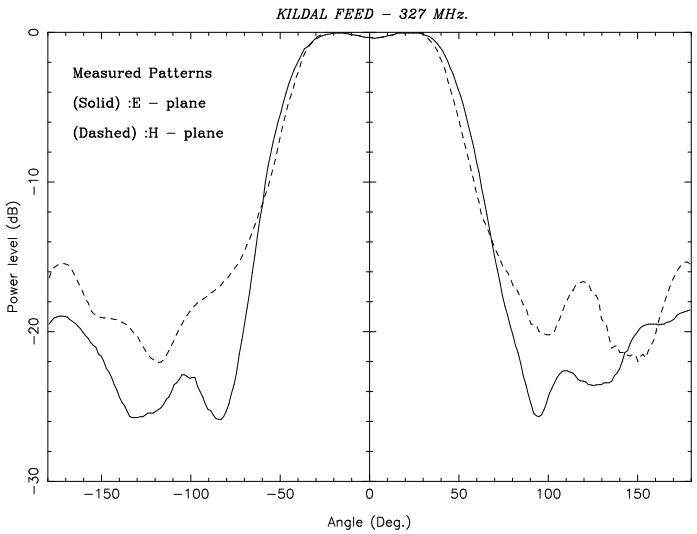


Figure 19.11: The measured antenna pattern at 327 MHz

19.6.4 Dual-Frequency Coaxial Waveguide Feed

The 610 MHz and 233 MHz feeds are dual frequency coaxial feeds. The single most attractive feature of coaxial waveguide feed is its' multi-frequency launching capability. Simultaneous transmission or reception of well separated frequencies is possible. Coaxial feeds have been used as on board satellite antennas to provide coverage at three separate frequency bands [11]. Coaxial feeds have also been used at the WSRT (operated by NFRA, The Netherlands). The prime focus feed system has at WSRT has two separate multi-frequency coaxial waveguides, covering 327 MHz, 2300 MHz in one and 610 MHz, 5000 MHz in another [12],[13].

The design of the GMRT 610 MHz/233 MHz waveguide feeds is based on an exhaustive theoretical analysis of the design of coaxial waveguide feeds [14],[15]. A constraint in such multi-frequency designs is that adjacent frequency bands should not overlap to within an octave. Thus at the GMRT either the 150 MHz or the 233 MHz could have been combined with 610 MHz. However the former choice was rejected since it resulted in unwieldy dimensions of the feed.

The fundamental mode of propagation in coaxial structures is TEM, hence the radiated field component along the axis is zero everywhere. Obviously for a feed this is the most undesirable characteristic. So propagation by an alternate mode (single or multiple) is essential. Coaxial waveguides must then be forced to radiate in TE_{11} mode. This can be achieved simply by exciting the probes in phase opposition[6].

In the dual frequency construction the outer conductor of the 610 MHz serves as the inner one for the 233 MHz. Quarter wavelength chokes are provided in both the frequency

[6]Low loss baluns are essential in such designs.

Dimensions	*610 MHz Coaxial*	*233 MHz Coaxial*
Aperture diameter	$0.9\,\lambda$	$0.85\,\lambda$
Waveguide cavity length	$0.95\,\lambda$	$0.73\,\lambda$

Table 19.3: Dimensions of the 610/233 MHz coaxial feed.

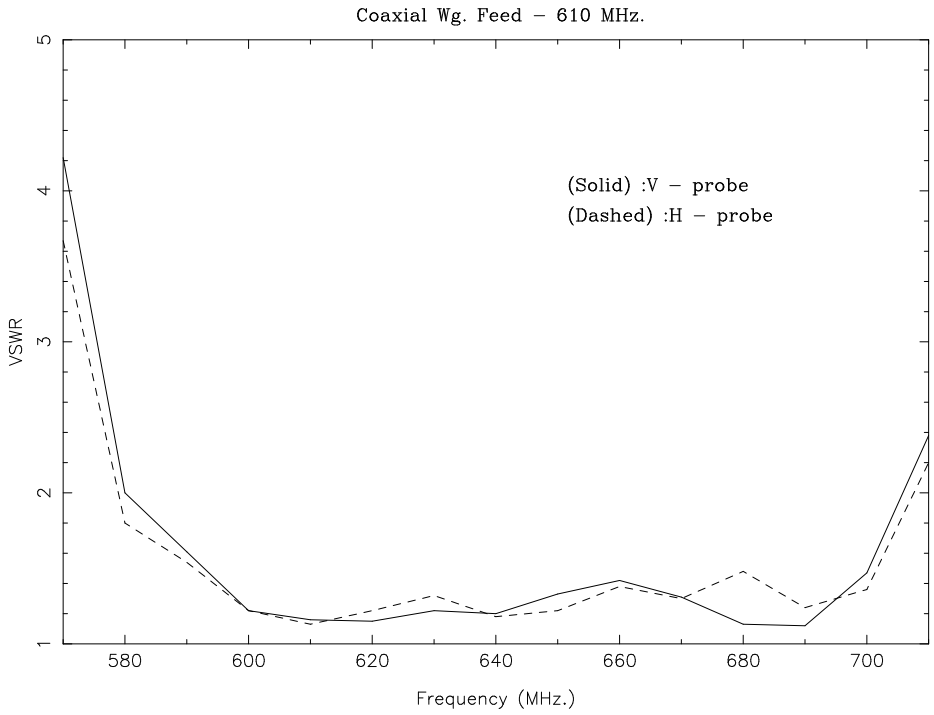


Figure 19.12: The VSWR as a function of frequency for the 610 MHz feed.

parts to cut down the surface currents on the outer conductor and thereby ensure pattern symmetry. The waveguide feeds have two pairs of probes. One pair supports a given plane polarization while the orthogonal pair supports the orthogonal polarization. Similar to the dipole feed discussed in the previous section, a quadrature hybrid at the back-end of the coaxial feed is used to convert the linear polarization to circular polarization. The rear-half of the 610 MHz feed, separated by a partition disc, is utilized to accommodate the baluns, quadrature hybrids and low-noise amplifiers of 610 MHz and the baluns of 233 MHz. The overall dimensions of the feed are given in Table 19.3

The phase center is not at the aperture plane, but at a point 60 mm in front of the aperture. A similar displacement of the phase center is also seen in the WSRT coaxial feeds [13]. Fig 19.12 shows the VSWR plot for an optimized probe geometry at 610 MHz For SWR ≤ 2.0, the band goes from 580 MHz to 707 MHz, i.e. a total bandwidth of 127 MHz. The feed patterns measured at 610 MHz are shown in Fig 19.13; the edge taper is -9.8 dB. The cross-polar maximum is -22.8 dB.

Fig 19.14 shows the VSWR plot of 233 MHz– part of the coaxial feed.

For SWR ≤ 2.0, the bandwidth is 12 MHz, i.e. this feed is rather narrow as compared to all other frequency bands. The effect of the inter-coupling of radiated power between

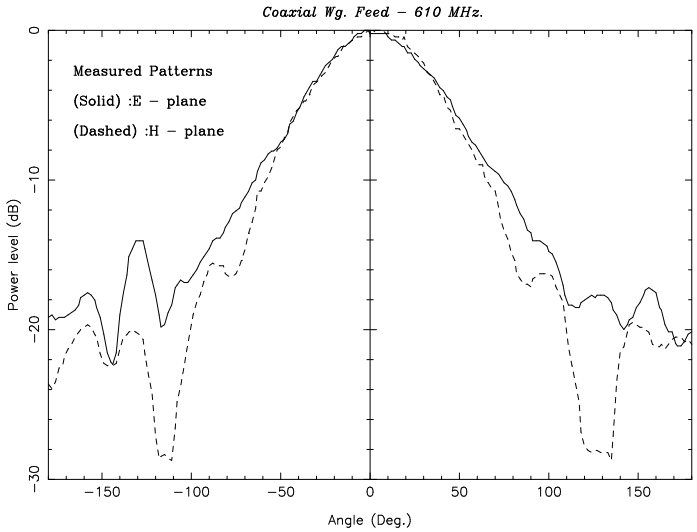


Figure 19.13: The feed pattern of the 610 MHz feed.

the two frequencies of the coaxial feed on the radiation patterns has been studied. The main lobe does not show any significant change due to the presence of the other coaxial waveguide part.

19.7 1000–1450 MHz Feed

This feed was designed and constructed by the Millimeter Wave Laboratory of the Raman Research Institute. It is of the corrugated horn type - known for its high aperture efficiency and very low cross-polarization levels. In any horn, the antenna pattern is severely affected by the diffraction from the edges which can lead to undesirable radiation not only in the back lobes but also in the main lobe. By making grooves on the walls of the of a horn, the spurious diffractions are eliminated. Such horns are called "*Corrugated horns*"[4]. Our feed at 1420 MHz. has fins instead of grooves, since the whole assembly is made out of brass sheets. The flare–angle of the horn is $120°$. The dimensions of the feed are:

- Aperture diameter : $3.65\,\lambda$
- Horn length : $4.48\,\lambda$

The phase center has been found out to be at the apex of the cone - at a depth of 200 mm from the aperture plane. This feed has an impressive bandwidth: 580 MHz, starting from 1000 MHz to 1580 MHz, as can be seen from Fig 19.15

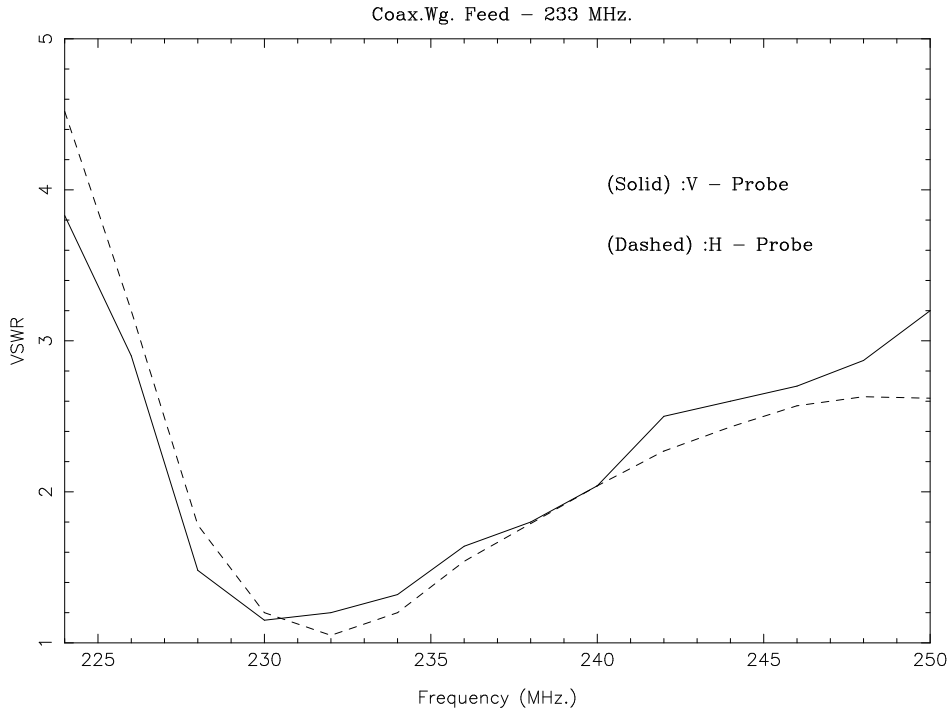


Figure 19.14: The VSWR as a function of frequency for the 233 MHz feed.

Radiation patterns, including the cross-polar pattern is shown in Fig 19.16.

The edge taper is -19 dB and the cross-polar peak is -24 dB. The front-end electronics is housed in a rectangular box,on the back side of the horn, forming one integral unit. The entire band is divided into 4 subbands, each 140 MHz wide and centered on 1390, 1280, 1170 and 1060 MHz. There is also a bypass mode in which the entire bandwidth is available.

19.8 GMRT Antenna Efficiencies

The efficiency relations shown in Section 19.5, have not considered the effect of aperture blockage by feeds and feed-support frames ("*quadripod legs*" in GMRT–parlance). Simple geometrical optics based models for such computation exist,[16] which are used along with GMRT–specific efficiency relations, to produce the following table. Limitations of this model are highlighted in [17].

Some of the loss terms can be expressed as equivalent noise temperatures (see Chapter 3). The spillover temperature is given by (see also Eqn 19.4.14)

$$T_{Sp} = T_g . \frac{\int_{\theta_0}^{\pi/2} |E(\theta)|^2 \sin(\theta) d\theta}{\int_0^{\pi} |E(\theta)|^2 \sin(\theta) d\theta} \tag{19.8.25}$$

where T_g is the ground temperature. Considering the reflectance of soil at microwave frequencies, it is presumed as 259° K.

Similarly, the mesh–leakage T_{ml}, scattered radiation by the feed– support frames T_{sc}, can also be expressed in terms of T_g. The overall system temperature (see Chapter 3) is

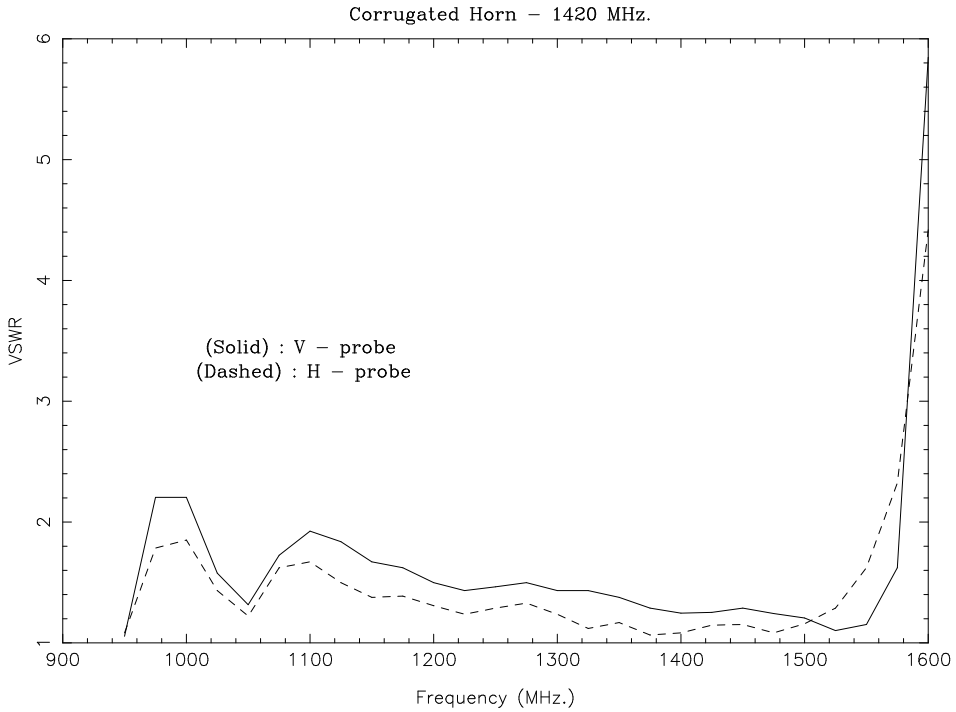


Figure 19.15: The VSWR as a function of frequency for the 1420 MHz feed.

the sum of all these and the receiver noise temperature, T_r and the sky temperature,T_{sky}, which is assumed to be,

$$T_{sky} = 3.0 + 20{\cdot}(408/f)^{2.75}, \tag{19.8.26}$$

where f is the frequency of the received signal (in MHz). Hence

$$T_{sys} = T_r + T_{sky} + T_{Sp} + T_{ml} + T_{sc}. \tag{19.8.27}$$

Finally the figure–of–merit of any radio antenna, is the gain-by-system temperature ratio, G/T_{sys}, expressed as :

$$G = \frac{SA_p\eta_a}{2k}, \tag{19.8.28}$$

where S is flux density in units of Jansky, A_p, is the physical area of the parabolic dish and η_a is the overall aperture efficiency. For a 1 Jy. source at the beam of the antenna and value of Boltzmann's constant k included in the above relation,

$$G = \frac{A_p\eta_a}{2760}. \tag{19.8.29}$$

Hence, the ratio G/T_{sys} is expressed in units of Jy^{-1}.

A summary of the relevant parameters for the GMRT antennas is given in Table 19.4. These have been computed based on the following assumptions.

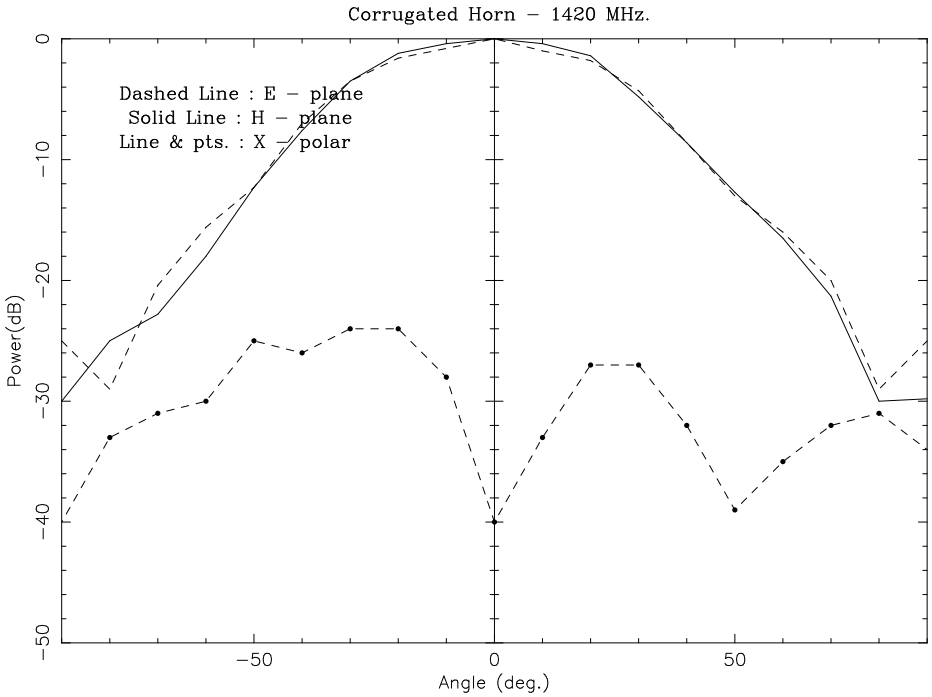


Figure 19.16: Radiation Pattern of the 1420 MHz feed.

1. $T_r = 100^\circ$ K for 150,233 and 327 MHz bands; 50° K for 610 MHz and 35° K for the 1000 to 1400 MHz bands.

2. The surface rms, $\sigma_1,\ \sigma_2,\ \sigma_3$ values are 8.0, 9.0, and 14.0 mm respectively.

The agreement between the observed HPBW, gain and system temperature and the computed values is in general quite good.

19.9 Further Reading

1. Swarup, G., Ananthakrishnan, S., Kapahi, V.K., Rao, A.P., Subrahmanya, C.R., and Kulkarni, V.K., 'The Giant Metre-wave Radio Telescope', Current Science, Vol.**60**, No.2, (1991).

2. Chengalur, J.N., Kher, R.K. 'Transmission through Wire Grids', TIFR Internal Report, 1987.

3. Kaplun, V.A., Babkin, N.I., Goryachev, B.G., 'Shielding Propeties of Wire Screens at SHF', Radio Engg. & Elec. Physics, **9**, (1964), pp.1428–30.

4. Clarricoats, P.J.B., and Olver, A.D., Corrugated horns for Microwave Antennas, Peter Peregrinus Ltd., London, (1984)

5. Stutzmann, W.L., and Thiele, G.A., Antenna Theory & Design, John Wiley & Sons. Inc., (1981)

	Frequency (MHz)						
Eff.	150	233	327	610	1000	1200	1400
Tap Eff.	0.689	0.823	0.715	0.775	0.566	0.533	0.592
Spill. Eff.	0.952	0.799	0.944	0.835	0.967	0.971	0.971
Mesh Eff.	0.999	0.999	0.998	0.991	0.943	0.941	0.94
RMS Eff.	0.997	0.992	0.986	0.948	0.88	0.835	0.78
Aper. Eff.	0.652	0.651	0.664	0.608	0.452	0.405	0.422
Tsys($^\circ$ K)	428	229	152	92	65	77	62
$\frac{G}{T_{sys}} \times 10^{-3}$	0.877	1.64	2.53	3.81	4.04	3.02	3.17
$HPBW$	2° 52'39"	1° 51'06"	1° 21'15"	0° 42'48"		0° 19'26"	

Table 19.4: Calculated aperture efficiencies and system temperatures for the GMRT antennas.

6. Kildal, P-S., "Factorization of the Feed Efficiency of Paraboloids and Cassegrain Antennas", IEEE Trans.on Ant.& Propg., Vol.**AP-33**, No.8, (1985).

7. Krishnan, T., "Analysis of TIFR / GMRT 45 m. dish performance at 327 MHz.", HAE/HSS Report 010/92, Sept.1992.

8. Sankar, G., Swarup, G., Ananthakrishnan.S., Sankararaman, M.R., Sureshkumar .S. and Izhak.S.M. "Prime focus feeds for GMRT Antenna"IAU - 6th Asian Pacific Regional Meeting on Astronomy, IUCAA, Pune. Aug.1993.

9. Guillou.L., Daniel.J-P., Terret.C., Madani.A., "Rayonnement d'un Doublet Replie Epais", Annales des Telecommunications, tome 30, nr 1-2.

10. Kildal, P-S., and Skyttemyr, S.A., "Dipole-Disk Antenna with Beam-Forming Ring", IEEE Trans.on Ant.& Propg., Vol.**AP-30**, No.4, (1982).

11. Livingston, M.L., "Multi-frequency Coaxial Cavity Apex Feeds", Microwave Journal, (Oct.1979) pp.51–54.

12. Van Ardanne, A., Bregman, J.D., Sondaar, L.H., Knoben, M.H.M., "A Compact Dual Polarized Coaxial Feed at 327 MHz.", Electronics Letters, Vol.**20**, No.18, (1984) pp.723–724.

13. Jeuken, M.E.J., Knoben, M.H.M, and Wellington, K.J., "A Dual Frequency, Dual Polarized Feed for Radioastronomical Applications", Rechnernetze und Nachrichtenverkehrstheorie, NTZ, Heft:**8**, (1972) pp.374–376.

14. Shafai, L. and Kishk, A.A., "Coaxial Waveguides as Primary Feeds for Reflector Antennas and comparison with Circular Waveguides", AEÜ, Band:**39**, Heft 1, (1985) pp.8–14.

15. Sankar, G. and Praveenkumar, A., "Dual Frequency Coaxial Waveguide Feed - Design calculations", Int.Tech.Report, **AG-02/90**, GMRT–TIFR, Pune.

16. Fisher, J.R., "Prime-focus Efficiency, Blockage, Spillover and Scattering Calculations on the HP 9825A Calculator", EDIN Report.174, NRAO, Nov.1976.

17. Chengalur, J.N. "Aperture Efficiency Calculations for GMRT Dishes", NCRA-TIFR Int.Tech.Report, Dec.1993.

Chapter 20

The GMRT Servo System

V. Hotkar

20.1 Introduction

The GMRT servo system is a dual drive position feed-back control system. It can track a source in the sky with an rms accuracy of $\sim 0.5'$. To realise such a system practically, the expertise from various engineering disciplines are put to work. In order to understand such a system, one has to become familiar with the theory of feedback control systems as well as its application for position control. This chapter discusses these issues. The material is presented in an simplified form and an effort has been made to use, wherever possible, graphical explanations instead of a mathematical treatment.

20.2 Objectives of the GMRT Servo System

The servo systems used for position control of the radio telescopes must meet following objectives.

1. Ability to point anywhere in the sky.
2. High pointing & tracking accuracy.
3. Able to accelerate rapidly in the direction of source.
4. Able to manoeuver remotely

The first requirement is met by making a two axes mount for the antenna. For large antennas like those used in the GMRT (i.e. with weight in excess of 80 tones) an alt-azimuth mount is preferred. In such a mount the antenna can be moved in two axes viz. azimuth and elevation. The azimuth axis movement is parallel to the horizon, while elevation axis movement is normal to the horizon. Alt-az mounts are mechanically simple, yet very stable.

Radio telescope antennas are required to point within +/- 10HPBW at any given wavelength of operation of the antenna. This means that the pointing accuracy of the antenna should be fairly good. The following issues are of concern when trying to achieve high accuracy pointing or tracking:

1. Structural deformation due to gravity.
2. Structural vibrations/deformations due to wind forces.
3. Servo positioning error.

Note that not only can the reflecting surface of the antenna be affected by gravity, the feed support legs too could deform, leading to a displacement of the feed from the focus of the antenna. The GMRT antennas are built using a novel technique (nicknamed "SMART") involving a stainless steel mesh which is attached to rope trusses by wires which are are tensed appropriately in order to achieve the desired parabolic reflecting surface. This results in a dramatic reduction in the gravitational and wind loading on the structure, as well as in the total weight of the dish.

20.3 The GMRT Servo System Specifications

A summary of the GMRT Servo Specifications is given in Table 20.1.

Table 20.1: GMRT Servo system summary

Dish mount	Altitude-Azimuth mount.
Drive	Dual drive in counter torquing mode.
Dish movement	Azimuth +270 to −270 deg.
	Elevation 15 deg to 110 deg.
Dish slewing speed	Azimuth 30 deg/min.
	Elevation 20 deg/min.
Minimum Tracking speed	Azimuth 5 arcmin/min.
	Elevation 5 arcmin/min.
Maximum Tracking speed	Azimuth 150 arcmin/min.
	Elevation 15 arcmin/min.
Tracking & pointing accuracy	1 arcmin for wind speed $<$20 kmph.
Gear reduction ratio	Azimuth 18963.
	Elevation 25162.
Antenna acceleration	Full speed in $\geq$ 3 sec for both axes.
Design Wind speed	40 kmph Operational.
	80 kmph Parking.
	133 kmph survival.
System operating voltage	415 VAC, 3 Phase, 50 Hz.
Antenna parking	Antenna parking using 96 V DC battery.

20.4 Control System Description

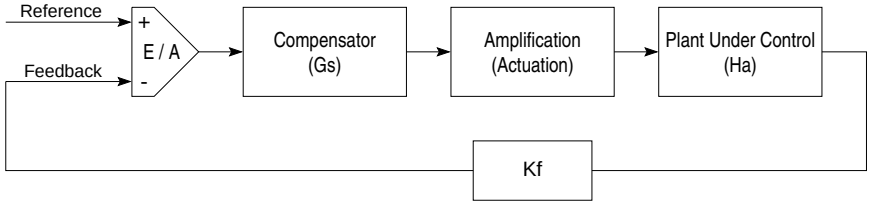
The GMRT servo system is a closed loop position feed-back control system, designed for tracking & positioning of the GMRT radio telescopes. The use of dual drive and counter-torque, eliminates non-linearity due to back-lash associated with the gearbox.

20.4.1 Closed Loop Control Systems

All automatic control systems use −ve feedback for controlling a physical parameter like position, velocity, torque etc. The parameter which has to be controlled is sensed by a

suitable transducers and fed back to the input, for comparison with the reference value (see Figure 20.1. This subtraction of the sampled output signal with that of reference input is called as −ve feedback. The difference signal, called the "error" is then amplified to drive the system (referred to as actuation) in such a manner that the output approaches the set reference value. In other words the system is designed to minimize the error signal.

All practical loads have inertia and spring constants due to which there is a delay in actuation. Hence, even though a system may be designed for −ve feedback, due to inherent time lags, the feedback may turn into +ve feedback at certain frequencies. If the loop gain is more than unity at some frequency at which the feedback is +ve, the system will oscillate. Hence, in designing control systems great care has to be taken to avoid such situations.

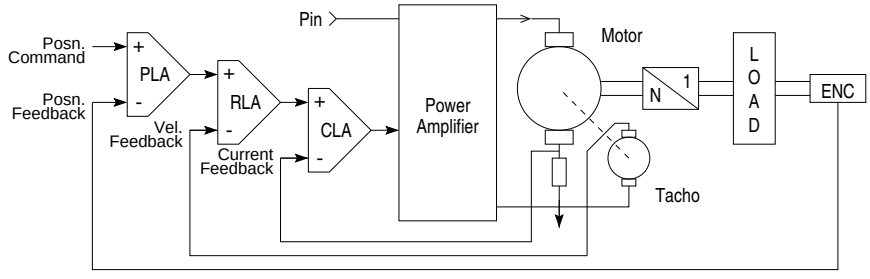


(Gs) = Compensator transfer function
Hs = Plant transfer function

Figure 20.1: Closed loop control system.

20.4.2 Principles of Position Control

For controlling a heavy load, one could, (as illustrated in Figure 20.2) use three nested feedback loops viz. a position loop, a velocity loop and a current loop. This configuration allows independent tunning of the loop parameters without affecting the adjacent loop. A current amplifier is used to amplify the current for driving the motor. The position is sensed by a suitable transducer. The velocity of the antenna is generally sensed by the tachometers mounted on the motor shaft.



GB = Gear box. ENC = Encoder. PLA = Position Loop Amplifier. RLA = Rate Loop Amplifier.

Figure 20.2: Three nested feed back loop.

The block diagram shown above can not be directly used in all position control applications. The back-lash which is inherent in any gear box, introduces a non-linearity in the position loop. Such a system exhibits a phenomena called as "limit cycle hunting". This affects the positioning accuracy of the antenna.

20.4.3 Position Loop Amplifier

The position loop amplifier (PLA) has two inputs viz. command input and feedback input. In an automatic position control system, the output of the position sensor is filtered, scaled and then applied to the PLA. The command signal is applied to the other input of the PLA. The PLA (which can be either analog or digital) subtracts its two inputs to generate an error signal. This error signal is then applied to the compensator.

A compensator is designed depending on the application. For example the GMRT antennas are used for tracking of stellar radio sources which are moving at constant speed in the sky (15^o/hr, the speed of the earth's rotation). For such an application, a position system having type II response is required. With a type I position compensator and with the use of rate loop in the position control, the overall system response is of type II .

Type of position system	Pointing Error	Tracking Error
Type O	Finite	Finite
Type I	Zero	Finite
Type II	Zero	Zero

Parameters like the structural natural resonant frequency (Wc) and the frictional (Bc) constants of the structure are required for the design of the position loop compensator . The main objective while designing the position compensator is that it should offer enough attenuation at the natural resonant frequency of the structure.

The output of the PLA acts as velocity command. If the target's angular position is far removed from the current position, then the error is very large and could saturate the PLA . The saturation of the PLA is considered as a fixed velocity command to the rate loop. The rate loop moves the antenna with a constant velocity towards the target position. As the antenna approaches the target position, the error at the output of the PLA goes on reducing, which commands the rate loop to reduce the speed of the antenna. When the antenna is at the target position the error at the output of the PLA goes to zero, which translates to a zero speed command to the rate loop. The sign of the error signal at the output of the PLA decides whether the antenna is to be moved forward or reverse .

20.4.4 Rate Loop Amplifier

The function of the Rate Loop Amplifier (RLA) is to control the velocity of the antenna. In position control applications, the rate loop improves the transient response of the position loop by adding a pole in the position loop.

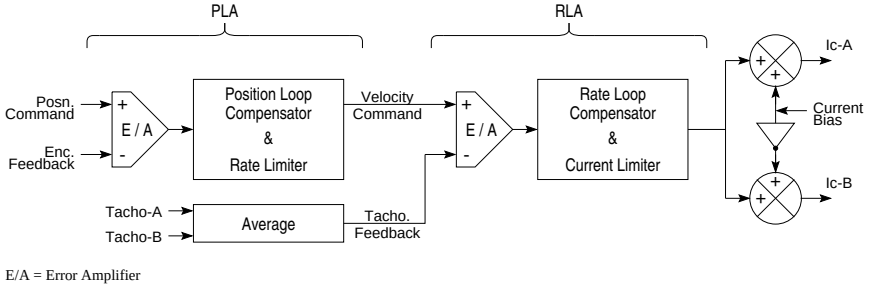


Figure 20.3: Rate loop amplifier.

The output of the PLA which acts as a velocity command, is applied to the one input while tachometer signal is applied to the other input. The RLA subtracts both the input signals and generates an error signal which is then applied to the compensator. For

position control applications like the GMRT the rate loop compensator can be of phase flag type (Type O) which avoids limit cycle hunting. The electro-mechanical time constant of the combined motor and load determines the bandwidth of the compensator. The output of the RLA acts as a command to the current loop. If the command speed is more than the actual speed, then the error at the output of the RLA becomes large, which commands the current loop to pass more current through the motor.

For GMRT antennas, where a dual drive system is used, the rate loop controls the antenna velocity by sensing the tacho signal from both the motors. Both these tacho signals are averaged and then applied to RLA as feedback. A voltage corresponding to torque bias is added/subtracted at the output of the rate loop, to generate two current commands. These two current commands are applied to the two current loop amplifiers, for controlling currents in accordance with the rate loop.

20.4.5 Current Loop Amplifier

The function of the Current Loop Amplifier (CLA) is to control/regulate the current of the motor which results in the control of the motor torque. The current of the motor is sensed either by a resistive shunt or with a Hall effect sensor. The control of over current should be fast in order to protect the power semiconductors during starting/stopping of the motor or in the event of fault. Also the steady state error of the current should be zero (as any error in torque affects the speed). These requirements can be met by using a "PI" (Proportional Integral) compensator.

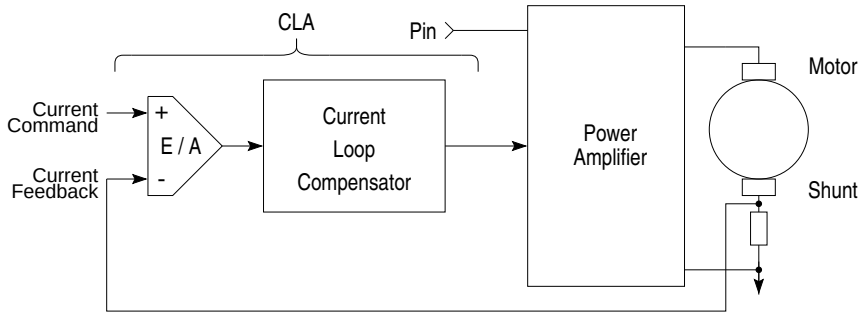


Figure 20.4: Current loop amplifier.

The current signal is filtered, scaled and then applied to the CLA. The output of the RLA which acts as a current command, is applied to the other input. The CLA subtracts both the input signals and generates the error signal. The error signal is applied to the proportional-integral (PI) compensator. In a 3-phase SCR amplifier like one used at the GMRT, the motor current has a 150 Hz component along with the DC component. As the current is sampled and fed back to the loop amplifier, the 150 Hz component of the current gets injected into the loop. This is like injecting a noise into a system. In order to avoid oscillations in the loop, the current loop compensator is designed to heavily attenuate the 150 Hz signal component. The filtered output of the error amplifier is applied to the 4-quadrant power amplifier.

20.5 Servo Amplifiers

Servo amplifiers are 4-quadrant, regenerative power amplifiers, supplying appropriate power to the motor as commanded by a control voltage. These amplifiers are capable of

Table 20.2: Servo amplifier specifications

Type	3-Phase, SCR based, 4-quadrant fully regenerative.
Control Type	Phase angle control with current loop.
Input Volts	275VAC L-L, 50 Hz, 3-Phase, 4-wire.
Command Volts	+/- 10 Volt.
Maximum Current	+/- 80 Amp.
Protection	Over current & over speed.

suppling energy to the load, as well as absorbing energy from the load. They are designed to convert the kinetic energy of the combined motor load, into electrical energy while the load is decelerating.

The GMRT servo amplifier is a three phase, half wave, four-quadrant, fully regenerative, SCR CLA for the control of permanent magnet DC brush type motors. A CLA is a device, which keeps the current through the motor proportional to a commanded input signal.

20.6 Servo Motors

Servo motors are special category of motors, designed for applications involving position control, velocity control and torque control. These motors are special in the following ways:

1. Lower mechanical time constant.
2. Lower electrical time constant.
3. Permanent magnet of high flux density to generate the field.
4. Fail-safe electro-mechanical brakes.

For applications where the load is to be rapidly accelerated or decelerated frequently, the electrical and mechanical time constants of the motor plays an important role. The mechanical time constants in these motors are reduced by reducing the rotor inertia. Hence the rotor of these motors have an elongated structure. For DC brush type motors, the permanent magnets are mounted on the stator, while the armature conductors are on the rotor. The rotating conductors make contact with the stationary electrical source via a brush-commutator assembly. A DC tacho is mounted on the motor shaft, for indicating the shaft speed in-terms of a voltage. These motors also come with fail-safe electro-mechanical brakes. In the event of failure of the utility mains, the antennas are stopped by these brakes.

20.7 Gear Reducers

Generally the motors which are commercially available deliver low torque at high speed and can not be used for driving the load directly. Gear reducers are used to increase the torque so as to meet the torque demand of the load . For servo application i.e. for positioning the load, the gear reducers should possess following characteristics.

1. Bi-directional energy flow

Table 20.3: Servo motor specifications

Type	DC brush type, permanent magnet field.
Horse Power rating	6 HP.
Rated motor voltage	150 V (DC).
Rated motor current	80 Amp (Continuous).
Rated motor speed	2250 rpm.
Continuous stall torque	47 N-m.
Peak Torque	111 N-m.
Torque Sensitivity	0.56 N-m / Amp.
Back E.M.F. Constant	59 V / krpm.
Armature resistance	0.045 Ohm.
Armature inductance	0.33 mH.
Tacho sensitivity	17 V / krpm.

2. Low back-lash

3. Low moment of inertia

4. High efficiency

The bi-directional reducers means that, the energy can be transferred from input to output as well as from output to input. During deceleration, the motor is forced to act like a generator, converting the kinetic energy of the load into electrical energy. The deceleration of the load is decided by the rate of consumption of the electrical energy produced. Planetary gear boxes meets this requirement and are hence used at the GMRT.

20.8 Position Sensors

Optical position sensors are the sensors of choice for highly accurate positioning of antennas. There are two broad styles of the encoders viz. incremental and absolute. An incremental encoder is made of a glass disc and a light interrupter. Transparent and opaque markings are put on the outer periphery of the glass disc. Light emitted from a lamp or LED is interrupted by the glass disc and received by a photo diode. As the disk rotates, the light falling on the photo detector is interrupted by the opaque markings, leading to pulses in the photdetector. These pulses are counted to determine the change in position. The disk has an index marker, is used to provide a reference. Though incremental encoders are simple in construction and provide a cheap solution for position sensing, they suffer from one drawback. On the failure of the power to the encoder or the electronic circuit, the electronic counter looses its count value, and hence all information as to the current position. Hence, upon the resumption of the power to the antenna, one would need to move the antenna until the index marker pulse is received, a procedure called “homing”. For large antennas like those at the GMRT, this is unacceptable and hence absolute encoders have to be used.

In an absolute encoder, a pattern corresponding to a gray code is printed on the glass disc. The glass disc moves through a light emitter and a set of light detectors. The number of light detectors are in proportion with the number of bits of the encoders. This enables the encoder to generate a binary word corresponding to the angular position of its shaft. The electronics housed inside the encoder converts the gray code to the natural

Table 20.4: Encoder specifications.

Type	Optical, absolute shaft encoder.
Resolution	17 bit (10 arcsec).
Max . Shaft speed	600 rpm.
Max. Data rate update	100 kHz.
Illumination	light emitting diode.
Input Power	+ 5V DC at 300 mA.
Output Code	Natural binary.
Output data format	Serial.
Data transmission	RS -422 differential line driver.
Serial output	MSB first, LSB last & then parity bit.
Count Direction	CW increasing.
Operating temp.	0° C to +70° C.

binary . Also the parallel code gets converted into serial format for transmitting over long distance cable. The encoder is directly mounted on each axis of an antenna.

20.9 Dual Drive

For a large antenna, the torque required to move the antenna is high, hence the large ratio gear reducers are used to meet the required torque demand. It is almost impossible in practice to manufacture a gear box which can deliver a large power with no back-lash. Any effort to reduce back-lash by tight coupling of pinions increases the friction of the gear box which reduces its efficiency. With the use of large gear ratios the backlash, hysteresis, and between the motor shaft and the load shaft increases. With the increase in these parameters the nonlinearity in the position loop increases, which leads to position loop instability. There are various ways to reduce the back-lash mechanically but they are inefficient and are unsuitable for a giant antennas like those at the GMRT. Instead one uses a dual drive. Here a pair of motors, gearbox and pinion are used to drive the common load.

Two amplifiers individually drive the motors. When the load is to be held at some position, the torque produced by two motors are equal and opposite, thereby eliminating the backlash. The net torque on the load is zero hence it does not move. For a slight movement of the load in a given direction, one motor increases its torque in that direction while the other reduces its torque. The load will be subjected to a net torque which causes small movement of the load.

20.10 Digital Controller

The digital controller for GMRT antennas, is built around Intel's 8086 processor running at 8 MHz and is called as the "Station Servo Computer" The 8086 is a bus master, controlling two slave processors 8031, for analog and encoder interface. The position loop of both the axes of the GMRT servo system is implemented digitally in this servo computer. The elevation and azimuth axes angles along with time, are fed to the servo computer by the antenna base computer (ABC, see Section 24.2.4). The servo computer computes the error of both the axes and performs necessary filtering (compensation). The compensator

output is converted into analog signal by using 16 bit DAC and then applied to the rate loop.

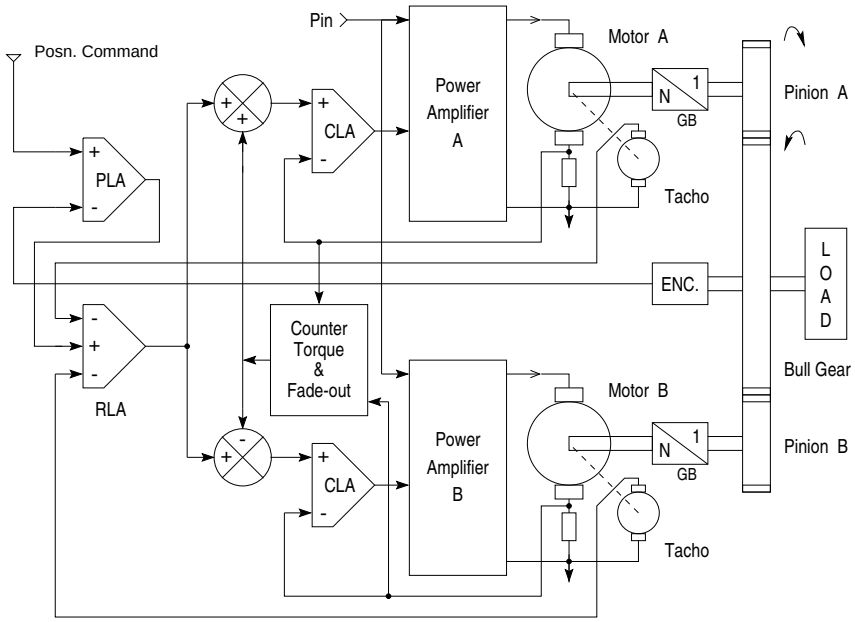


Figure 20.5: Dual drive position control system.

For the digital implementation of a position loop, the sampling rate must be large enough. The "S" domain transfer function of the compensator is converted into a "Z" domain transfer function, by using the "Tustins approximations". The Z-domain transfer function is further converted into a difference equation, to be solve recursively at a regular interval. Tustin proposes that the sampling frequency must be greater than 10 times the compensator bandwidth. With 1.5 Hz as a structural resonant frequency of the GMRT antennas, the position loop bandwidth can be around 0.4 Hz to 0.5 Hz . For a 0.5 Hz loop bandwidth the sampling rate should be more than 5 Hz. This sets the lower limit of the sampling rate. The upper limit of the sampling rate is determined by the processor speed, other tasks of the processor, the transport lag etc. We have chosen 10 Hz as a sampling rate. The processor is interrupted at regular interval of 100 ms to run the real time programme.

20.11 Servo Operational Commands

The central control station sends commands to a group of antennas via an optical fiber link (see Chapter 24). Some of the operational commands, related to the servo is described next.

1. COLDSTART: On receiving this command, the servo system removes the stow-lock pins, releases the motor brakes, enables the servo amplifiers, holds both the axes at the current angle and waits for next command.

2. MV arg1,arg2: Move along the azimuth and elevation axes to the angles arg1 and arg2 respectively. The servo system releases the motor and moves the antenna.

3. TRACK arg1,arg2,arg3: Track in azimuth and elevation axes with the destination angle as arg1 and arg2 and the time parameter as arg3.
4. HOLD: Holds both the axes. On receiving this command, servo system releases brakes of both axes motors and holds the antenna in position.
5. STOP: Stops both the axes. On receiving this command, servo system disables amplifiers & applies brakes to both axes motors.
6. CLOSE: Close the observations. On receiving this command, servo system positions the elevation axis to 90:00:00 deg., disables all amplifiers, applies brakes to all motors & inserts the stow-lock pin.
7. STOW: Inserts the stow-in pin in the elevation axis and locks the axis.
8. SWRELE: Releases stow-in pin from the elevation axis and frees the axis.
9. RSTSERVO: Resets the station servo computer.

Chapter 21

GMRT Receivers

Praveen Kumar

21.1 Introduction

This chapter discusses the GMRT receiver system chain. This chain starts from the Multi-frequency RF Front-Ends and ends at the Baseband system. The major blocks in this chain along with their various possible configurations are described.

A detailed analysis of the noise contributed by the various components of this chain is presented. The length of the fiber optic cables linking the antennas to the CEB varies from about 600m for the nearest antennas to about 21km for the most distant ones. Since the transmission loss increases with increasing fiber length, different antenna systems will have different signal to noise ratios at the CEB. However by optimally adjusting the operating power levels at different points of the receiver chain one can ensure that the maximum degradation of the system noise temperature is less than 1% for all antennas.

21.2 Overview of the GMRT Receiver Chain

The GMRT receiver chain is shown schematically in Figure 21.1. The first block is the multi-frequency front end. This is located in a rotating turret at the prime focus. All the feeds and low noise RF front-ends have been configured to receive dual polarization signals. Lower frequency bands (from 50 to 610 MHz) have dual circular polarization channels, i.e. left hand circular and right hand circular polarizations which have been labelled as CH1 and CH2 respectively. The L-band (1000-1450 MHz) system has dual linear polarization channels, i.e. vertical and horizontal polarizations (also labelled CH1 and CH2 respectively).

The first local oscillator (I LO, situated at the base of the antenna, inside a shielded room) converts the RF band to an IF band centered at 70 MHz. After passing the signal through a bandpass filter of selectable bandwidth, the IF at 70 MHz is then translated (using II LO) to a second IF at 130 MHz and 175 MHz for CH1 and CH2, respectively. The maximum bandwidth available at this stage is 32 MHz for each channel. This frequency translation is done so that signals for both polarizations can be frequency division multiplexed onto the same fiber for transmission to the CEB.

At the CEB, these signals are received by the Fiber-Optic Receiver and the 130 and 175 MHz signals are then separated out and sent for base band conversion. The baseband converter section converts the 130 and 175 MHz IF signals first to 70 MHz IF (using III

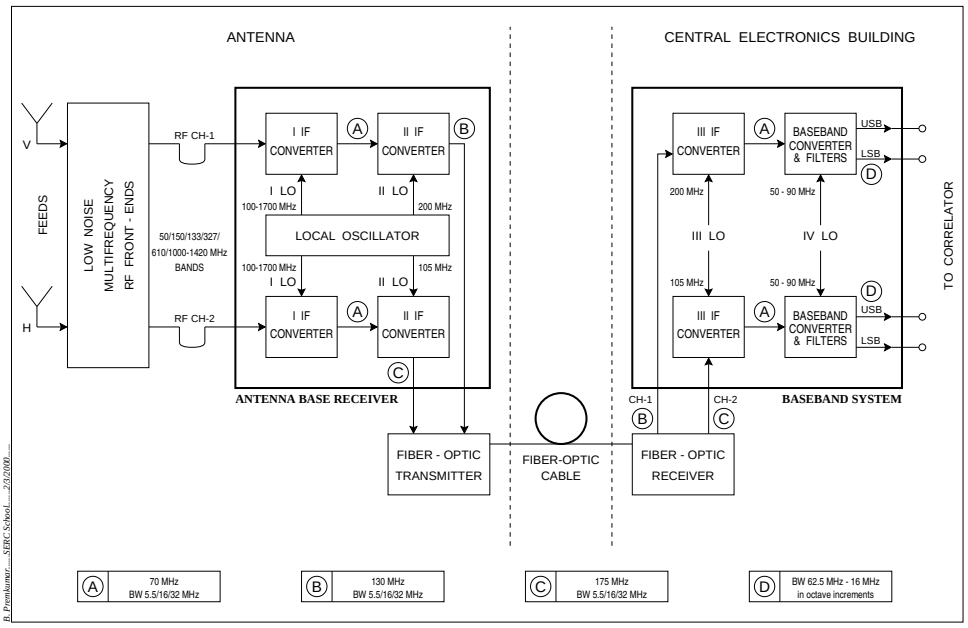


Figure 21.1: Schematic block diagram of the GMRT receiver. See the text for more information.

LO), these are then converted to upper and lower side bands (each at most 16 MHz wide) at 0 MHz using a tunable IV LO. The various local oscillators and baseband system are discussed in more detail in Chapter 23. There are also two Automatic Level Controllers (ALCs) in the receiver chain (not shown in Figure 21.1 but discussed in more detail below). The first is just before the Fiber Optic transmitter and the second is at the output of the baseband unit.

21.3 Receiver Design Considerations

Each of the various blocks in the receiver chain has some gain (or loss) associated with it. The receiver chain hence has distributed gain. There are several considerations involved in determining exactly how to distribute the gain across the RF, IF and BB electronics, viz.

1. The response of the entire system must remain linear over a wide range of noise temperatures from cold sky to the high antenna temperatures anticipated when observing strong sources like the Sun.

2. The entire receiver system should remain linear even in the presence of strong interference signals. In particular the inter-modulation distortion (IMD) products should be below a critical threshold[1]. Also the receiver should have a high desensitization

[1]Basically one needs receiver with high enough Compression and Spurious Free Dynamic Range (CDR and SFDR) to handle the range of astronomical signals and interference signals present. In communications receiver

dynamic range[2] so that a single dominant out of band interfering signal does not reduce the receiver SNR by saturating the subsystems in the receiver.

3. The RF Front End gain should be such that no more than 1 K noise is added to the Low Noise Amplifier (LNA) input noise temperature by the rest of the receiver chain.

4. The gain should be so distributed that no more than 1% gain compression should occur at any stage of the receiver chain.

5. The level of signals at the input of the cables that run from antenna turret to the base of antenna should be sufficiently high compared to any extraneous interference signals that might be picked by these cables.

6. Components whose contribution to the signal phase needs to be kept constant should preferably be located at the antenna base room where the temperatures are relatively stable compared to that at the prime focus.

7. Internally-generated spurious products (if any) in the receiver, must be very low compared to the receiver noise floor.

8. The Antenna Base Receiver (ABR) input (which receives the the RF signals from the front end through long lengths (about 100 m of cable) should be well matched for the full RF band i.e. 10 MHz to 1600 MHz. A poor match would result in passband ripples.

9. The receiver should have a good image rejection (at least 25 dB). Further since the RF pass band in the common box electronics (see below) has 10 MHz - 2000 MHz coverage, a 70 MHz signal may find a path past the amplifiers and mixer and be coupled into the 70 MHz IF circuitry. The units have to be optimally configured such that a good IF rejection[3] is achieved.

10. The ALCs should be active over a large signal amplitude range.

21.4 The Multi Frequency Front Ends

A block diagram of the Multi Frequency Front Ends is given in Figure 21.2. There are six possible observing bands centered at 50 MHz[4], 150 MHz, 233 MHz, 327 MHz, 610 MHz and an L-band extending from 1000 to 1450 MHz[5]. The L-band is split into four sub bands centered at 1060 MHz, 1170 MHz, 1280 MHz and 1390 MHz, each with a bandwidth of 120 MHz. The L-band receiver also has a bypass mode in which the entire RF band can be brought down to the ABR.[6] The 150 MHz, 233 MHz, 327 MHz and 610 MHz

parlance, the SFDR is defined as the power ratio between the receiver thermal noise floor and the two tone signal level that will produce third order IMD products equal to the noise floor level. The CDR is defined as the power ratio between the receiver thermal noise floor and the 1 dB compression point. However, for radio astronomical receivers it is customary to define the upper limit for the CDR as the signal level where 1% gain compression occurs and in the case of SFDR, the upper limit as the two tone signal levels which produce IMD products 20 dB below the noise floor.

[2]The desensitization dynamic range is defined as the power ratio between the level of the strong undesired signal which reduces the SNR by 1 dB and the receiver noise floor.

[3]IF rejection is a measure of attenuation between the receiver input and the IF circuit.

[4]The 50 MHz feed is as of yet not commissioned.

[5]Some of the L-band feeds have coverage up to 1750 MHz to allow observations of the OH molecular lines.

[6]This mode is useful in for example making observations at frequencies below 1000 MHz the nominal bottom of the L band.

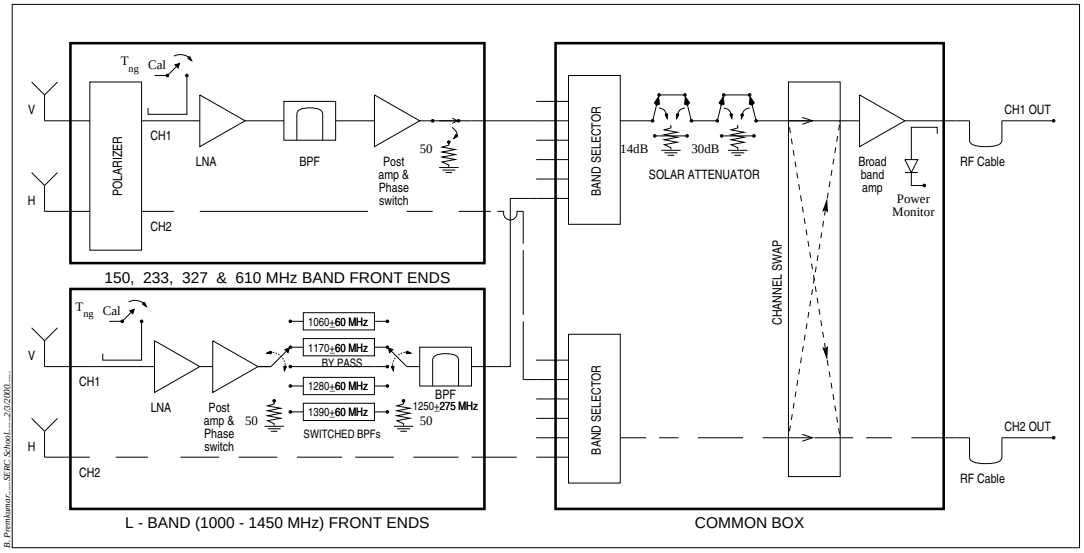


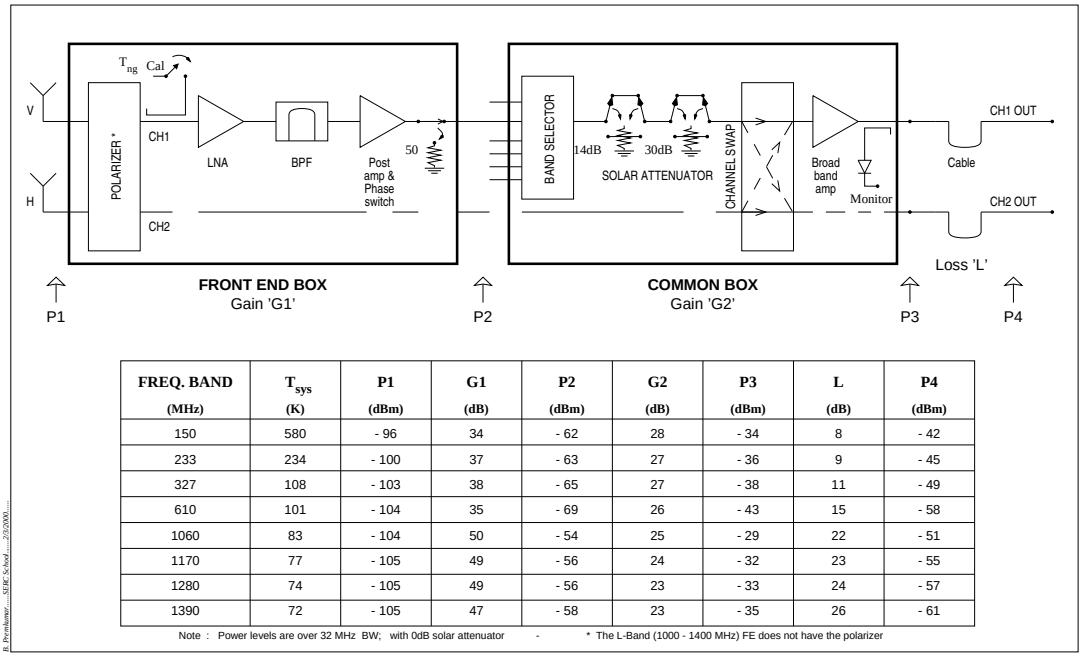
Figure 21.2: Block diagram of the GMRT multi-frequency front end. For illustrative purposes, the upper left side block has been shown as any of the 50 MHz to 610 MHz RF receivers, the lower block is an L band RF receiver. See the text for more information.

bands have a nominal bandwidth of 40 MHz[7]. Table 21.1 gives a break up of the contribution from various sources to the system noise temperatures at the different frequency bands. Figure 21.3 gives the expected RF power (for a 32 MHz bandwidth) at different stages of the multi-frequency front end.

At the lower frequencies (50 MHz to 610 MHz) there is a polarizer before the LNA which converts the received linear polarization to circular. At L band, in order to keep the system temperature low, this element is not inserted into the signal path, and the linear polarized signals are fed directly to the LNA. To calibrate the gain of the receiver chain, it is possible to inject an additional noise signal (of known strength) into the input of the LNA. It is possible to inject noise at any one of four levels. These are called Low cal, Medium cal, High cal and Extra high cal and are of monotonically increasing strength. To minimize crosstalk between different signals a phase switching facility using separate Walsh functions for each signal path is available at the RF section of the receiver. It is also possible to connect a 50 ohm resistor instead of the LNA to the output of the RF receiver. This is primarily of use in debugging.

At all bands the signals go through one additional stage of amplification in the 'Common Box'. The common box has a broad band amplifier which covers the entire frequency range of the GMRT. The band selector in the common box can be configured to take signals from any one of the six RF amplifiers. The common box (and the entire receiver system) has the flexibility to be configured for receiving either both polarizations at a single frequency band or a single polarization at each of two different frequency bands. It is also possible to swap the polarization channels (i.e. to make LCP come on CH2 instead of CH1 and RCP on CH1 instead of CH2), this is primarily for debugging use. For observing strong radio sources like Sun, solar attenuators of 14 dB, 30 dB or 44 dB are available in the common box. In addition there is a power monitor whose output can be continuously monitored to verify the health of the subsystems upstream of the common box.

[7]But the usable bandwidth is often somewhat larger.



FREQ. BAND (MHz)	T_{sys} (K)	P1 (dBm)	G1 (dB)	P2 (dBm)	G2 (dB)	P3 (dBm)	L (dB)	P4 (dBm)
150	580	- 96	34	- 62	28	- 34	8	- 42
233	234	- 100	37	- 63	27	- 36	9	- 45
327	108	- 103	38	- 65	27	- 38	11	- 49
610	101	- 104	35	- 69	26	- 43	15	- 58
1060	83	- 104	50	- 54	25	- 29	22	- 51
1170	77	- 105	49	- 56	24	- 32	23	- 55
1280	74	- 105	49	- 56	23	- 33	24	- 57
1390	72	- 105	47	- 58	23	- 35	26	- 61

Note : Power levels are over 32 MHz BW; with 0dB solar attenuator - * The L-Band (1000 - 1400 MHz) FE does not have the polarizer

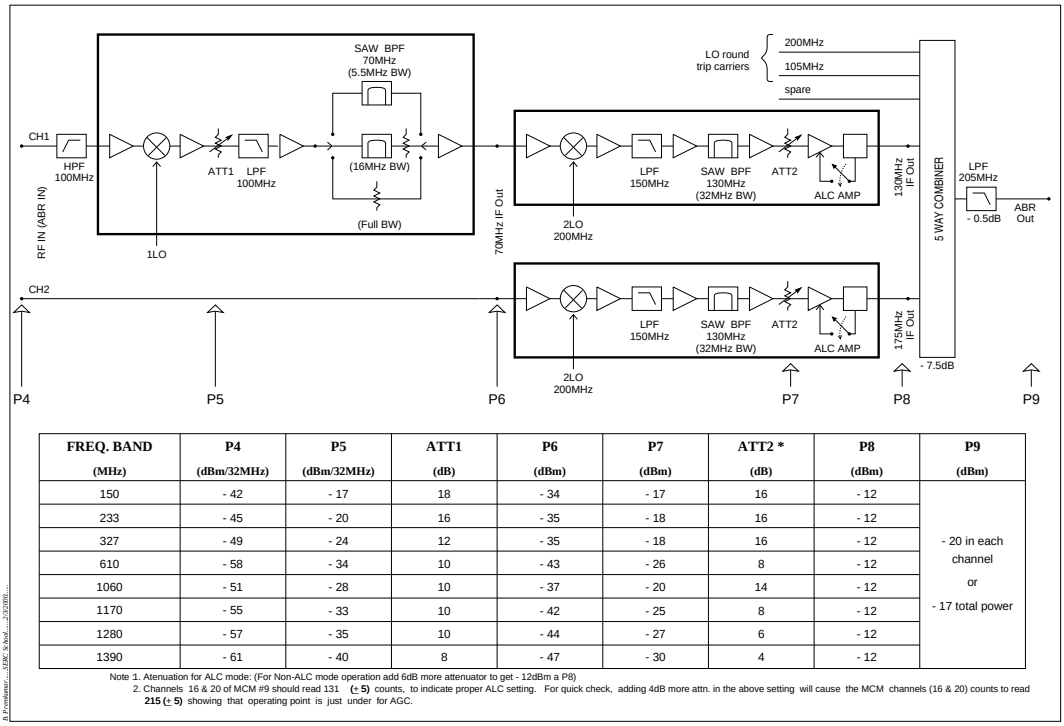
Figure 21.3: The RF power levels for a 32 MHz bandwidth at various locations in the RF receiver chain for the different GMRT frequency bands and sub bands. It is assumed that the solar attenuator has been bypassed and that the noise diode is off.

Table 21.1: Contributions from different sources to the system temperature at the various GMRT bands.

Band (MHz)	BW (MHz)	Cable Loss (dB)	Pol Loss (dB)	LNA (K)	Receiver (K)	Ground (K)	Sky (K)	Sys (K)
50	40	1.33	0.80	895	1651	19	6500	8170
150	40	0.2	0.75	150	260	12	308	580
235	40	0.55	0.25	35	103	32	99	234
327	40	0.13	0.18	30	55	13	40	108
610	40	0.22	0.15	30	59	32	10	101
1060	120	0.22	-	35	53	25	5	83
1170	120	0.22	-	32	49	24	4	77
1280	120	0.22	-	30	47	23	5	74
1390	120	0.22	-	28	45	23	5	72

21.5 The Antenna Base Receiver

From the Common Box, the signal is brought down via a coaxial cable[8] to the Antenna Based Receiver (ABR), which is housed in a shielded room inside the antenna shell[9] Figure 21.4 shows a schematic block diagram (and also gives the expected signal levels for a 32 MHz bandwidth) at different stages of the ABR, and the nominal values to be set for the pre and post attenuators (see below).



FREQ. BAND (MHz)	P4 (dBm/32MHz)	P5 (dBm/32MHz)	ATT1 (dB)	P6 (dBm)	P7 (dBm)	ATT2 * (dB)	P8 (dBm)	P9 (dBm)
150	- 42	- 17	18	- 34	- 17	16	- 12	- 20 in each channel or - 17 total power
233	- 45	- 20	16	- 35	- 18	16	- 12	
327	- 49	- 24	12	- 35	- 18	16	- 12	
610	- 58	- 34	10	- 43	- 26	8	- 12	
1060	- 51	- 28	10	- 37	- 20	14	- 12	
1170	- 55	- 33	10	- 42	- 25	8	- 12	
1280	- 57	- 35	10	- 44	- 27	6	- 12	
1390	- 61	- 40	8	- 47	- 30	4	- 12	

Note 1. Atenuation for ALC mode: (For Non-ALC mode operation add 6dB more attenuator to get - 12dBm a P8)
2. Channels 16 & 20 of MCM #9 should read 131 **(± 5)** counts, to indicate proper ALC setting. For quick check, adding 4dB more attn. in the above setting will cause the MCM channels (16 & 20) counts to read **215 (± 5)** showing that operating point is just under for AGC.

Figure 21.4: Schematic block diagram of the antenna base receiver. The nominal values that the attenuators should be set to, as well as the expected power levels at different stages are also shown. See the text for more information.

The High pass filter (HPF) at the input of the ABR has a rejection of about 40 dB at 70 MHz and provides the IF rejection[10]. After mixing the signal power level can be adjusted using a variable attenuator[11] (which can be set from 0 to 30 dB in steps of 2 dB). After this the the signal passes through a SAW filter where one of three bandwidths (32 MHz, 16 MHz and 5.5 MHz) can be chosen. The net gain of the filter is independent of the chosen bandwidth due to the incorporation of bandwidth compensating gain circuitry. The signal is then up-converted to either 130 or 175 MHz (depending on which polarization it is), passed through further gain and an attenuation[12] and then an Auto-

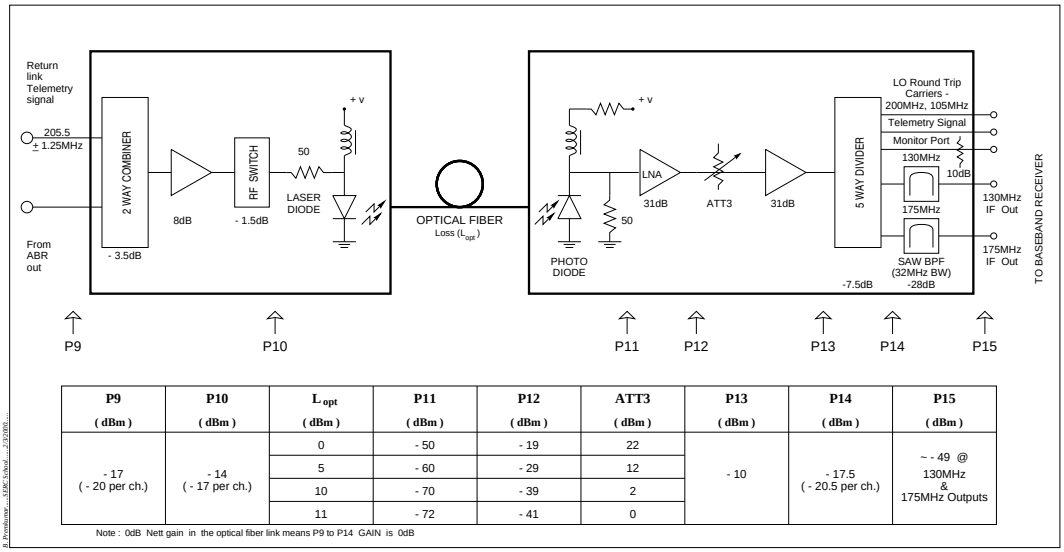
[8]The loss in this cable is a strong function of frequency. This fact can be used to advantage in the bypass mode for image rejection. In the bypass mode if one places the I LO above the RF of interest, the image frequency is suppressed due to the greater attenuation at higher frequencies.

[9]The concrete structure on which the dish rests is called in local parlance the “antenna shell”.

[10]i.e. prevents passage of a 70 MHz signal from the RF directly through to the 70 MHz IF stages.

[11]Which in local parlance is called the “pre-attenuator”.

[12]Also settable from 0 to 30 dB in steps of 2 dB, and called the “post-attenuator”.



P9 (dBm)	P10 (dBm)	L_{opt} (dBm)	P11 (dBm)	P12 (dBm)	ATT3 (dBm)	P13 (dBm)	P14 (dBm)	P15 (dBm)
- 17 (- 20 per ch.)	- 14 (- 17 per ch.)	0	- 50	- 19	22	- 10	- 17.5 (- 20.5 per ch.)	~ - 49 @ 130MHz & 175MHz Outputs
		5	- 60	- 29	12			
		10	- 70	- 39	2			
		11	- 72	- 41	0			

Note : 0dB Nett gain in the optical fiber link means P9 to P14 GAIN is 0dB

Figure 21.5: Schematic block diagram of the Fiber Optic Link. See the text for more information on each block.

matic Level Controller, (ALC). The ALC ensures that the even if the RF signal level varies (for eg. because you point at a bright source) the signal level at the IF remains at the optimum level for transmission through the optical fiber[13] The IF powers are continuously monitored and the monitoring data is sent to the CEB. The recommended IF power level is -20 dBm per polarization.

At the final stage of the ABR, the LO round trip carriers, the monitoring data as well as the astronomical signal are combined in a 5 way combiner and sent to the CEB via the optical fiber link.

21.6 The Fiber-Optic Link

Figure 21.5 shows the schematic block diagram and the nominal powers at different stages of the the fiber-optic return link[14]. The link consists of a laser diode (which converts the input electrical signal into an optical signal), the optical fiber itself, a photo diode (which converts the optical signal back into an electrical signal) followed by an amplifier and a 5 way divider which separates out the monitoring data as well as the two polarizations of the astronomical signal.

The Fiber optic link is designed to provide a net gain of 0 dB from the input (P9 in Figure 21.4) to the output (P14 in Figure 21.5) which is also the input to the baseband system discussed in Chapter 23. The link is meant to have 0 dB gain irrespective of the length of the fiber optic cable linking the antenna to the CEB. The attenuator (ATT3 in the Figure 21.5) can be varied in accordance with link optical loss to provide this no loss/gain configuration. The level diagram shows the attenuator settings for 0, 5, 10 and 11 dB of

[13]The ALC has a time constant of the order of 0.1 seconds. This can produce artifacts in signals (eg. pulsars) whose short timescale structure is of interest. For such observations there is a provision to switch the ALC off.

[14]As discussed in Chapter 22 each antenna has two fibers connecting it to the CEB. One fiber is used to send control signals to the antenna, and is referred to as the forward link, while the other fiber is used to bring back the astronomical signal and monitoring data to the CEB and is called the return link.

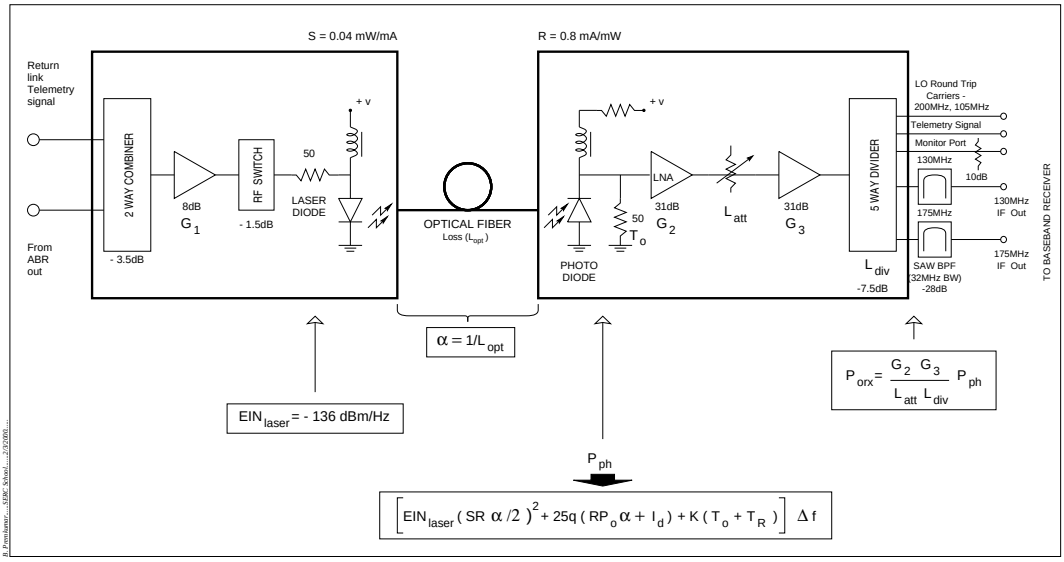


Figure 21.6: Values or expressions for the gain/loss at in the various stages of the fiber optic link. See also Figure 21.7.

optical loss (L_{opt}).

The fiber-optic receiver also contains 32 MHz SAW filters centered at 130 and 175 MHz to separate out the 130 and the 175 MHz IF signals for routing to the base band converter subsystem. The level of the signal at this point (P15) is nominally -49 dBm[15]

An ideal communications link would transfer signals unaltered from the input to the output. Any real link however introduces both additional noise as well as distortions into the signal it transports. In the GMRT fiber-optic link, these non idealities include the laser intensity noise, shot noise and thermal noise of the laser diode, loss and reflections in the optical fiber, as well as shot noise and thermal noise in the photo-diode. Figure 21.6 gives expressions for these various noise terms, and Figure 21.7 and Table 21.7 give the expected values for the various noise terms for the GMRT fiber optical link. The largest loss is for the most distant antennas, and turns out to be ~ 11 dB. From Table 21.2 (or Fig 21.8) the corresponding equivalent input noise (EIN) is ~ -41 dBm. The nominal input power level (P9) of -20 dBm would hence give a signal to noise ratio of ~ 20 dB, i.e. 100. In this case, the system temperature is degraded by 1% due to noise added by the link.

[15]The fiber-optic receiver has a monitor point at the front panel in order to allow measurements of the IF signals and other carriers using a Spectrum Analyzer.

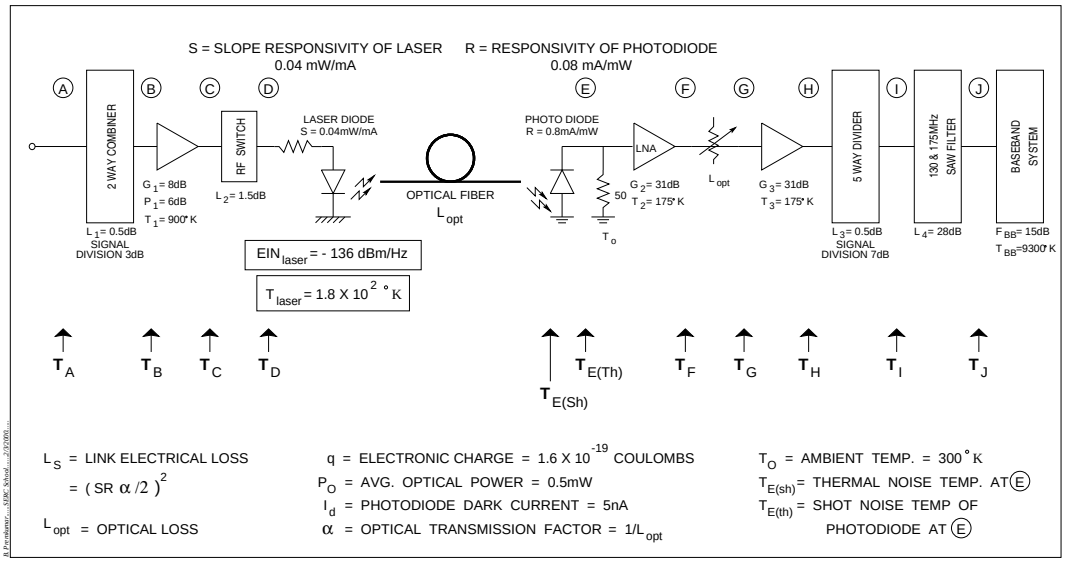


Figure 21.7: Equivalent system noise temperatures at various stages of the GMRT fiber optic link. The link has been designed to have a net gain of 0 dB and to increase the system temperature by less than 1%.

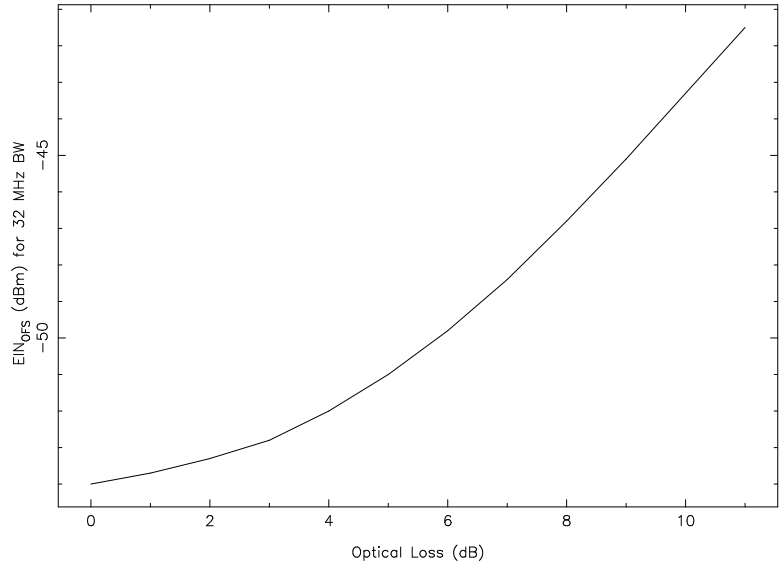


Figure 21.8: Equivalent input noise as a function of optical loss for the GMRT fiber optic link. The maximum optical loss (which occurs for the most distant antennas) is ~ 11 dB.

Table 21.2: Equivalent system noise temperatures as a function of link loss at various stages of the GMRT fiber optic link. See also Figure 21.7.

EIN_{OFS} (dBm)	T_A 10^6K	T_B 10^6K	T_C 10^6K	T_D 10^6K	L_S dB	L_{opt} dB	T_{ORX} K	$T_{E(sh)}$ K	$T_{E(th)}$ K	T_F 10^4 K	L_{att} dB	T_G 10^2 K	T_H 10^6 K	T_I 10^6K	T_J 10^2 K
-54.0	9.00	4.02	25.38	17.97	-36	0	4062	116	3946	437	22	271.82	34	6.06	93
-53.7	9.61	4.29	27.06	19.16	-38	1	2751	92	2659	275	20	271.82	34	6.06	93
-53.3	10.55	4.71	29.69	21.02	-40	2	1922	73	1849	173	18	271.82	34	6.06	93
-52.8	12.07	5.39	34.04	24.10	-42	3	1407	58	1349	110	16	271.82	34	6.06	93
-52.0	14.36	6.41	40.47	28.65	-44	4	1069	46	1023	69	14	271.82	34	6.06	93
-51.0	18.10	8.08	50.96	36.08	-46	5	861	37	824	44	12	271.82	34	6.06	93
-49.8	23.74	10.60	66.90	47.36	-48	6	722	29	693	28	10	271.82	34	6.06	93
-48.4	32.44	14.48	91.39	64.70	-50	7	629	23	606	17	8	271.82	34	6.06	93
-46.8	46.99	20.98	132.38	93.72	-52	8	580	18	562	11	6	271.82	34	6.06	93
-45.1	70.76	31.59	199.32	141.11	-54	9	555	15	540	8	4	271.82	34	6.06	93
-43.3	106.74	47.65	300.62	212.82	-56	10	530	12	518	6	2	271.82	34	6.06	93
-41.5	160.94	71.85	453.34	320.94	-58	11	506	9	497	3	0	271.82	34	6.06	93

$T_A = T_{OFS} = 2.24T_B + (L_1 - 1)T_0$ $T_B = T_C/G_1 + T_1 \simeq T_C/G_1$ $T_C = L_2\, T_D + (L_2 - 1)T_0 \simeq L_2 T_D$

$T_D = L_S T_{ORX} + T_{laser}$ $T_{ORX} = T_{E(th)} + T_{E(sh)}$ $T_{E(th)} = T_0 + T_2 + T_F/G_2$

$T_{E(sh)} = \frac{25a(RP_0\alpha + T_0)}{K}$ $T_F = L_{att} T_G + (L_{att} - 1)T_0$ $T_G = T_H/G_3 + T_3$

$T_H = 5.6T_1 + (L_3 - 1)T_0$ $T_I = L_4 T_{BB} + (L_4 - 1)T_0$ $T_J = T_{BB}$; $L_S = 1/2(SR\alpha)^2$

Chapter 22

The GMRT Optical Fiber System

M. R. Sankararaman

22.1 Introduction

The Giant Metrewave Radio Telescope (GMRT) consists of a distributed array of antennas all connected to a Central Electronics Building (CEB) via optical fiber links. Optical fibers are superior to the more traditional co-axial cables or waveguides in a variety of respects. Optical fibers have lower transmission losses, higher bandwidth and have better isolation against radio frequency interference. More quantitatively, while the loss in co-axial cables are several 10 s of dB/km, the loss in optical fibers is as low as 0.2 dB/km. Further 100 GHz-km bandwidths are routinely achievable in single mode optical fibers, while the achievable bandwidth for co-axial cables is only ~ 20 MHz-km.

The optical fiber link between the CEB and a given antenna has two major functions:

1. Transmission of local oscillator (LO) as well as control signals from the CEB to the antenna, and
2. Transmission of the astronomical signal as well as monitoring data from the antenna to the CEB.

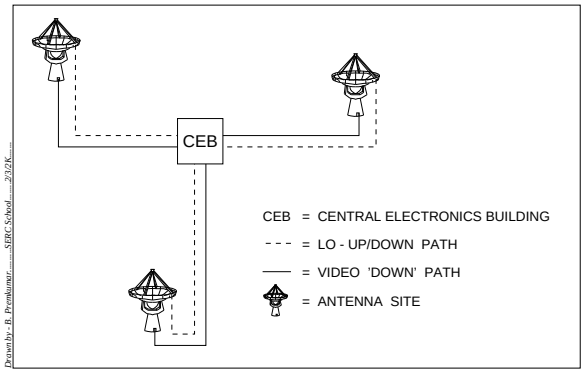


Figure 22.1: Schematic of the optical link at the GMRT. Each antenna is connected to the central electronics building by two fibers, one for the *forward* link, and the other for the *return* link.

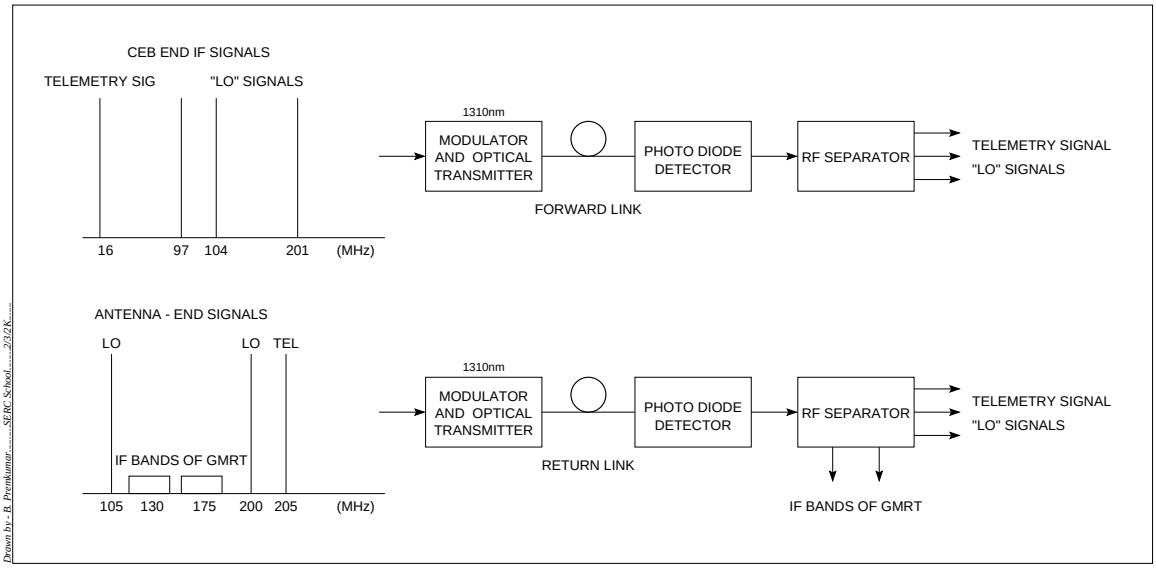


Figure 22.2: Configuration of the GMRT optical communications link. The upper panel shows the *forward* link that takes control signal and LO signals from the CEB to the antenna. The lower panel shows the *return* link that brings the astronomical signal (at the IF frequency) as well as the telemetry and return LO signals from the antenna to the CEB. The frequencies of the various signals transported by the link are also indicated.

As shown in Figure 22.1 there are two fibers between each antenna and the CEB, one of which forms part of the *forward link* and carries the control and LO signals to the antenna, and the other of which forms part of the *return link* and carries the astronomical signal (at the IF frequency, see Chapters 21, 23) and the monitoring data (also referred to as telemetry data) and the return LO[1] back to the CEB. Each link consists an optical transmitter (a laser diode) the fiber itself (which is a single mode glass fiber), and a receiver (a photo diode). A block diagram of the GMRT optical Link is show in Figure 22.2 and the frequencies of the different signals that are transported by the link are also indicated. We now discuss the various elements of the GMRT optical link in some more detail.

22.2 The Laser Transmitter

A block diagram of the GMRT optical transmitter is shown in Figure 22.3. The optical signal that is transmitted down the fiber is generated by appropriately modulating a laser diode, which is essentially a forward biased p-n junction diode (typically InGaAsP). The edges of the p-n diode are cleaved such that they act as mirror resonators. Photons travel between the mirrors and for the wavelengths which bear the following relationship with distance between the mirrors, longitudinal mode oscillations occur:

$$\nu_q = q(c/(2 \times n \times l)) \tag{22.2.1}$$

where q is an integer, l is the length of cavity, n is the refractive index of the medium and ν_q is the longitudinal mode frequency. An active medium within the diode provides positive feedback to these photons thus providing amplification.

[1]The return LO is useful in measuring the phase stability of the system as well as in correction for the phase

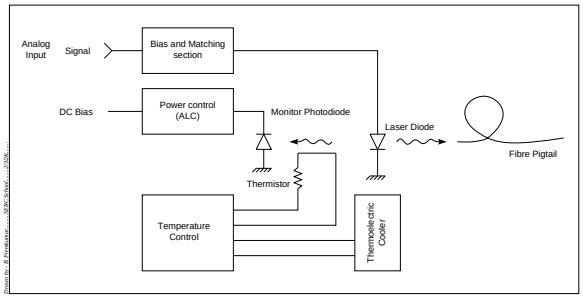


Figure 22.3: Block Diagram of the GMRT optical transmitter.

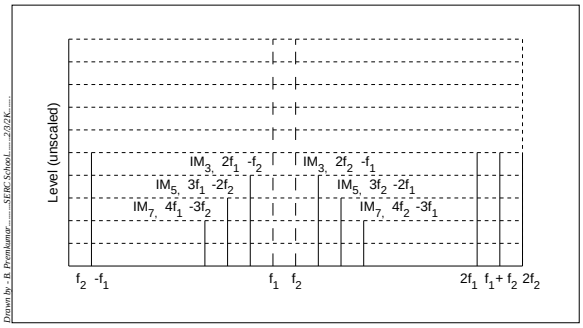


Figure 22.4: The relationship between the frequencies of a few low order inter-modulation products (bold lines) and the the fundamental input frequencies (dashed lines).

The laser used in GMRT is of multi-mode type. The (nominal) peak wavelength is 1300 nm and spectral width is 2 nm (rms). Multi-mode lasers are appropriate for "low" (i.e < 10 GHz) bandwidth applications. At higher bandwidths multi-mode lasers are not acceptable, since they lead to more dispersion and also to *inter-modulation* products. Inter-modulation (IM) products are essentially a particular kind of non linear response. When two pure sine waves are fed to a non ideal device, the output will have additional frequency products that are related to the frequencies of the two input sine waves. These are called IM products of different orders. Figure 22.4 shows a few low order inter-modulation products. The amplitudes for these products is a non linear function of the amplitudes of the input sine waves (see Figure 22.5). The figure also illustrates *gain compression* where beyond a critical input power the output is no longer linearly related to the input, even at the fundamental frequency.

The laser intensity is modulated according to the signal that is to be transmitted, i.e at the GMRT one uses analog modulation. There are two types of analog modulation, *direct* and *external*. In direct modulation the signal is applied directly to an optical carrier generator whose light output varies as per the applied signal. In external modulation the modulating signal is applied outside the device for changing the intensity of the light carrier. In GMRT the simpler direct modulation method is employed.

In the linear regime, the optical power output, P_{opt} by the laser is proportional to the input current i_n, the constant of proportionality is the slope of the characteristic curve and is usually denoted by S.

---

introduced during the LO transmission process.

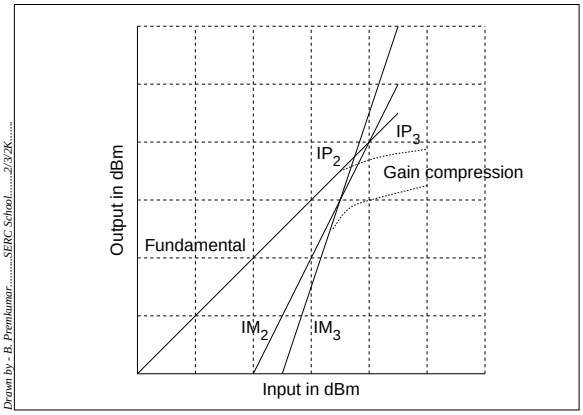


Figure 22.5: Output of a slightly non-ideal optical transmitter showing 2nd and 3rd order inter-modulation products as well as gain compression

22.2.1 Laser Specifications

Rated power o/p	:	1 mW @ 51 mA bias current
Threshold current	:	28 mA
Peak wavelength	:	1306 nm
Slope of the transfer curve	:	0.04 mW/mA (appendix 2).
EIN (Eqv. Input Noise)	:	-137.57 dBm/Hz.

22.3 The Optical Fiber

An optical fiber is essentially a dielectric (silica glass) waveguide consisting of a core and cladding. The core is usually has a circular cross-section (although elliptical or other cross-sections are also used) and is made of doped silica of refractive index slightly higher than that of the cladding (which is made of pure silica). Light waves are guided along the fiber via total internal reflection. If light is launched at an angle greater than the critical angle, the rays are reflected back into the core from the surface separating the core and cladding. The rays travel along the length of the fiber by continuous reflections of this type. Rays launched at different angles travel along different paths (or modes) and arrive at the receiver at different times, leading to *inter-modal dispersion*. Fibers are classified as *single-mode* or *multi-mode* depending on whether they support one or more. Single mode fibers have narrow cores, typically 10νm. Multimode fibers have core dimensions ~ 5 times larger. The number of modes a given fiber can support is characterized by the V number, which depends on the frequency, the core radius and the refractive indices of the core and the cladding.

$$V = \frac{\omega}{c} a \sqrt{{n_1}^2 - {n_2}^2} \tag{22.3.2}$$

where $n_1,\ n_2$ are the refractive indices of the core and cladding, a is the radius of the core and ω is the angular frequency of the light being transmitted through the fiber. The number of modes N is given by $N = V^2/2$. Multimode fibers have bandwidths that are ~ 100 times smaller than single mode fibers and are best suited to short haul applications. In addition to the number of modes supported, the polarization properties of the fiber are also of interest. One can make fibers that maintain the polarization state of the transmitted light by proper choice of core cross-section and refractive index gradient

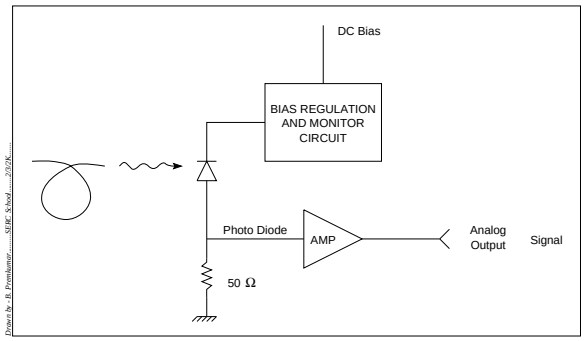


Figure 22.6: Block diagram of the GMRT optical receiver.

across the core and the cladding.

Dispersion is an important characteristic of an optical fiber, it determines the bandwidth and channel carrying capacity of the fiber. here are three kinds of dispersion viz: inter-modal dispersion, material dispersion and waveguide dispersion. Inter-modal dispersion occurs because of the different modes in which the light propagates in the fiber travel different paths. This causes differences in the arrival time of the rays at the receiver and hence a distortion of the signal. Inter-modal dispersion is less in fibers which have a parabolic refractive index parabolic profile in the core region. This change in refractive index causes a change in the light travel time in different parts of the core which partially compensates for the different path lengths. Material and waveguide dispersion are wavelength dependent. Material dispersion arises because of variation in the refractive index of the core material (i.e. silica) across the transmission band. Waveguide dispersion is due to the propagation constant (i.e. the inverse of the group velocity) dependent property of the medium. The derivative of the propagation constant w.r.t frequency is dependent on the frequency itself, even in the absence of material dispersion.

Dispersion affects both the temporal and spectral characteristics of the signals and it is essential to minimize it as far as possible. This can be done by

1. Choosing the 1300 nm window where dispersion is minimum. It may be noted that dispersion for silica fiber is minimum in the 1300 nm band(typically 2 ps/km-nm) compared to that at the 1550 nm band (15 ps/km-nm). However the attenuation is higher in the 1300 nm band (0.31dB/km) than that in the 1550 nm band (0.15 dB/km).

2. Choosing a laser with line-width as small as possible(< 1 nm), like a single longitudinal mode type or DFB laser.

3. Using external modulation. Unlike direct modulation, external modulation does not affect the physical mechanism of the laser and does not introduce spreading of frequency or chirping.

4. Using dispersion compensation. This is essential achieved by proper design of the refractive index gradient across the fiber.

22.4 The Optical Receiver

Photo detection is the process of conversion from optical to electrical domain. A block diagram of the GMRT optical receiver is shown in Figure 22.6. The basic detector is a

reverse biased p-n junction diode. In this bias condition a reverse leakage current (the *dark current*) flows. The threshold for detection is determined by the dark current. The other important characteristic of a photo-detector is its *responsivity* R. The responsivity is a measure of the efficiency with which light is converted to electrical current and it is related to the width of the depletion region of the diode and to the spectral response of the receiver. A larger depletion region leads to a better responsivity. PIN diode detectors made of InGaAsP and grown on InP are popular photo-detectors as they have low dark currents and high responsivity. In the case of the GMRT, the detector used has a dark current of 5 nA and R of 0.8 mA/mW.

In order to match the device to the electrical output device, care has to be taken to maintain a wide frequency response and to keep the thermal noise contribution of the detector low. In the case of the GMRT the laser noise is more than the thermal noise contribution of the photo-detector.

22.5 Link Performance

The relation between the power delivered at the output of the detector to the power input to the laser is:

$$P_o = P_i\big[\frac{SRl}{2}\big]^2 \tag{22.5.3}$$

Where S is the slope of the laser diode characteristic curve, R is the responsivity of the photo-diode and l is the loss in the fiber. The total loss is the combination of losses due to attenuation in the fiber, splices, bending of the fiber and couplers. Measurements show that the optical losses of the links vary between 0.3 to 8.7 dB for the various antenna stations.

In addition to this change in signal power level, the link also introduces noise. Noise is introduced by the laser diode, the photo- diode as well as all resistive elements in the signal path.

The laser diode introduces noise due to quantum fluctuations even under conditions of constant bias current. This is called *Relative Intensity Noise* (RIN) and is define as:

$$RIN = \frac{\big\langle \Delta P^2 \big\rangle}{\big\langle P^2 \big\rangle} \tag{22.5.4}$$

where ΔP are the fluctuations in the laser diode output power, and P is the instantaneous laser diode output power. The laser diode noise is also often characterized by the *Equivalent Input Noise* (EIN) which is defined as $EIN =< \Delta I^2 > R_i$ where ΔI is the input current fluctuation that would correspond to the output power fluctuations ΔP. It can be shown that EIN $= \mathrm{RIN}(I_{\mathrm{bias}} - I_{\mathrm{threshold}})^2 \times R_i$, where R_i is the input resistance.

The noise generated within the the photo detector is called shot noise. As the name suggests, it is due to the discrete nature of light and its interaction of photons with materials. Shot noise is present in the detector even in the absence of illumination and increases with illumination of the detector with light. All resistive elements contribute to thermal noise. The total noise power(N) is the sum of the laser, shot and thermal noise components. The Signal to Noise Ratio (SNR) of the link can be shown to be

$$SNR = \frac{P_i\big[\frac{srl}{2}\big]^2}{\big[EIN(\frac{srl}{2})^2 + 25e(RP_0l + I_d) + FkT\big]B} \tag{22.5.5}$$

where F is the noise figure of the detector amplifier, T is the temperature of the resistive elements, B is the bandwidth of the link, I_d is the dark current, P_0 is the average output power of the laser, and e is the electron charge. The analog optical fiber communication system of the GMRT has been designed to ensure a minimum SNR of 20 dB .

In addition to this intrinsic additive noise, there are various other imperfections in the fiber optic link. Discontinuities in the refractive index near the connectors, couplers, bends in the fiber and impurities along the length of the fiber could cause part of the light to get reflected back into the laser. This leads to the formation of a resonant cavity between the discontinuity and the laser hence to ghosts. To overcome this problem, optical isolators and low reflection connectors are used. An optical isolator is a unidirectional device with highly reduced signal transmission in the reverse direction. Low reflection connectors are special devices with refractive index matching and focusing arrangements.

The other important characteristic of the optical link, apart from the SNR is the dynamic range, i.e. the range over which its response is linear. The dynamic range of the GMRT optical fiber link is ~ 14 dB were the input to be purely Gaussian random noise, and ~ 19 dB for quasi-sinosidal input.

Chapter 23

Local Oscillator and Base-band Systems

T. L. Venkatasubramani

23.1 Requirement for a Local Oscillator System at the GMRT

The GMRT Analog Receiver, in its simplest form, can be considered as a 2-terminal black box, as given in Figure 23.1.

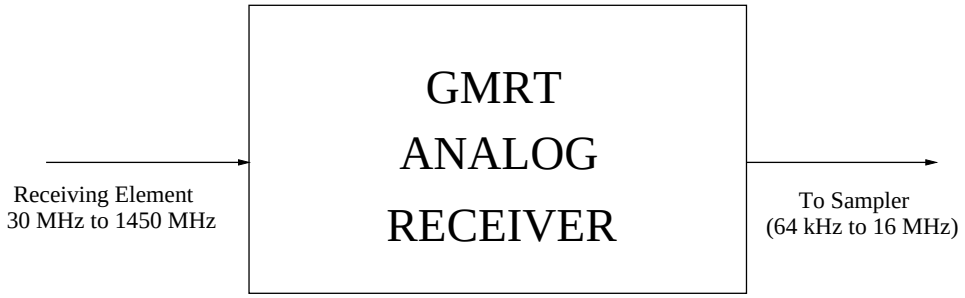


Figure 23.1: A Two-terminal representation of the GMRT analog receiver system.

The receiving element (i.e. any of the dual linearly polarised feed systems of the GMRT) is connected at the input of the black box and provides a signal consisting of:

1. Thermal noise power kTB.[1]
2. The astronomical signal, which is usually much weaker than the thermal noise power.
3. Unwanted Radio Frequency Interference (RFI), which could occur anywhere in the frequency spectrum and is often much stronger than the thermal noise power.

[1]where k is the Boltzman constant, T is the system temperature and B is the bandwidth of the signal. At the GMRT for a 32 MHz bandwidth this power is typically of the order of -100 dBm.

The output of the black box is in base-band frequency range, which is typically from DC to a maximum of 16 MHz. The upper value determines the maximum instantaneous bandwidth of GMRT. The base-band signals are then digitized by a sampler. The nominal power level needed for the sampler is 0 dBm. Since the typical input power level is -100 dBm, the gain within the black box is about 100 dB.

This large amplification has to be achieved while simultaneously providing the desired band-limiting[2] and spurious free dynamic range[3] in the presence of strong RFI. For this, the electronics system within the black box has been implemented as a heterodyne receiver, where the RF signal from the receiving element is converted to the base-band signal via different stages of frequency translation (see also Chapter 3). This frequency translation requires multiple Local Oscillator signals.

23.2 The Frequency Translation Scheme used at the GMRT

The simplified block diagram of frequency translation scheme used to convert the RF signals from each antenna of the GMRT to the base-band signals required by the sampler is given in Figure 23.2. A schematic of the typicaly used mixing scheme at the GMRT is shown in Figure 23.3.

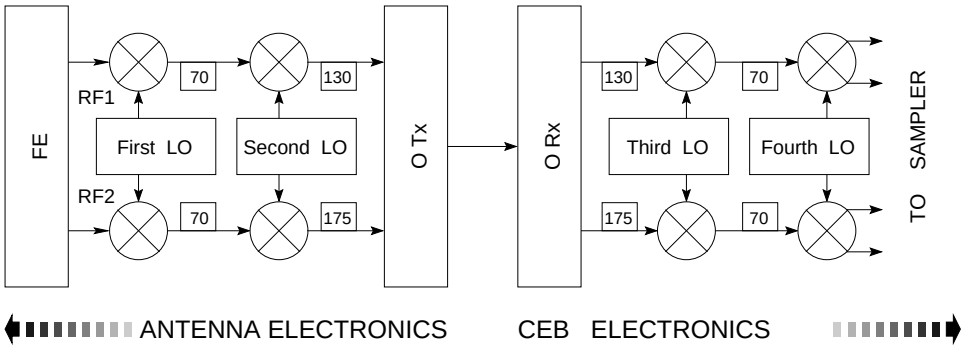


Figure 23.2: Block diagram of frequency translation Scheme at the GMRT. The numbers in the boxes are the Centre Frequency of the signal at that point.

The pair of RF signals from the Front End (FE), which are typically in the range of 30 to 1660 MHz are initially converted to a first IF signal centered at 70 MHz. The choice of the first IF frequency has been decided by the availability of commercial sharp cut-off band-pass filters with a wide choice of bandwidth at this frequency[4]. This translation from RF to the first IF needs a first LO signal, which should be tunable at least over the range of 100 to 1590 MHz. The GMRT is designed to simultaneously process two RF signals (RF1 and RF2, also called the 130 signal and the 175 signal respectively). These signals could be either (a) two polarisation signals at the same RF band, or (b) one polarisation signal in each of two different RF bands. To cater to case (b), we need two independent first LO sources.

The pair of first IF signals are brought from each antenna to the central electronics building (CEB) by a single optical fibre for further processing. Hence, we need to separate

[2] i.e. one needs to filter the signal so that only frequencies within the band of astronomical interest are accepted.

[3] i.e. one needs to ensure that the entire system is sufficiently linear so that RFI at one frequency does not produce spurious spikes at other frequencies.

[4] At the GMRT, IF bandwidths of 32, 16 and 5.5 MHz are available. See Chapter 21 for more details

the two first IF signals (centered at 70 MHz) in the frequency domain before they can be combined and fed to the optical transmitter (OTx) unit. For this, one of the first IF signals is translated to a second IF signal centered at 130 MHz and the other, to another second IF centered at 175 MHz. The choice of these centre frequencies has been decided by (a) The maximum bandwidth of the first IF signal and (b) the need to keep the overall band occupancy on the fibre to within an octave. The value for the two second LO signals chosen for the GMRT are hence 200 and 105 MHz. At the CEB the *reverse* translation is done, after the optical receiver (ORx) unit, to produce a pair of third IF signals, centered at 70 MHz. This requires two third LO signals, at 200 and 105 MHz respectively.

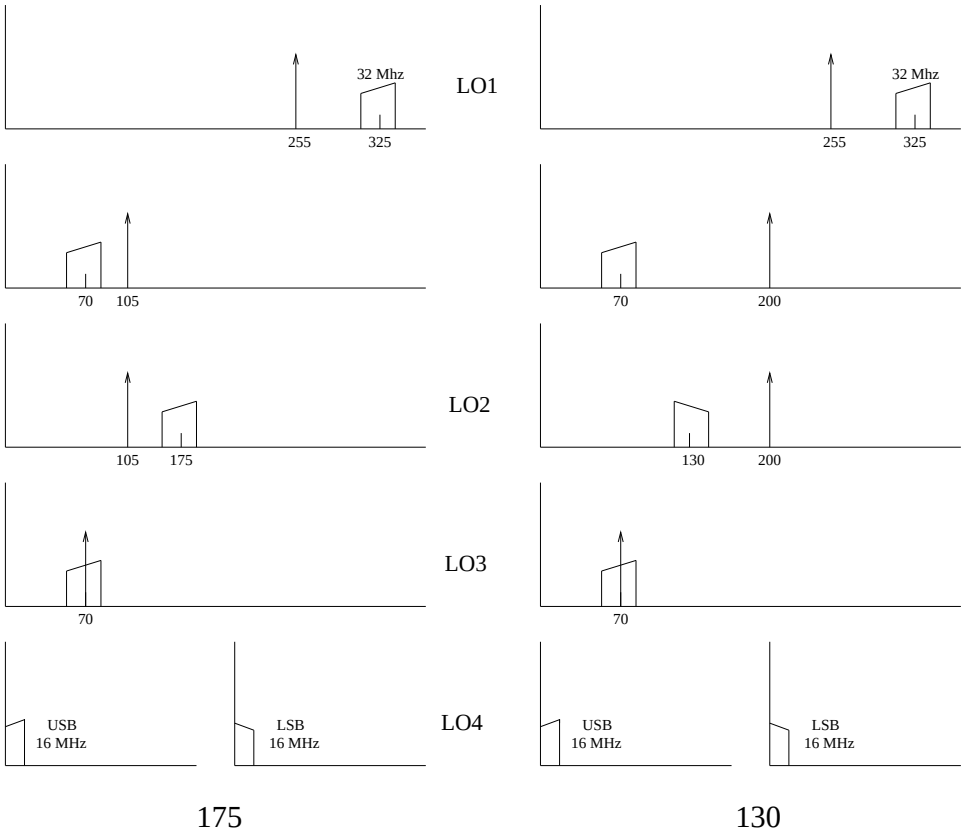


Figure 23.3: A typical mixing scheme at the GMRT. LO1 is tunable in the range 30 - 1590 MHz, in steps of 1 MHz below 350 MHz and 5 MHz above that. LO2 and LO3 are not tunable. LO4 is tunable in the range from 50 - 90 MHz, in steps of 100 Hz. As can be seen from the figure, if $\nu_{LO1} > \nu_{RF}$ then the sky frequency increases with correlator channel number for the USB and decreases with increasing correlator channel number for LSB. This is true for both the 175 and the 130 signals. If $\nu_{LO1} < \nu_{RF}$, then for both 175 and 130 signals, the sky frequency decreases with increasing correlator channel number for USB and increases with increasing channel number of LSB.

The last stage of frequency translation is to base-band, using a fourth LO signal, which can be set to any frequency from 50 to 90 MHz in 100 Hz steps. The step size is determined by the need to incorporate online doppler tracking, so that a spectral line under observation can be confined to a specified channel in the correlator throughout the entire observation.

23.3 Generation of Phase-Coherent Local Oscillator Signals

The GMRT is generally used in the interferometric mode where the correlation between the electric field vectors received by each feed element is measured. To maintain the relative phase between the electric field vectors incident on different antennas, it is essential that the local oscillators used must be phase-coherent[5]. This implies that the frequency of the LOs at the various antennas must be identical and the variation of phase of the LO at a given antenna (with respect to the phase of some reference antenna) must be precisely known during an observation so that necessary correction can be made in real time. This is acheived by having all the local oscillator signals be generated from a single reference frequency source using phase-locked loop (PLL) techniques.

In detail, the third and fourth LO signals are generated in the CEB from an ultra-high stable and pure quartz oscillator at 5 MHz. The second LO generation uses a voltage controlled crystal oscillator (VCXO) in a PLL, while the first LO needs a phase-coherent frequency synthesiser. Signals at 106 MHz and 201 MHz are broadcast from the CEB to each antenna, and these are used at the antennas to generate the 105 MHz and 200 MHz second LOs. The 105 MHz second LO signal is in turn used to generate the first LO. In this way the phase coherence of all the LOs at all the antennas is maintained.

Despite being derived from a common signal, there are still phase variations of the LOs at the different antennas for a variety of reasons. The physical length of the optical fibre link to various antennas varies from a few hundred meters to about 20 kms. As the temperature coefficient for expansion of the fibre is not zero, there will be a variation of the phase of LO signal broadcast from CEB and received at an antenna. The receiver system in each antenna is housed in an air-conditioned environment and undergoes independent cyclic variation in temperature. This also causes the LO phase between antennas to vary in a random manner. All of this would make it desirable to have a system for estimation of phase of the LO signals at all antenna locations. This could be achieved by bringing back the second LO signals from each antenna to the CEB and comparing its phase with that of the signal originally generated at CEB. From this information, the phase variation introduced by the transmission process could be estimated. Of course, this needs the pair of optical fibre to an antenna to be reciprocal and non-dispersive, which has been independently confirmed. However, this scheme is not yet fully implemented.

23.4 Noise Calibration and Walsh Switching

As discussed in Chapter 21, all the GMRT receivers have the facility for noise injection. By injecting noise of known power the system temperature can be measured. The noise at any antenna can be switched on and off (on sub second time scales) according to a pre-determined pattern, which is encoded in PROMs in the Antenna Based Receiver (ABR). By synchronously measuring the total power, it is possible to calibrate the system temperature. The synchronous total power measurment however has not yet been implemented.

Signals from one antenna could leak into another antenna at various points along the signal flow chain. This is normally referred to as *cross-talk*. This would cause a spurious corrrelation between the base-band signals from these two antennas. This leakage can be minimized by switching the phase of the RF signal of each antenna by a pattern that

[5]i.e.the phase difference between the LO signals at antenna i and antenna j should be constant with time for all antennas i, j. If this is not acheivable, then at least the phase difference between the LOs at different antennas should be calibratable in real time.

is ortho-normal to the pattern used for all other antennas. At the correlator the exact reverse phase switching is done for each antenna so that the original phase is recovered just before the cross-correlation is done. Such a scheme would greatly reduce the cross-talk at all points between the RF amplifier and the base-band. Typically the ortho-normal functions used are Walsh functions, and this scheme is called Walsh Switching. The required Walsh patterns for each antenna are also encoded in PROMs situated at the ABR. However the Walsh demodulation at the correlator is yet to be implemented.

23.5 The Base-band System

The base-band system of GMRT processes the IF signals received from the antennas and makes them compatible for the correlator.

The maximum bandwidth of the IF signal is 32 MHz. Considering the fact that the correlator system can run at a maximum clock speed of around 40 MHz, a Single Side-Band (SSB) conversion with image rejection approach is used in the base-band. This results in two base-band signals, (the Upper Side Band, (USB) and the Lower Side Band (LSB)) each with a maximum bandwidth of 16 MHz, for each of the third IF signal. There are hence 120 base-band outputs resulting from 60 third IF signals (typically one from each of two polarizations) from 30 antennas of the GMRT. A simplified block diagram for the system to handle one of the third IF signals is given in Figure 23.4

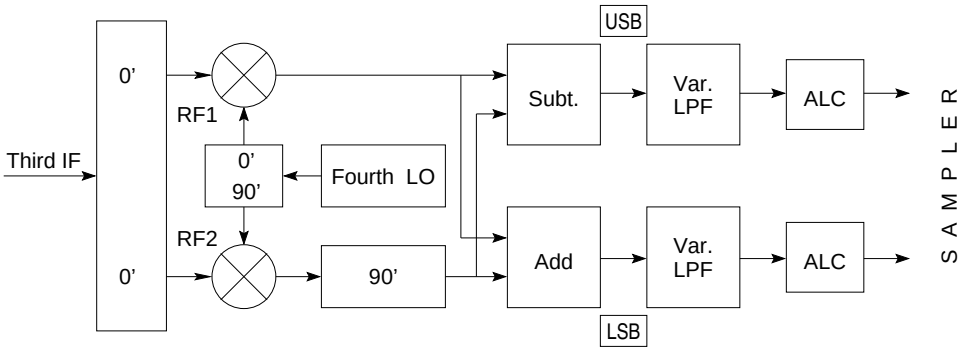


Figure 23.4: A simplified block diagram of the base-band processing at the GMRT.

The input third IF signal at 70 MHz is converted to two base-band signals, the Upper Side Band (USB) corresponding to $70+16$ MHz and Lower Side Band (LSB), to $70-16$ MHz. This single side-band mixers are based on in-phase and quadrature-phase power dividers for the third IF and fourth LO as well as a broadband quadrature network in the base-band. The typical image rejection which has been achieved is 25 dB.

The base-band system has facility for a wide choice for bandwidth, from 62.5 kHz to 16 MHz, in octave steps. This is achieved in the variable low pass filter block. The power output from the system is kept constant by automatically increasing, the gain as the bandwidth is decreased, to keep the product constant. In addition the ALC stage ensures that the sampler is supplied with a constant power. For some applications (eg. Pulsar observations) however, the finite time constant of the ALC produces undersireable distortions of the astronomical signals. For these observations, the ALC can be switched off.

23.6 A Summary of Important Specifications

23.6.1 Array Frequency Reference

Model used	Frequency and Time Standard (FTS) make 1000B Quartz TCXO
Output frequency	5 MHz, 2 Hz, (adjustable externally)
Aging per day	1×10^{-10}
Short-term Stability over 1 to 100 sec	1×10^{-12}
SSB Phase noise	-116 dBc/Hz at 1 Hz offset -140 dBc/Hz at 10 Hz offset -150 dBc/Hz at 100 Hz offset
Harmonics	-40 dBc
Spurious	-100 dBc

23.6.2 First LO Synthesiser

Frequency Coverage	100 MHz to 1795 MHz
Step Size	1 MHz from 100 MHz to 354 MHz 5 MHz from 350 to 1795 MHz
Power output	$+11$ dBm -3 dB
Harmonics	Better than -20 dBc
Spurious	Better than -60 dBc
Phase Noise	Better than -60 dBc/ Hz at 10 kHz offset, (corresponding to a peak-to-peak phase jitter of better than 0.1 deg in time scales of 0.1 msec)
Line Frequency modulation	Better than -20 dBc

23.6.3 Second and Third LO Sources and Offset Frequency Sources

Oscillator circuit	Transistorised VCXO with 5th overtone crystals for sources around 105 MHz and 7th overtone crystal for 200 MHz., operating in a PLL with loop bandwidth of 70 Hz.
Maximum frequency deviation	2 kHz for 10 V range.
Spurious	Better than -50 dBc
Harmonic	Better than -70 dBc
Phase jitter	Less than 1 nsec
SSB Phase noise	Better than -90 dBc/Hz at 100 Hz offset.

23.6.4 Fourth LO Synthesiser

Frequency Coverage	50 to 90 MHz
Step Size	100 Hz
Inband spurious	Better than -60 dBc
Phase noise	-80 dBc/ Hz at 1 kHz offset

23.6.5 Base-band System

Lower cut-off frequency	~ 10 kHz
Choice for Upper cut-off frequency	62.5 kHz, 125 kHz, 250 kHz, 500 kHz, 1 MHz, 2 MHz, 4 MHz, 8 MHz and 16 MHz
Pass-band ripple	< 0.5 dB
Stop band rejection	48 dB/octave
Image Rejection	minimum 20 dB
Fourth LO leakage to output	Better than 60 dB
Third IF input level	-65 dBm to -55 dBm
Base band output power level	0 dBm (ALC ON mode)
Typical Third IF level in ALC OFF mode:	-50 dBm, to give 0 dBm output.

Chapter 24

A Control and Monitor System for the GMRT

R. Balasubramanian

24.1 Introduction

Modern radio telescopes are complex assemblies of electronic and electro-mechanical subunits. To allow a successful observation, all of these sub-units have to be "set" as per the users requirements. For example, the antennas have to track the selected source, the front-ends have to be tuned to the chosen frequency band, all the amplifiers along the signal path have to be set at the value which would give the optimum signal to noise ratio, the local oscillators have to be tuned to select the desired frequency, the correlator has to be set up to do the appropriate fringe and delay tracking etc. In an interferometer like the GMRT this means that one has to, in a co-ordinated manner, control sub-systems which are several tens of kilometers separated from one another. In addition, it would be highly desirable for the health of the critical sub-systems to be able to be periodically monitored, so that should any subsystem fail, the affected data can be flagged, and also of course remedial action could be taken to fix the faulty unit. Further since it is not humanly possible to remember all the various safety limits of each of the sub-systems, one requires the telescope control system to not permit operations which could endanger the safety of the telescope or the human operators. The GMRT Monitor and control system was designed with all these different requirements in mind.

The GMRT control and monitor system (also referred to as the "telemetry system") allows one to

1. Rotate of all the thirty antennas in azimuth and elevation, and/or to track a celestial source.
2. Bringing the required feed in the feed turret to the focus via the Feed Position System(FPS).
3. Select front end system parameters like the observing frequency band, desired noise calibration, etc.
4. Sets the IF sub-systems including the LO frequencies, the IF bandwidths and attenuation, the ALC[1] operation etc.

[1]Automatic Level Controller

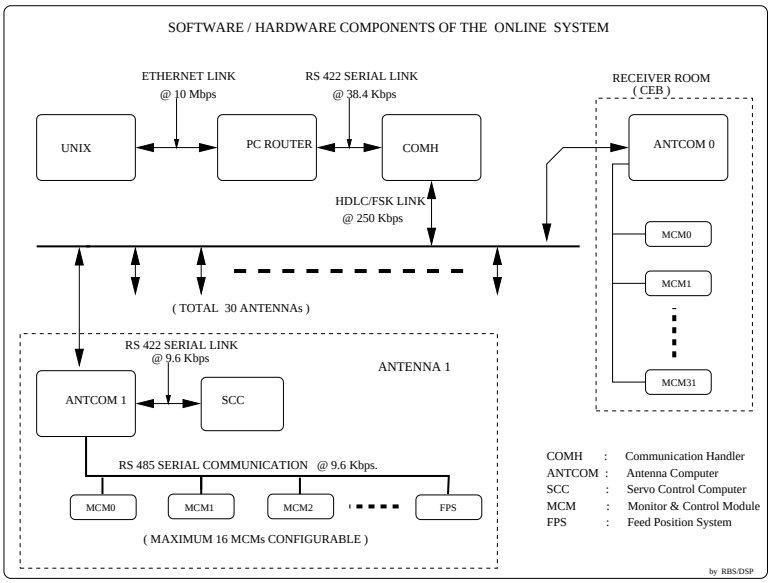


Figure 24.1: Block diagram of the GMRT monitor and control System. See the text for more information.

5. Sets the baseband bandwidth and attenuation.

6. Monitor, literally, hundreds of system parameters at all points along the signal flow path.

7. Have a voice link between each antenna shell and the control room in the CEB (Central Electronics Building).

24.2 Overview

The major components of the Monitor and Control system (see Figure 24.1) are ONLINE (a unix level program that provides the user interface), PCROUTER, a PC based router, COMH, a communications handler that deals with the packet based communication between the Unix workstation on which ONLINE is running and the Antenna Based Computer (ABC, also called ANTCOM) located in each antenna shell, and finally several Monitor and Control Modules, (MCM) which provide the monitoring and control interfaces to the various sub-units (i.e. the LOs, amplifiers, etc.).

We now look at each of these sub systems in slightly more detail.

24.2.1 ONLINE

The ONLINE software running on a UNIX workstation provides the user interface for the control and monitor system. The commands typed by the user are sent to the relevant antenna(s) by the telemetry system. The monitoring data from all the various GMRT subsystems are also logged by ONLINE. Should some critical subsystem fail, ONLINE will raise an appropriate alarm so that remedial action can be taken.

24.2.2 PCROUTER

This converts the data from the format used for TCP/IP (ethernet) communication to one suitable for the serial communication links used by the GMRT telemetry system. As the name suggests this is PC based.

24.2.3 COMH

COMH is the communication handler and it handles all the communication between the UNIX workstation on which ONLINE is running and the various Antenna Base Computers (ABCs). COMH operates in a time division multiplexing (TDM) mode i.e. it sends the formatted user commands to the first antenna and then waits for an acknowledgment. If it receives an error free reply before the timeout period it selects the next antenna and the operation continues. In case COMH doesn't get a reply before the timeout period or if the reception is erroneous then it tries the same antenna again. After a total of three failures COMH passes on a Timeout or Checksum error (as appropriate) to ONLINE and then moves on to the next antenna.

24.2.4 ANTCOM or ABC

There is an ANTCOM (also called an ABC) located in each antenna shell. All communication between the antenna and ONLINE is routed through the ANTCOM in that antenna. The ANTCOM receives various parameters sent by COMH, performs some computations if necessary, and passes on the commands to the appropriate sub system of the antenna. In detail, the ANTCOM has three communication links, viz. (a) the main link between COMH & ANTCOM which operates at 250 kbps, (b) an asynchronous 9.6 kbps RS 422 communication link between ANTCOM & the Servo Control Computer (SCC) and (c) an asynchronous 9.6 kbps RS 485 communication link between ANTCOM & upto 16 Monitor and Control Modules (MCMs).

In addition to the ANTCOMs in the various antennas, there is also an ANTCOM (called ABC0) in the receiver room of the CEB. ABC0 handles the configuring of the baseband system in the receiver room.

24.2.5 Servo Control Computer

In addition to an ANTCOM, each antenna has a Servo Control Computer (SCC) which is responsible for controlling the motion of the antenna. The SCC accepts movement commands, position information etc. from the ANTCOM, checks that the command is sensible, and if so obeys it. It also returns the antenna status information periodically through the same link. This information is passed on by the ABC to ONLINE and is displayed on a monitor in the CEB.

24.2.6 Monitor and Control Modules

MCMs are a general purpose Micro-controller based card which provides 16 TTL Control O/Ps and can monitor upto 64 analog signals. These MCMs are the interface to all the settable GMRT subsystems, like the front ends, the LOs the attenuators etc. In detail, at each antenna, MCM 5 is the interface to the front end system, while MCMs 2,3, and 9 are the interface to the LO and IF systems.

The Feed Positioning System (FPS) which is used to position the feed turret so that the desired feed is at focus is also controlled by the ANTCOM.

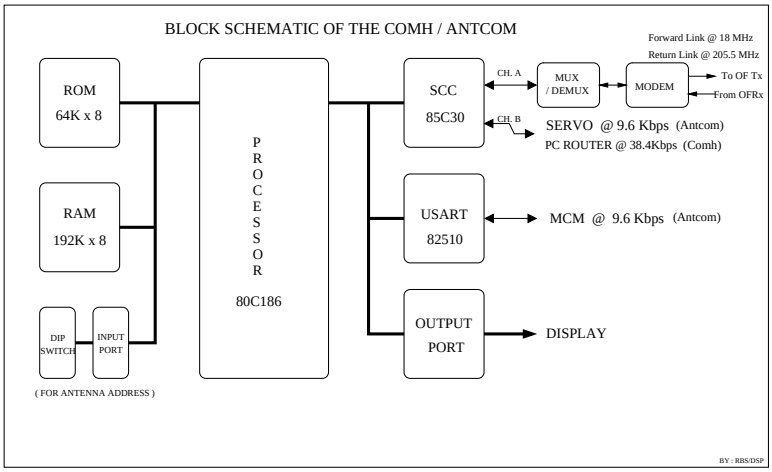


Figure 24.2: Block diagram of the GMRT communication handler (COMH). The antenna computer (ANTCOM) has essentially the configuration.

24.3 Signal Flow in the GMRT Control & Monitor System

User commands for various antennas are processed by ONLINE running on a UNIX workstation and are sent to the COMH via the PCROUTER. The PCROUTER acts as a buffer and accepts the TCP/IP data on a 10/100 Mbps (i.e. a standard ethernet) link, strips the TCP/IP header and sends the data to COMH on a 38.4 kbps link. This uses a standard RS232-C link on the PC side and a conversion to RS 422 signals (differential TTL signals) on the COMH side.

COMH (see Figure 24.2) is basically an 80C186, 16 bit micro-controller based card, which works at a clock speed of 6 MHz. This card also contains a Zilog 85C30 dual channel communication controller. The two channels are respectively for SDLC/HDLC communication at 125 kbps (for communication with the ANTCOMs at the different antennas) and a asynchronous communication at 38.4 kbps (for communicating with the PCROUTER). COMH also has an Intel 29C17 CODEC (voice coder-decoder) to handle voice communication at 62.5 kbps, circuitry for digital Phase Lock Loop and other combinational logic to handle clock recovery and bit interleaving functions, as well as FSK modem chips NE 5080 and NE 5081 to handle FSK modulation and demodulation. COMH multiplexes command data, digitized voice, synchronization pulses, dial pulses and two aux channels into a single bit stream. This bit stream is then converted (via FSK, see section 24.4.1 for details) to an analog signal at 18 MHz.

The block diagram for this multiplexing of voice and data is shown in Figure 24.3. The structure of the multiplexed bit interleaved data frame is shown in Figure 24.4. At the bottom of this figure is shown the flow diagram for the synchronous detector state machine.

The FSK analog signal is sent via the fiber optic link to the ANTCOM at the antenna base. The ANTCOM has the same circuitry as COMH but unlike COMH it handles two serial communication links (using an INTEL 82510 Communication Controller) i.e. the ANTCOM-MCM communication link and a serial link to the Servo Control Computer (SCC). ANTCOM demodulates the FSK signal into 250 kbps data, regenerates the 250 kHz clock using a digital Phase Lock Loop, looks for sync bits and if it finds a match with no error or one bit error then it demultiplexes the data into command, voice, dial pulse and aux data and passes each to the appropriate circuit for further processing (see Fig-

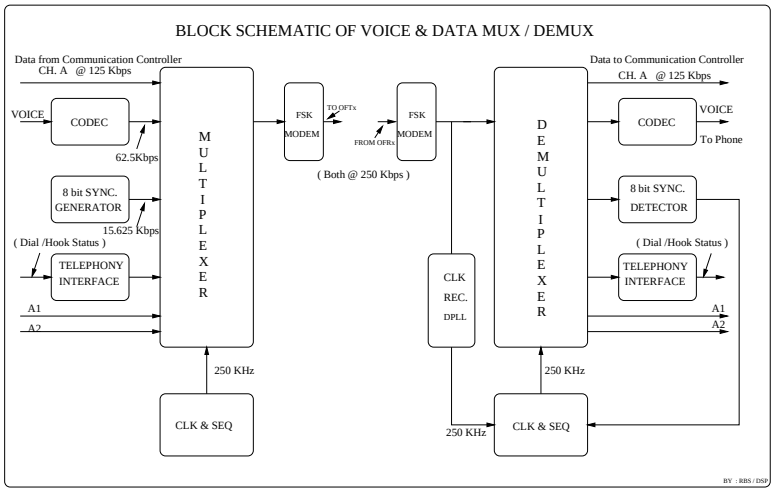


Figure 24.3: Block diagram of the voice and data multiplexer.

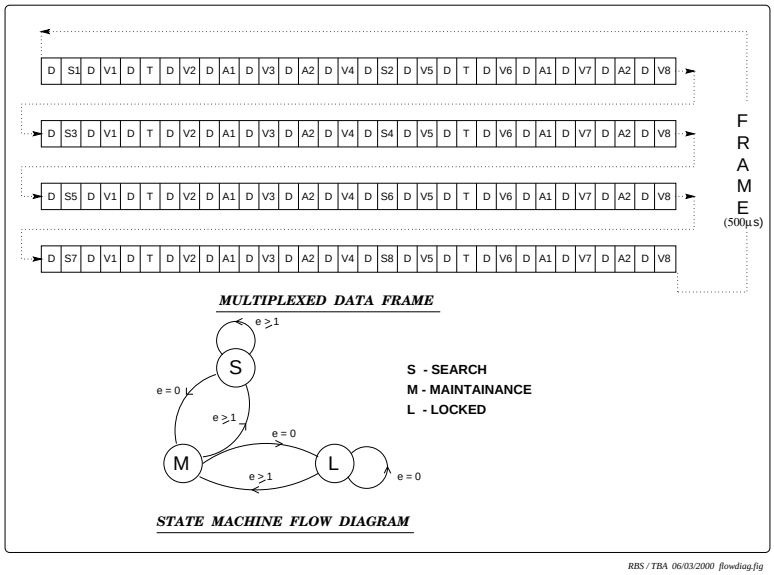


Figure 24.4: Structure of the multiplexed voice and data frame. Shown at the bottom of the figure is the flow diagram for the synchronous detector state machine.

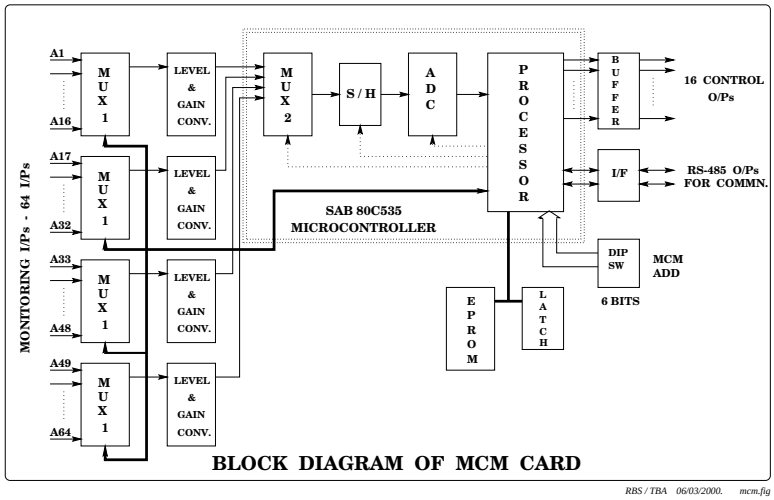


Figure 24.5: Block diagram of the monitor and control (MCM) cards.

ure 24.3). The ANTCOM communicates with and controls the various subsystems in the antenna (other than the servo subsystem for which there is the dedicated SCC) via the MCM cards. A block diagram of the MCM card is shown in Figure 24.5. The MCM card is a general purpose 80C535 Micro-controller based card which provides 16 TTL Control O/Ps and monitors upto 64 analog signals. It also has an RS485 communication link for communicating with the ANTCOM.

For the return link, the ANTCOM takes the monitoring information from SCC, MCMs and FPS forms a packet of SDLC/HDLC data and multiplexes with voice, hook status and aux channels into a single bit stream. This bit stream is converted into an FSK analog signal at 4.5 MHz, which is then up converted to 205.5 MHz using the regenerated 201 MHz as the LO (see Figure 24.8). This analog signal is sent along with the astronomical signals to the CEB. At the CEB thirty CEBCOMs (one for each antenna) demodulate the FSK signal to convert it back into a digital 250 kbps data stream which is passed on to COMH via a 32 way multiplexer (MUX 32) card. The voice signals from the antennas are routed to the EPABX (telephone exchange) system. The block diagram for this telephonic communication is shown in Figure 24.6. The voice signals are digitized using an INTEL 29C17 CODEC IC using a 7.8 kHz clock to produce a 62.5 kbps data stream. The CODEC uses “A law” for data companding/expanding.

24.3.1 Error Detection

The error detection uses both Cyclic Redundancy Check (CRC) and checksum methods. SDLC/HDLC supports 16 bit CRC error detection. CRC can detect all the single errors, double errors and burst errors up to 16 bits in length and can also detect 99% of burst errors of lengths greater than 16 bits.

The way this works is as follows. A cyclic code message consists of a specific number of data bits $G(X)$ and a Block Check Character (BCC). Let n equal the total number of bits in the message, k equal the number of data bits, i.e. $n-k$ is the number of bits in the BCC. The code message is derived from two polynomials which are algebraic representations of two binary words, the generator polynomial $P(X)$ and the message polynomial $G(X)$. The generator polynomial $P(X)$ is a type of code used in CRC-12, CRC-16 and CRC-CCITT.

For example, n bits of binary data can be represented as a message polynomial of

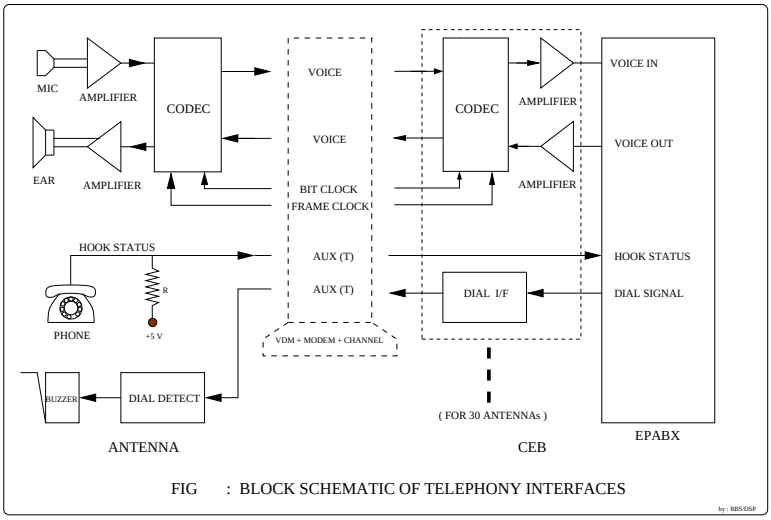


Figure 24.6: Block diagram of the telephony interfaces.

degree $n-1$. Thus, an eight-bit long message 10101010 is represented as

$$G(X) = X^7 + X^5 + X^3 + X^1. \tag{24.3.1}$$

The code message can be constructed as follows:

1. Multiply the message $G(X)$ by X^{n-k} where $n-k$ is the number of bits in the BCC.
2. Divide the resulting product $X^{n-k}[G(X)]$ by the generator polynomial $P(X)$.
3. Disregard the quotient and add the remainder $C(X)$ to the product to get the code message polynomial $F(X)$, which is represented as $X^{n-k}[G(X)] + C(X)$.

The division is performed in binary without carries or borrows. The code message $F(X)$ is transmitted as binary data and the receiver at the other end retrieves the message using the same generator polynomial and accepts the data if the remainder is zero.

24.4 Signal Modulation

As described above, the Control and Monitor system hardware essentially consists of a digital part, an analog part and the Optical Fiber system (see Figure 24.7).

The optical fiber is a single mode analog link operating at 1310 nm, and can carry signals from a few MHz to about 1 GHz. There are two fibers (an 'forward link' and a 'return link') between the Central Electronics Building (CEB) and each antenna. In the forward link the telemetry signals use an 18 MHz carrier, and the return link has a 205.5 MHz carrier. See Figure 24.8 for a schematic of the different signals carried by the forward and return links.

24.4.1 Frequency Shift Keying

As mentioned above, the digital data that the telemetry system generates is converted to an analog signal using Frequency Shift Keying (FSK). FSK is a special type of modulation

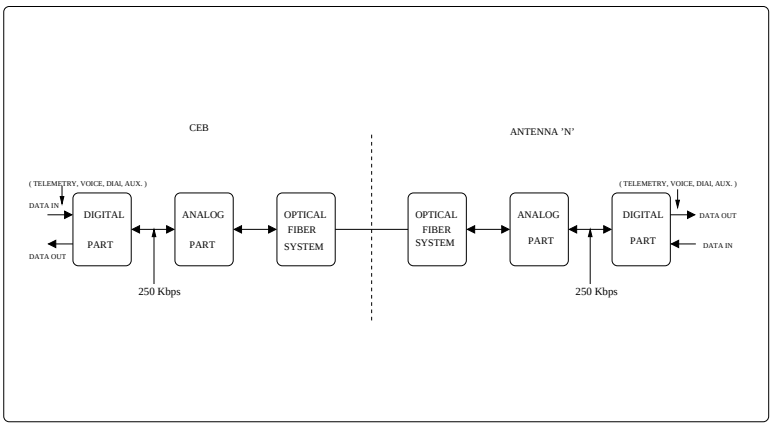


Figure 24.7: Schematic of the GMRT telemetry system. See the text for more information.

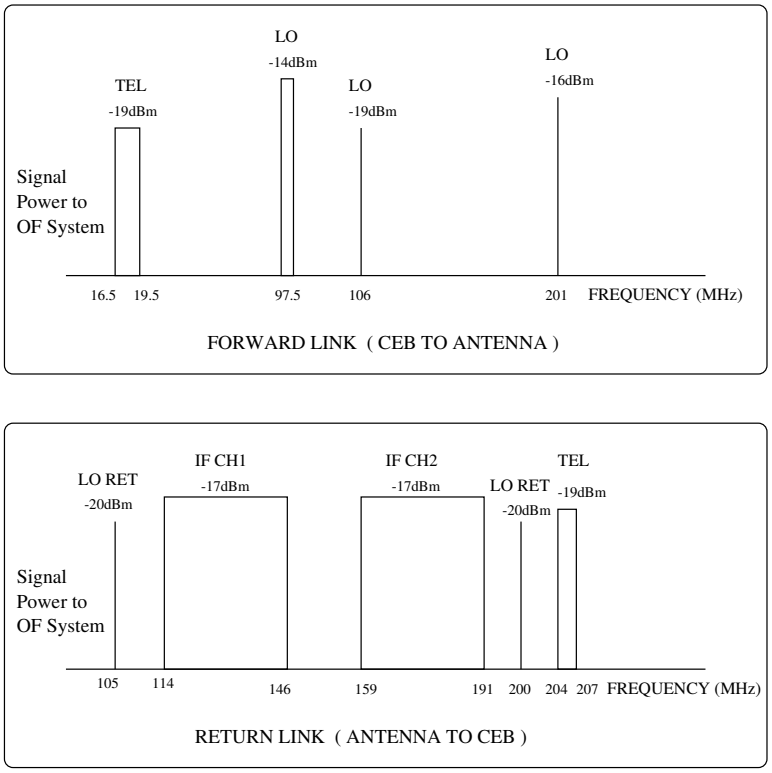


Figure 24.8: Schematic of the signals carried by the forward and return link. See the text for more information.

where the digital signals ("0" & "1") changes the frequency of the pseudo carrier to one of two frequencies, usually denoted as MARK and SPACE respectively.

If T is the duration of a bit, then the bandwidth (BW) occupied by the FSK signal is:

$$\left[\nu(\text{MARK}) + \frac{1}{\text{T}}\right] - \left[\nu(\text{SPACE}) - \frac{1}{\text{T}}\right] = \nu(\text{MARK}) - \nu(\text{SPACE}) + \frac{2}{\text{T}}. \quad (24.4.2)$$

For example, in the Forward Link, $\nu(\text{mark}) = 19$ MHz, $\nu(\text{space}) = 17$ MHz and $t = 4$ microseconds (i.e. corresponding to a data rate of 250 kbps). Therefore, the bandwidth of the FSK signal in the forward link is

$$\Delta\nu = (19 - 17) + \frac{2}{4 \times 10^{-6}} = 2.5 \text{ MHz}. \quad (24.4.3)$$

24.5 System specifications of the Control & Monitor system

24.5.1 Overview

1. Non-coherent FSK is used for data transmission over the optical fiber links The baud rate is 250 Kbits/sec.
2. Bit interleaving is used for multiplexing the Telemetry, Voice, Sync., Dial and Aux channels.
3. Data integrity is checked using polynomial and checksum error detection with ARQ capability.
4. The Bit Error Probability is 10^{-10}.

24.5.2 Bit rates available for various services

	Service	Bit Rate (kbps)
1	DATA (COMH - ANTCOM COMM)	125.000
2	VOICE (TELEPHONY VOICE)	62.500
3	DIAL (TELEPHONY SIGNALING)	15.625
4	SYNC (SYNCHRONIZATION PATTERN)	15.625
5	AUX1 (AUXILIARY CHANNEL1)	15.625
6	AUX2 (AUXILIARY CHANNEL2)	15.625
	TOTAL BIT RATE	250.000

24.5.3 Details of the various communication links

Type	Subsystems	Involved		Rate
Ethernet	UNIX WS	<==>	PC ROUTER	10.0 Mbps
Asynchronous RS232-C 10 bit	PC ROUTER	<==>	COMH	38.4 kbps
SDLC/HDLC	COMH	<==>	ANTCOM	125 kbps
Asynchronous RS485 11 bit	ANTCOM	<==>	MCM	9.6 kbps
Asynchronous RS422 10 bit	ANTCOM	<==>	SCC	9.6 kbps
FSK MODEM	COMH	<==>	ANTCOM	250.0 kbps
VOICE	CEB	<==>	ANTENNA	62.5 kbps

Chapter 25

The GMRT Correlator

D. Anish Roshi

25.1 Introduction

Chapters 8 and 9 covered the basics of correlator design and implementation. Recall that there are two popular types of correlators, viz. the FX and XF types. The FX design has a number of advantages including (a) low cost, (b) digital fringe stopping and fractional delay compensation and (c) minimal closure errors. The GMRT correlator is an FX correlator. The integrated circuit (IC) used for performing the FFT and the correlation is an application specific IC (ASIC) designed by the NRAO for the VLBA correlator. This chapter provides an overview of the GMRT correlator and discusses its various modes of operation. The material is meant as a guide the correlator users (i.e. astronomers). For details of hardware implementation see Tatke (1997).

The main considerations while designing the GMRT correlator were the following:

1. *Astronomical requirement :* Briefly, the correlator should have the capability to make continuum radio maps of all the stokes parameters as well as spectral line radio maps.

2. *Radio Frequency interference :* As a low frequency telescope the GMRT is highly susceptible to man made interference. To observe weak celestial sources in the presence of strong radio frequency interference (RFI), the dynamic range of the receiver system and the correlator should be large. If the RFI spectrum is narrow band, it may also be possible to edit it out from the data if the visibility spectrum is measured with sufficiently high resolution.

3. *Cost :* The overall cost of the correlator system should be kept at a minimum.

The last two requirements favor an FX configuration. Since the FX correlator inherently measures the visibility spectrum, any narrow band RFI can be edited out. To improve the dynamic range 4-bit sampling is used.

Recall that the GMRT has 30 antennas and that each antenna provides signals in two orthogonal[1] polarizations. The maximum operating bandwidth at all frequency bands is 32 MHz, which is provided as two 16 MHz wide baseband signals (corresponding to the two sidebands) for each polarization (see Fig. 25.1). From the basic block diagram

[1]All the frequency bands of GMRT except the L band are circularly polarized. At L band two linearly polarized signals are provided.

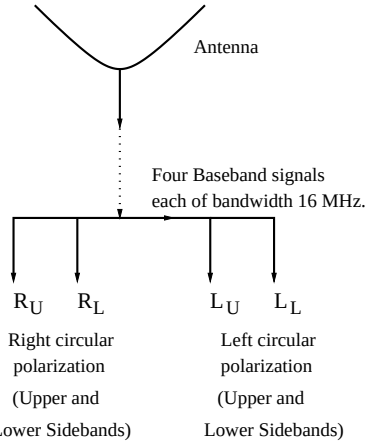


Figure 25.1: Schematic showing the four baseband outputs from each GMRT antenna. Each antenna is dual polarized and each polarization signal (which is of maximum bandwidth 32 MHz) is split into two sidebands, each of maximum bandwidth 16 MHz. At all frequencies of operation, except L band, right (R) and left (L) circular polarizations are measured. At L band two orthogonal linear polarizations are measured. The two sidebands are called the Upper Side Band (U) and Lower Side Band (L) respectively. So R_U is the upper side band of the right circular polarization, and so on.

of an FX correlator (see Fig 9.4 in Chapter 9) it is evident that the GMRT correlator should have $120\ (= 30 \times 4)$ ADCs, integral delay compensation units, number controlled oscillators, FFTs and fractional delay compensation units.

The total number of multiplier units required for the GMRT can be calculated as follows. The total number of cross products for a n element array is $n \times (n-1)/2$. If the self products are also computed then the total number of products is $n(n-1)/2+n = n(n+1)/2$. In an FX correlator these products have to be measured for each spectral channel. Since the GMRT correlator provides 256 spectral channels, the total number of multiplier units required is $n \times (n+1)/2 \times 256$. Further since, as discussed above, there are four baseband signals for each antenna, the number of multiplier units required goes up by a factor of 4. To measure all the four Stokes parameters the cross products between different polarizations need to be measured (see chapter 15), this causes the required number of multiplier units to increase by another factor of 2. Thus for $n = 30$ the total number of multipliers required is 9,52,320. However, to lower the cost and to simplify the hardware design the number of multiplier units in the GMRT correlator is only 2,38,080. To minimize the impact of this reduction in multipliers, the GMRT correlator has a a highly configurable design. Depending on the astronomical requirement the correlator can be configured to minimize the loss of information, for example in may spectral line observations it is not neccessary to measure all four stokes parameters. The following sections give an overview of the GMRT correlator and also discuss these different correlator configurations.

25.2 An overview of the GMRT Correlator

A simplified block diagram of the GMRT correlator is shown in Fig. 25.2. The basic units are the analog to digital converters (ADC), the Integral Delay compensation (Delay-DPC) subsystem, the Fourier transform and fractional delay compensation (FFT) subsystem and the multiplier-accumulator (MAC) unit. The data from the MAC output is acquired

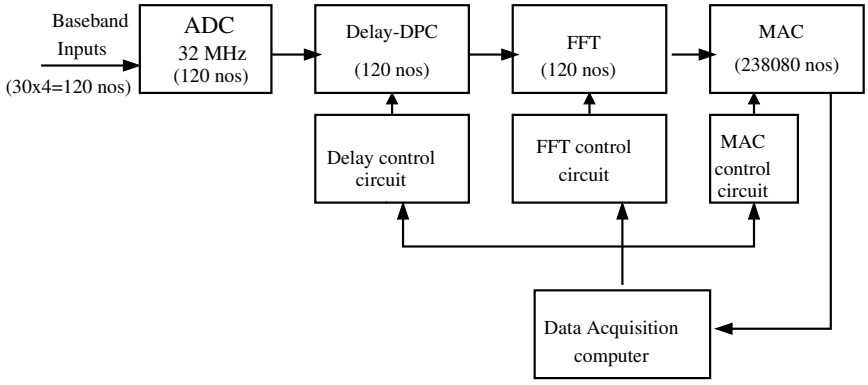


Figure 25.2: A simplified block diagram of the GMRT correlator

using a special purpose PC add on card. All of the subsystems, except the ADC, have DSP (digital signal processor) based control circuits. These control circuits are in turn controlled by the data acquisition computer, (i.e. the same machine which acquires the data via the add on card; see Chapter 26 for more details).

25.2.1 ADC

The GMRT correlator uses 6 bit, uniform quantization ADCs. The ADCs are designed such that a Gaussian random signal of 0 dBm power will have minimum distortion (see Chapter 8) and operate at a fixed clock frequency of 32 MHz. This means that when the input signal has a bandwidth of 16 MHz the digitized signal is Nyquist sampled. However at the GMRT, the input signal could have a bandwidth less than 16 MHz[2], for these signals the Delay-DPC effectively resamples the digitized signal so that down stream of the Delay-DPC unit the signal is Nyquist sampled.

25.2.2 Delay-DPC

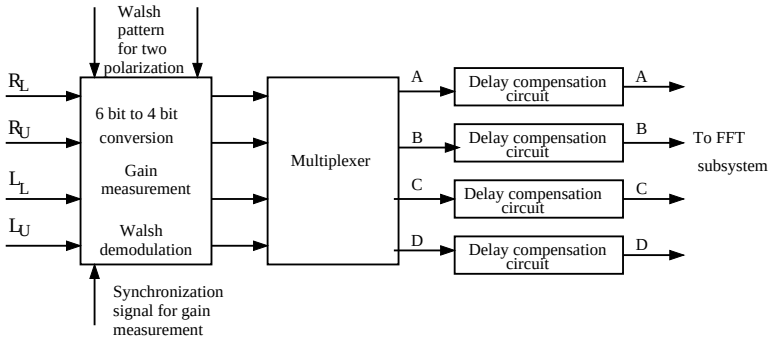


Figure 25.3: Block diagram of the delay-DPC unit of the GMRT correlator

The block diagram of the delay and data preparation card (delay-DPC) is shown in Fig. 25.3. Each basic unit of the delay-DPC takes the four outputs of ADCs corresponding

[2]Available bandwidths go from 64 kHz to 16 MHz in steps of 2, (see also Chapter 23).

to (see Fig 25.1) the signals R_U, R_L, L_U and L_L from a given antenna. These 6 bit quantized signals are rounded off to 4 bits and then sent to a multiplexer. The multiplexer has various modes; for example any one of the four inputs of the multiplexer can be mapped to all four of its outputs (A, B, C, D in Fig. 25.3). Other mappings include (a) $A = R_L$, $B = L_L$, $C = L_L$, $D = R_L$, and (b) $A = R_U$, $B = L_U$, $C = L_U$, $D = R_U$, which are used for polarization observations with the correlator. The multiplexer outputs are passed through a memory based integral delay compensation circuit (see Chapter 9). The delay compensated outputs are then fed to the FFT subsystem.

The rate at which data is written to the memory in the dly-DPC card is tunable. In particular it can be any one of $32/2^k$ MHz, where $k = 0$ to $k = 7$. This rate is chosen to be the Nyquist rate for the input signal bandwidth, i.e. for bandwidths smaller than 16 MHz, the rate is less than 32 MHz. However, the data is always read out at a constant rate of 32 MHz[3]. To maintain the data throughput, data from the memory hence has to be read out in an 'overlapping' fashion. This way of reading the data provides the facility to perform 'overlapping' FFTs (and hence an improvement in the signal to noise ratio) when the input bandwidth is less than 16 MHz.

The two other functions of the delay-DPC system are (a) gain measurement (b) Walsh demodulation.

25.2.3 FFT

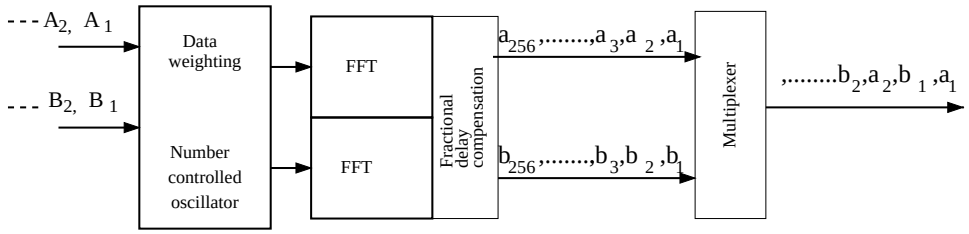


Figure 25.4: Block diagram of the FFT unit of the GMRT correlator

The block diagram of the FFT subsystem is shown in Fig. 25.4. The basic unit of the FFT subsystem takes two data streams (either A,B or C,D in Fig. 25.3) from the Delay-DPC. In the first stage, a weighting function can be applied to the 4 bit time series. The weighting function is software selectable, and can be chosen to be one of the standward "window functions" discussed in Chapter 8. This is followed by a number controlled oscillator (NCO), which does the fringe stopping (see Chapter 9). The two fringe stopped time series are passed through two sets of FFT engines, realized using VLBA ASICs, to perform Fourier transforms. Phase gradients are then applied to the spectrum of the signal to correct for delays smaller than the sampling interval (FSTC).

Each FFT engine can perform a Fourier transforms of maximal length 512 points. This length is software selectable to be 256, 128 or 16 points; it is even possible to bypass the FFT operation altogether. A 512 point FFT gives 256 channels, however in the next stage of the correlator (MAC) there are only enough multipliers for 128 channels per sideband per polarization. In the standard mode of operation, two adjacent FFT channels are hence averaged together in the MAC. A single MAC also acquires data from two FFT engines in a time multiplexed fashion. The data is multiplexed as shown in Fig. 25.4, where a_i and b_i are the spectral channels from the two FFT engines.

[3]In the final correlator it will be at 32.25 MHz.

25.2.4 MAC

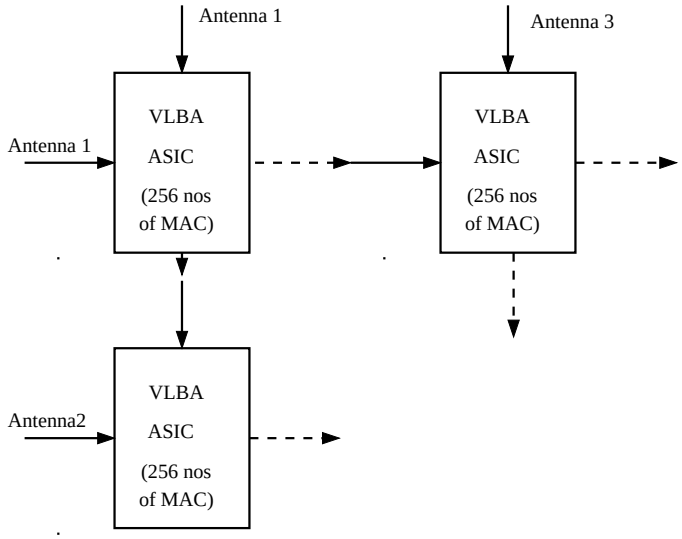


Figure 25.5: Block diagram of the Multiplier-Accumulator (MAC) unit for the GMRT correlator

The Multiplier and Accumulator (MAC) is, hardware wise, the most complex subsystem of the correlator. The MAC takes the FFT outputs computes the cross and self products and accumulates them for a maximum of 128 ms and a minimum of 4 ms. A schematic of the configuration of the multipliers is shown in Fig 25.5. Each MAC unit consists of 256 accumulators. The MAC can be configured in several different modes. As described in more detail below this flexibility allows the GMRT corrrelator to be used to make a wide variety of observations. Data from the MAC unit is read out by the Data Acquisition System (DAS) using a special purpose add on card on a PC (see Chapter 26 for more details).

25.3 Modes of operation of the GMRT correlator

As mentioned above the total number of multipliers available in the correlator is less than that required for the measurement of all four stokes parameters in all spectral channels for all sidebands of all antennas. Instead ,the correlator is configurable in various ways. Some configurations would sacrifice polarization measurements for improved spectral resolution, while others allow the measurement of all four stokes parameters at the expense of total bandwidth. The most commonly used configurations of the correlator are described in some more detail below.

25.3.1 Non-Polar Mode

In this mode of observation the visibility for only one of the two polarizations can be measured in each of 256 spectral channels for all baselines (including self correlations). The maximum bandwidth possible is 2×16 MHz. Thus the observation will have half the total sensitivity of the GMRT.

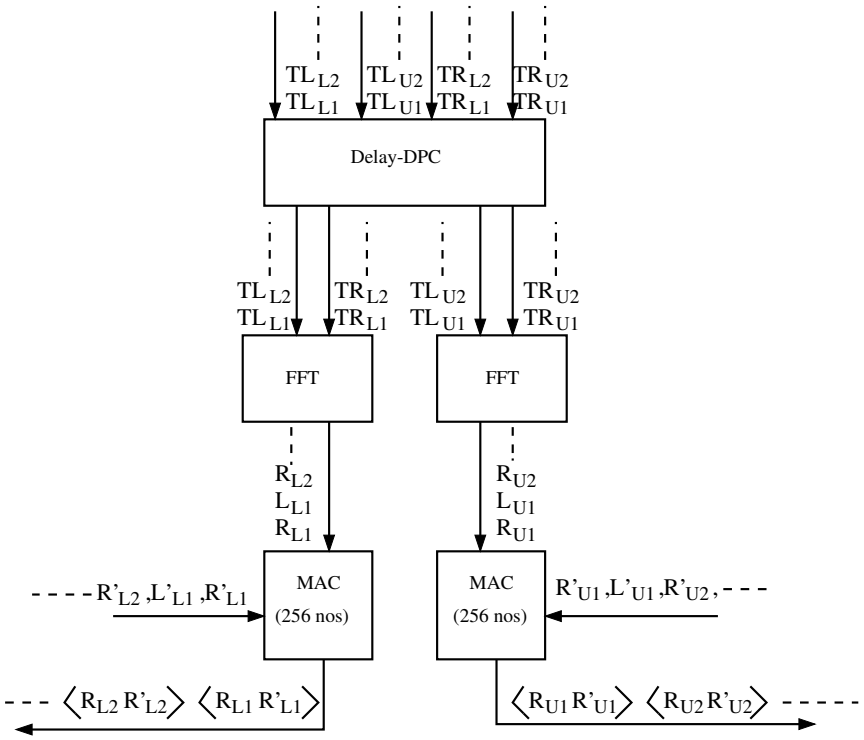


Figure 25.6: The signal flow in the GMRT correlator for the Non-Polar mode. Signal names preceding with a T indicate time series and angular brackets denote time average. Numeric subscripts indicate the sample number for time series data and the frequency channel number for spectral data.

In the non-polar mode the required sampling frequency is selected in the delay-DPC system. The multiplexer in the delay-DPC is configured such that the data flow is as shown in Fig 25.6. $TR_{Ui}, TR_{Li}, TL_{Ui}$ and TL_{Li} are the time series of the four baseband signals for a given antenna. The FFTs of these time series are R_{Ui}, R_{Li}, L_{Ui} and L_{Li} respectively (for $i = 1$ to 256). $R^{'}_{Ui}$, $L^{'}_{Ui}$ are the corresponding signals from a second antenna. The MAC mode is selected such that the 256 channels in one of the sidebands of one of the polarizations (in this case R_{Ui}) is integrated in its 256 accumulators. A second set of MACs integrate the signals from the second sideband of the same polarization (in this case R_{Li}).

25.3.2 Indian-Polar Mode

In this mode the visibility from both polarizations of all 30 antennas is measured but the number of spectral channels per baseline is limited to 128. Thus the spectral resolution is half that of the Non-Polar mode but the maximum bandwidth that can be observed in this mode is 32 MHz. Thus the observation will have the full sensitivity of the GMRT. For a non polarized source, this mode measures stokes I, and it is the most commonly used mode in interferometry.

The Delay-DPC configuration for this mode is similar to that of the Non-Polar mode. The MAC is configured (see Fig 25.7) such that the adjacent channels of the same polarization are averaged together, thus reducing the number of channels from 256 to 128.

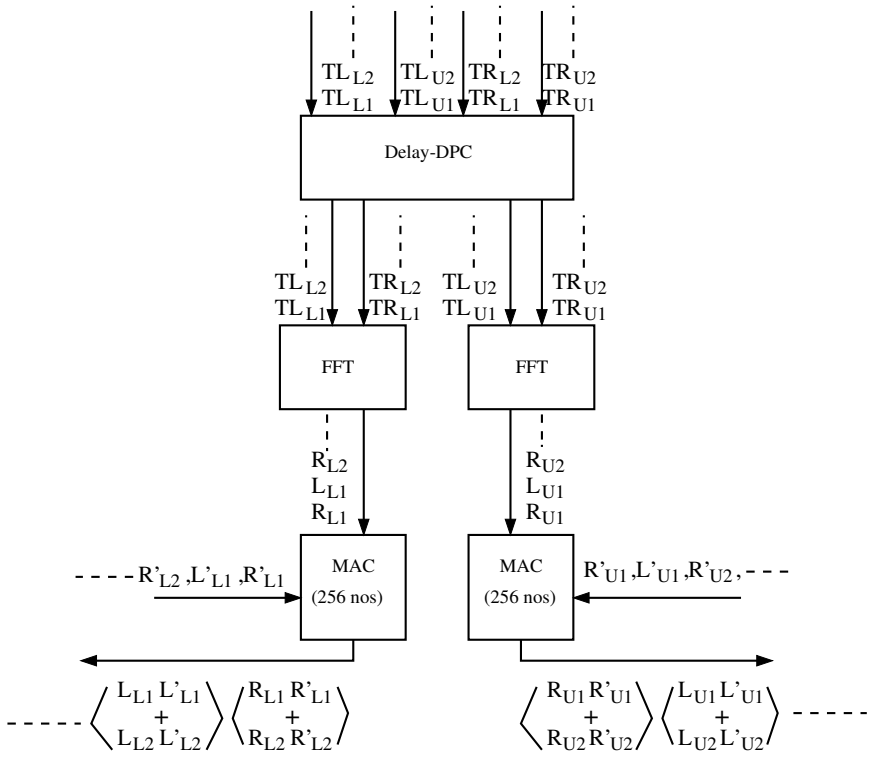


Figure 25.7: Signal flow in the GMRT correlator when it is configured in Indian-Polar mode. See the caption of Figure 25.6 for other details.

Since each MAC unit has 256 accumulators, the other 128 accumulators of the same MAC are used to integrate the data from the second polarization of the same baseline.

25.3.3 Polar Mode

All the cross products needed for measuring the four Stokes parameters (see Chapter 15) are measured in this mode. The number of channels available per baseline is again restricted to 128 and further one sideband from all 30 antennas is processed. Thus the maximum possible bandwidth in the Polar Mode is 16 MHz, as opposed to 32 MHz in the Indian Polar mode (which measures Stokes I for unpolarized sources), and the spectral resolution is also half of the maximum possible in the Indian Polar Mode.

The delay-DPC multiplexer is configured so that the data flow will be as shown in Fig 25.8. The data from one side band for both polarizations (in this case R_{Ui}, L_{Ui}) is multiplexed to get the required data sequences. The MAC is configured in the polar mode such that it measures the cross product of the two polarizations in addition to the cross products of a polarization with itself. Adjacent channels of the cross product of one of the polarizations (eg: $R_{Ui} \times R^{'}_{Ui}$) are averaged and integrated in 128 accumulators of the MAC. Unlike in the Indian-Polar mode, the second set of 128 accumulators integrate the cross product of the two polarizations (eg: $R_{Ui} \times L^{'}_{Ui}$). Similar measurements of the second polarization (i.e. in this case, $L_{Ui} \times L^{'}_{Ui}$ and $L_{Ui} \times R^{'}_{Ui}$) are made in the second MAC. Thus all required cross products are measured, from which, as described in Chapter 15 all four Stokes parameters of the source can be computed.

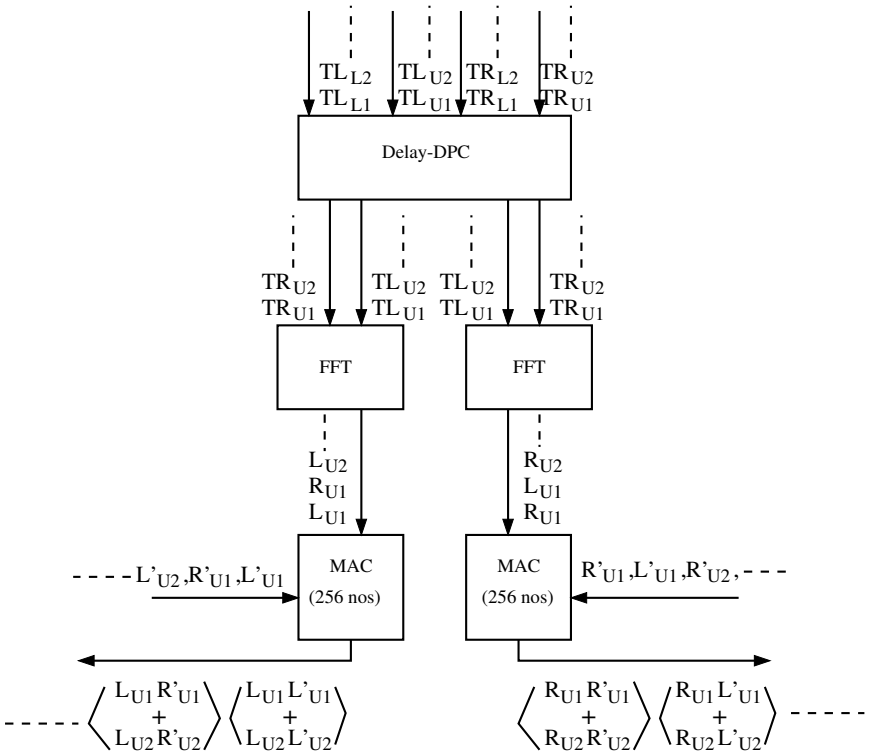


Figure 25.8: Signal flow in the GMRT correlator in the Polar Mode. Signal names preceded by a T indicate time series and angular brackets denote time average. See the caption of Figure 25.6 for other details.

25.4 Further Reading

1. Tatke, V. M., 1997, M.Sc.(Engg) Thesis, Indian Institute of Science, Bangalore.

Chapter 26

The Data Acquisition System for the GMRT

R. K. Singh

The Giant Meterwave Radio Telescope (GMRT), an array of thirty antennas, situated at Khodad, Narayangaon, is by and large completely controlled by observers, sitting at their desks, with the help of a centralized networked system of computers at the Central Electronics Building (CEB).

The software system that achieves this can be broadly divided into two parts, (i) one that deals with the control and monitoring of the antennas (ONLINE) and (ii) one that deals with the control, monitoring as well as data acquisition (and processing) of the digital backends (DAS). Both of these systems have a large number of small embedded subsystems that can be controlled across the network. In discussion in this chapter is limited to the DAS.

Any reasonable Data Acquisition System (DAS) for the GMRT must fulfill the following major requirements.

1. Interfacing with the hardware for acquisition of data.
2. Control of the correlator both for configuration as well as required periodic updates of parameters.
3. Monitoring of the status of the correlator and flagging the data appropriately.
4. Online processing of the data for reduction and archiving.

A functional description of the GMRT correlator can be found in chapter 9, and it is assumed that the reader is familiar with it.

26.1 Data Acquisition

Let us first estimate the maximum data rate produced by the GMRT correlator. For each antenna there are four associated data streams (two sidebands USB and LSB, in each of two polarizations RR and LL) each of 16 MHz bandwidth. Therefore, one requires four sampling points per antenna with sampling rate of 32 MHz. The present correlator has only 60 sampling points, and handles only one side band. For each stream there is a corresponding FFT unit in the correlator. This FFT unit carries out a 512 data point transform in real time, i.e. one 512 point transform every 16μsec. The 512 point transform

corresponds to 256 complex numbers, i.e. 256 frequency channels. The time to acquire 512 data samples, i.e. 16μsec is called a FFT cycle. The number of distinct pairs of antennas (including an antenna with itself, i.e. self correlations) that can be made from 30 antennas is $30 \times (30+1)/2 = 465$. Therefore, 465 Multiplier and Accumulator (MAC) units are required to correlate all data from 30 antennas. Each MAC unit accepts 4 data streams i.e. two polarizations from two antennas (it makes no sense to correlate USB with LSB) and multiplies them to produce 128 complex numbers each for two polarizations, or 256 values for one polarization. In either case, it is 256 complex numbers, which the MAC units sum for a duration of a STA (Short Term Accumulation) cycle, which is 4096 FFT cycles. One STA cycle is equivalent to $4096 \times 16\mu$sec $= 66$ms. The MAC data format is such that 4 bytes encode one complex number. The actual number of MACs in the correlator is 176×3 (there are 3 Racks with 176 MAC units each) or 528, i.e. there are $528 - 465$ redundant MACs. Therefore the total amount of data produced per second per side band is $528 \times (1\text{sec}/66\text{ms}) \times 256 \times 4$ bytes $= 8$MB. The total data including both side bands would be $= 16$MB/sec.

16MB/s is a huge data rate to be sustained on any general purpose machine, or to be stored on any media. This means that would would need another piece of hardware in the correlator to carry out Long Term Accumulation (LTA). Such a hardware element was planned, but has not been implemented so far. Instead, in the present correlator system, the STA cycle has been configured for 8192 FFT cycles, i.e. the STA cycle duration is 132 ms. This brings down the sustained data rate per side band to 4 MB per second. Even this requires a special interface cards to input the data into the general purpose machine for processing. The host computer currently used for the DAS is a pentium based machine running the Linux operating system.

26.2 Correlator Control

The correlator system for the GMRT serves the following four major functions depicted in the schematic shown in the Figure 26.1.

1. Analog to Digital Conversion.
   As we mentioned earlier that the full GMRT would require $30 \times 4 = 120$ sampling points, for thirty antennas with two side bands in each of two polarizations. The current correlator system suffices half of this requirement. The sampling takes place at 32 MHz in order to have data for a maximum of 16 MHz bandwidth. There is no control required for this unit of the correlator.

2. Delay DPC unit.
   Signals originating from a given point in the source, and received via two different antennas, traverse different path lengths before they arrive at the samplers. This different path lengths arise because of the different locations of the two antennas as well as the different cable lengths between the two antennas and the correlator (and are called the geometric delay and the fixed delay respectively). As discussed in chapter 4 the geometric delay changes with the Hour Angle of the source. In order to compensate for the differential delays that the signals from different streams have suffered, the Delay unit has to be periodically updated with the current values of time delays that have to be applied. Therefore, delay values are required to be transmitted from the host computer down to this unit of the correlator periodically. The Delay unit of the correlator can delay only for the integral number of sampling time intervals, which is 1/32MHz. Finer delay corrections are made in the FFT unit, where delay values are converted into a phase gradient across the band.

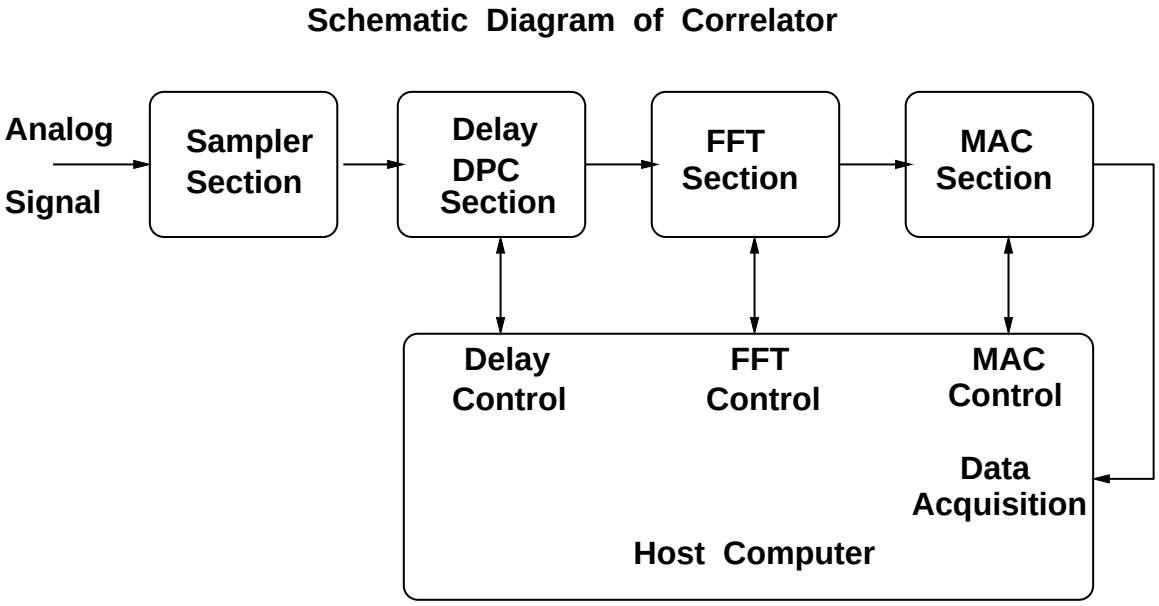


Figure 26.1: The schematic block diagram of the correlator. The four major blocks of the hardware as well as the host computer are shown.

3. FFT unit.

   This section of the correlator carries out 512 points FFT every $1/32\text{MHz} \times 512 = 16\mu$sec. The two other major tasks that this unit performs are, 1) fringe phase subtraction, and 2) fractional sampling time delay correction (FSTC, see chapter 4). This requires periodic updates of the phase and FSTC values that are to be applied, which has also to be supplied by the host computer.

4. MAC unit.
   The MAC unit can be configured in a variety of modes. This configuration is usually done by the host computer during the initialization sequence.

All these units (except the samplers) need to be initialized. The exact initialization required depends on the observing mode. To achieve this, at the start of the observation appropriate pieces of programs are loaded into the controllers, which then behave like embedded systems.

26.3 Monitoring the health of the correlator

The quality of the data acquired, depends on the health of the correlator for the duration of the observation. It is desirable to have flag bits in the recorded data indicating the state of the correlator at the time of the observation. There exists a planned set of parameters that are to be monitored and a method of transmitting such information to the host computer exists, but the actual monitoring has not been implemented yet. Hopefully, this will be done sometime in the future.

26.4 Online processing of the data

26.4.1 The network of acquisition and processing

The actual data acquisition and online processing of the data are carried out over a network of computers. The connectivity of the network is shown in Figure 26.2.

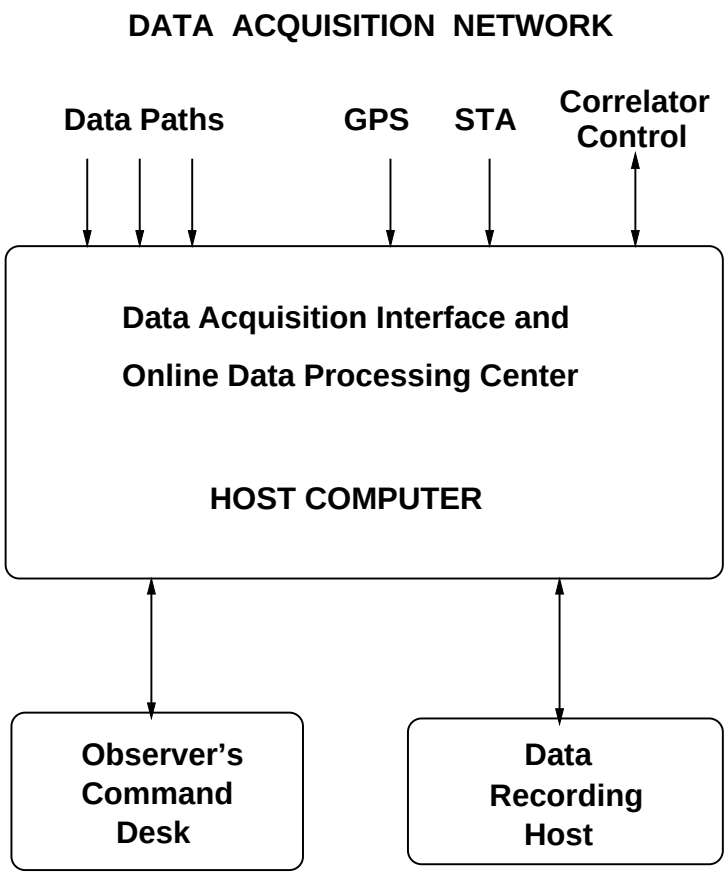


Figure 26.2: The various connections that exist to the correlator host computer. The connections on the upper side are to the correlator hardware, while on the lower side are to other computers on the observatory network.

The present correlator handles one side band of 16 MHz bandwidth. The MAC unit consists of three racks, and the data is hence received over three channels. The host computer uses a custom made card to accept data at three separate ports. In order to maintain standard time for reference for the data blocks, the host computer is also provided with minute pulses from the Global Positioning System (GPS). One of the fundamental cycles of the correlator operation is the Short Term Accumulation (STA) cycle, over which the MAC unit accumulates the data; it is equivalent to 8192 cycles of FFT. The clock edges that signify the beginning of STA cycles are also provided to the computer. The STA cycle timings are used to maintain the synchronization in the communication between the host computer and the correlator components. This STA clock edge also signifies the beginning of the STA data cycle. The actual communication for the control of the correlator components takes place via another custom made, in-house developed

card which can communicate to sixteen different components sequentially.

So far, the connections were described on the correlator side of the interface (host) computer. The major data processing and correlator control is carried out by the same host computer. On the other side of the interface computer is the 'network' i.e. a connection via the standard Internet Protocol over Ethernet. One of the computers on this network side is used by the observers to control and monitor the entire GMRT system. Commands are issued by this computer ('Observers Command Desk' in Figure 26.2) to the host computer which in turn manages the correlator. Another network component, the 'Data Recording Host' (figure 26.2) is used for data storage and off-line processing. This computer is the main compute server and has a huge data storage capability.

26.4.2 The software layout

A large number of separate programs work together in order to achieve the parallelism that is a natural requirement for a real time acquisition, processing, and control. These units will now be described one by one. The Figure 26.3 shows the important components of the whole package.

The low level programs

There are three low level components of the package that establish communications with correlator.

1. The data collector.
   As stated above each MAC racks passes data over a separate channel, therefore, this low level software, the data collector, accepts the data from three channels. Each MAC rack is composed of 11×16 MAC chips each of which gives out 256 complex numbers per STA cycle. Each complex number is coded in 32 bits, i.e. two 16 bit real numbers. Each MAC chip has two buffers; one is used for accumulating the incoming data over a STA cycle, and another is used for transmitting data accumulated during the previous STA cycle. Therefore, one MAC rack outputs $11 \times 16 \times 256 \times 4$ bytes $= 176$ KB of data per STA cycle. One of the bits of the 16 bit word is reserved to indicate the beginning of a fresh STA cycle.

   We use a custom made PCI bus based interface card to input this data. The interface card has four ports, one of which is unused. These ports are 16 bits wide data ports with separate control lines. Three of the ports are configured to accept the data on an independent external clock. The interface card has total 64/128 KB total memory, which is equally distributed among the 4 ports. Three ports are configured for streaming mode operation, where the 16/32 KB memory for a port is divided into two halves, such that when one half is full, an interrupt is generated for the host computer's CPU for transferring this data into the memory of the computer. While this transfer continues from one half of the data, the external data continues to fill the other half of the memory for each port. This process continues endlessly.

   The user level software programs are expected to process the data at this rate on the average. This low level program receives the data from all the three channels and stores them in separate local buffers. For every block of data corresponding to the half of port memory (8 KB/16 KB, depending on the port memory size), the PC time is noted. This time corresponds to the end of the block of data just arrived, and can also be thought of as the time corresponding to the beginning of the next data block. This time will be used for time stamping of data by the next level software referred to as 'acq' in Figure 26.3.

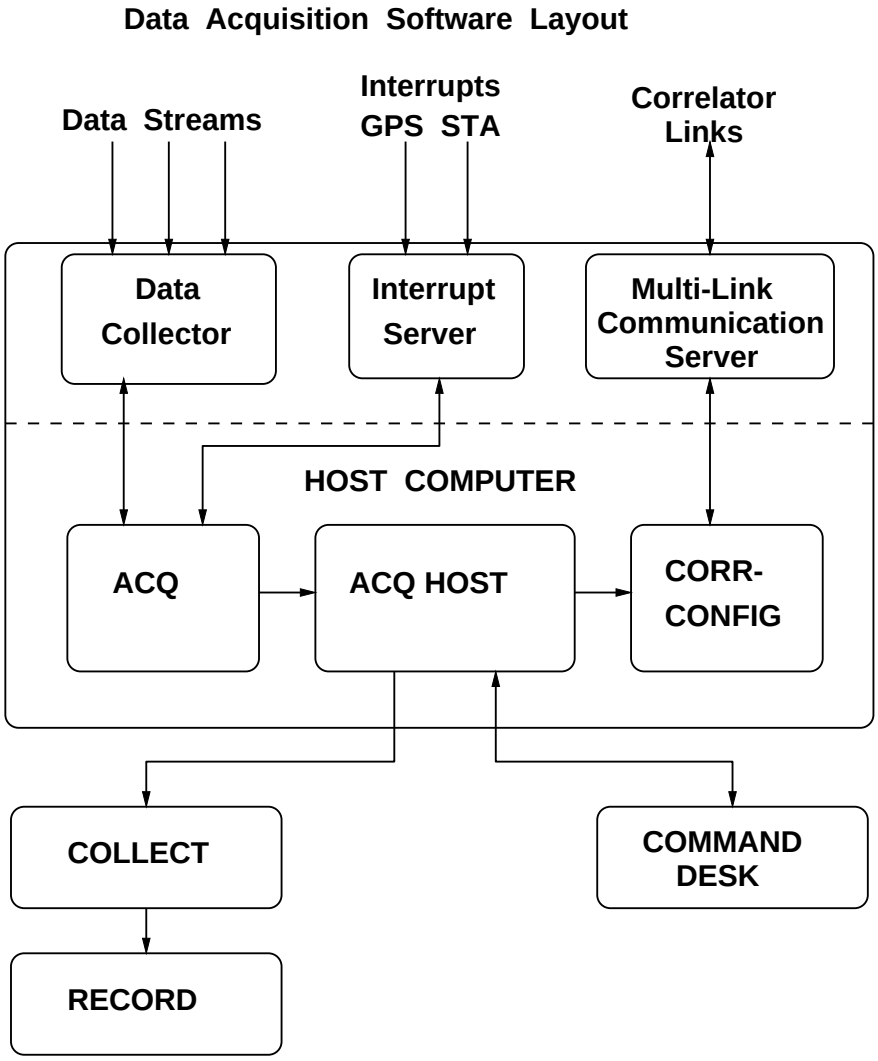


Figure 26.3: All the major programs used in DAS are shown here. The components in the solid box and above the dashed line are 'low level programs' that deal with the hardware interface. The ones inside the solid box but below the dashed line are 'high level programs' which deal with controlling the hardware as well as collecting and processing the data. The programs below the solid box are 'peripherial' or 'network' programs.

The low level program makes no effort to look inside the data buffers and synchronize three streams into one logical stream.

The data collector stores a good fraction of a second's worth of data in three separate circular buffers. Hence it is highly unlikely that the higher level program 'acq' (even if it is temporarily busy elsewhere) will be unable to drain out this data before the buffers overflow.

2. GPS and STA interrupt server.
The GPS minute pulse, and the STA clock edge are used to generate interrupts to the host processor. When an interrupt occurs, the interrupt server program notes

down the PC time and maintains a list of last few such times. On request from the higher level programs, the interrupt server simply supplies this list (for the GPS or STA interrupts, as requested). For both STA and GPS the time interval between two successive interrupts should in normal circumstances be a fixed constant, and therefore, the interrupt server can optionally be requested to build statistics on time intervals at which the interrupted occurred and attempt to interpolate over temporary glitches. Note that as discussed further below the PC time noted by the interrupt server could in principle be wrong for a variety of reasons. These errors need to be recognized and corrected for by the higher level programs.

3. Multi link communication server.
   This low level program communicates with the correlator through a home made custom communication protocol. It is primarily used for configuring the correlator. A higher level correlator configurator program sends and receives data through this low level program.

The high level programs

There are three high level processes that run on the correlator host computer which are responsible for the real time processing of the data as well as correlator control.

1. The acquisition center, the 'acq' program.
   This program has three major responsibilities.

   (a) It acquires the data from the data collector, and looks for the STA cycle markers described above.

   (b) It still maintains the data in three separate streams, but marks each data block in each stream with a sequence number. The sequence number is derived from the time noted as the beginning of STA cycle for that block. This basically synchronizes the three streams.

   (c) The program places the data (in three separate circular buffers), and also considerable book keeping information in a common memory resource (shared memory). This shared memory can be accessed by other programs for further real time processing. Note that more than one program can simultaneously read the shared memory, but only one program, viz. 'acq' can write to the shared memory.

2. The correlator configurator, the 'corr_config/fstop' program.
   This program is has two main functions

   (a) To initialize the embedded correlator controller cards and prepares them for the specified observation mode.

   (b) To periodically set the model values (delay, fringe stop, FSTC). The delay value is transmitted to the Delay section of the correlator and the fringe stop and FSTC to the FFT section.

   The current model values have to be down loaded to the embedded systems at a time and in a manner such that the normal workings of those systems are not affected. Further the protocol should be such that higher level program must know in advance at what instant of time these new values will be made effective by the low level software. For this purpose, two basic norms are followed.

(a) The model values are down loaded at times which are not close to the beginning of a new STA cycle. At the start of an STA cycle the embedded systems have several time critical tasks to perform.

(b) The embedded systems are initialized with a STA cycle sequence number. This sequence number is incremented by all the embedded systems at the start of new STA cycle, so that all the embedded systems are properly synchronized. When model values are down loaded, the configurator also passes the STA sequence number at which these values are to be effected. The activity across all the components of correlator are hence synchronized. The protocol takes into account the fact that erroneous conditions which require corrective action could arise.

3. The acquisition manager, the 'acqhost / acq30' program.
This program (which is the the central program) receives interactive commands on one side (i.e. over the network) and on the other side (i.e. the correlator host computer and the correlator's embedded systems) controls all activities much like a band master. It is the responsibility of this program to continuously ensure the logical coherence of the observation.

When 'acq30' is started, it scans a set of files (that could also be specified via user commands) that describe the correlator hardware configuration. These files contain information such as what are the coordinates of the antennas, which antenna is connected to which samplers, which samplers is connected to which FFT pipeline, and which FFT pipelines are connected to which MAC chips. From this information 'acq30' builds a complete mapping of all the connectivities in the correlator. 'acq30' reads the data from the shared memory created by 'acq'. Recall that in this shared memory one has 'raw' data, i.e. the data is organized into three streams (one per rack) and in each stream the ordering depends on the ordering of the MAC chips in that rack. 'acq30' merges these three streams into one stream using the sequence number information planted in each data block by 'acq' (see above). Further it also notes the common IST arrival time of the data corresponding to a given STA sequence. It was originally intended that the merging of the three streams of data into one stream would be effected by the embedded software of the correlator, but this was not realized. Once the data has been merged into one stream, 'acq30' uses its knowledge of the correlator connectivity, to determine which piece of the data corresponds to which antennas, and which samplers, and which FFT pipelines. The 'acq30' program however has no way of determining whether the specific antennas whose data is being read are all pointing to the specified source or not. 'acq30' simply trusts that when a scan is started (see below) the antennas are pointing at the specified source. Given the source position and the antenna co-ordinates, 'acq30' computes the appropriate delay, fringe stop and FSTC values. The calculation of such quantities for a given point of time is referred to as 'model calculation', and the values as 'model parameters' or 'model values'. For as long as a given scan continues, 'acq30' computes the model parameters periodically and sends them to the appropriate program (corr_config/fstop) to be set. From its mapping of antennas to samplers to FFT pipelines, 'acq30' can determining which delay value to send to which delay control card etc.

'acq30' also converts the data format from the 'MAC format' described above (i.e. 4 bytes for each complex word) to the IEEE standard floating point format. The Long Term Accumulation (LTA) is also performed by this program. The observer specifies the number of STA cycles for which the data should be further summed before recording. This summation is carried out at the same point as the conversion

to IEEE floating point format. After performing LTA 'acq30' sends the resulting data on the network to a separate program 'collect' which runs on another computer (the 'Data Recording Host' in Figure 26.2). 'collect' makes this data available to other monitoring and recording programs.

'acq30' accepts (and passes on to down stream programs) several user commands, of which the four major ones are:

(a) Init:Initiate an observation session. During initialization time parameters which will remain constant during the observation session (eg. the bandwidth, the polarization mode, the number of channels to record, which antennas to record etc.) have to be specified.

(b) StartScan: Start a scan on a specified source. Required parameters include source position, observing frequency, observer's name etc.

(c) StopScan: This indicates the end of the present scan. At this stage one can start a new scan on the same or different source using the 'StartScan' command.

(d) End Close the observation session. On receiving this command 'acq30' informs all other dependent programs to shut down and then shuts itself down.

The peripheral programs on the network

There are three other peripheral programs on the network that complete the entire data acquisition chain. They are:

1. 'dassrv'.
   A program 'dassrv' running on the 'Observer's Command desk' (see Figure 26.2) interfaces between the main GMRT control program 'ONLINE' and 'acq30' running on the 'Host Computer'. An observer sitting at this desk hence has therefore control over both the antenna system and the data acquisition system. Commands can be entered from this desk interactively or via an observation file.

2. 'collect'
   As discussed above once the LTA of the data has been done by 'acq30', it is sent (along with the associated model values, times, and flags) to 'collect' (which runs on the 'Data Recording Host' in Figure 26.2). 'collect' performs the simple job of creating a shared memory resource and placing the data it receives into it in a well specified format. Any number of programs can then access this data either for recording, archiving, or online monitoring.

3. 'record'
   This is the last in the chain of Data Acquisition System (DAS). As stated earlier, this program takes the data from the sharable resource created by 'collect' and writes the data in a specific format (locally called the 'lta format') onto the hard disk. From here it can be transferred to tape for long term storage.

Sequence of program execution

When the correlator host computer boots, the three low level programs (in Figure 26.3 are started. The correlator itself is usually started before this computer is booted, but in principle the correlator can be brought up at any time before higher level programs are executed. Next, 'acq' is started, and it connects to two low level programs, the data collector, and the interrupt server. As long as 'acq' is alive it takes data from the data

collector, time stamps it using the information from the interrupt server (for more details on time-stamping see below) and then places the time-stamped data in a shared memory. 'acq' does not care whether any such programs exist to use this shared memory.

At this point, the correlator configurator program (corr-config), is run to setup the correlator in the required mode. Ideally, this should be done when 'acq30' is instantiated, but as of now this is not recommended. 'acq30' and the peripheral programs ('collect', 'dassrv', 'record') can then be started. At this point the entire data acquisition chain is up. The 'record' program can be started anytime after 'collect' has been started. Any number of 'record' programs could run at a time. Similarly, any number of monitoring and online display programs can run simultaneously to follow the progress of the observation.

A large number of small tools are developed to look at the raw data as placed in the shared memory of 'acq' on the host computer. These programs provide invaluable tools to debug the correlator problems, and also to make consistency checks at run time. These programs do not interfere with the normal observations, therefore, any number of them could be run at a time.

26.4.3 Time stamping of data and it's accuracy

The time at which a given astronomical signal was received by the GMRT is a fundamental parameter. At the GMRT we attempt to stamp the data with an accuracy of better than 100μsec. This time resolution is very good given our baseline lengths and operating frequencies.

A complex algorithm is followed to keep the time information to within 100μsec accuracy. As stated earlier time information is collected from three sources, (i) the Host Computer Clock (usually called the 'PC Clock'), (ii) the GPS minute pulse and (iii) the STA pulse. All these sources could possibly be in error. Below we describe in more detail the time information available.

1. The most basic source of time is the data itself. Recall that the low level program, 'data collector', reads the PC time at the end of every 16 KB of data that is received from the MAC. This time information is available to 'acq' along with every block of data. The 'data collector' has no knowledge about where in the data a new STA cycle begins because it simply collects the data coming at an uniform rate as a continuous stream. 'acq' figures out the beginning of the STA cycle, by looking for the synchronizing bit in the data (see above). Once the synchronizing bit is found 'acq' can use the time-stamping of each data block to associate a PC time to the beginning of the data block of a given STA cycle.

2. The GPS minute pulse is provided to the Host Computer as an interrupting pulse. The 'interrupt server' reads the PC time at this event and maintains a list of the last several such times. 'acq' goes through this list, looking for a set of contiguous events for which the interval between successive events is the same to within 100μsec. Once it finds such a set, it uses their time of arrivals to establishes a linear equation between IST and the PC time. This equation (which is communicated to all programs which require to know the exact time) is used to go back and forth between IST (which is essentially what is required for astronomical purposes) and the PC clock, (which is what is readily available to all programs). The linear equation accounts for an offset as well as a constant drift between the PC time and IST. The equation is continuously updated using the latest "good" GPS pulses, so higher order drifts are also taken care off.

3. At the start of a new STA cycle a synchronizing pulse is sent to all the components of the correlator as well as the host computer. This pulse is picked up by the 'interrupt server' which maintains a list of the time of arrivals of the last several such pulses and provides it to 'acq' on request. As discussed above, this same information is acquired also from the synchronization bit in the data. This redundancy allows for a consistency check to be made. The STA interrupts are used for two purposes (i) to make this redundancy check, and (ii) to set up a linear equation between the STA pulse time of arrivals and the PC time. Once this equation has been set up, a sequence number can be assigned to any given STA pulse based on its time of arrival.

There are several possible sources of error in the time keeping, viz.

1. It is possible that the system (OS) may be busy when a given event (say GPS or STA pulse arrival) took place, and by the time it registers this event and notes down the PC time an unknown delay has occurred.
2. The basic unit of scheduling for the OS is a unit called one CLOCK TICK (10ms), and occasionally the OS makes an error of exactly one CLOCK TICK. This error is the simplest to detect and correct for.
3. Sometimes, (mainly because of hardware glitches), an expected event does not take place, or there a number of spurious events.

The fundamental assumption that is used in correcting for these errors is that the GPS cycles, STA cycles, and the STA coded in data are all driven by external agents, therefore, even if there is a momentary glitch, sanity will prevail again after a while. Similarly drifts in the time of arrivals are expected to be slow and not discontinuous. The only discontinuous event that could occur is the missing of one CLOCK TICK, but since this leads to an error with a well determined signature (a jump in time by exactly one CLOCK TICK), it is very easily recognized. In practice this sort of error rarely occurs. 'acq' uses a number of complex heuristics as well as continuous redundancy checks to interpolate short term hardware glitches and experience has shown that it is quite robust to short term glitches.

While ignoring short term glitches, 'acq' should nonetheless track slow drifts of the PC time. To do this, after every 16 STA cycles it makes an attempt to arrive at a new equation between the STA sequence number and PC time. This equation is not accepted unless it is found to be matching within 100 microseconds of the previous equation. Similarly, every minute when a new GPS pulse is received, an attempt is made to arrive at a new equation.

Sometimes 'acq' cannot update its equations for more than a threshold time interval (because the updated equations are very different from the existing equations). This indicates that a much more serious problem has occurred. Generally what has happened at this point is that the PC time has jumped. 'acq' then attempts to reestablish equations afresh from a reasonably long sequence of good time intervals for STA and GPS. For example, it is demanded that a contiguous set of 4 GPS pulses must be within 100 microseconds of error, before the IST equation is accepted. Since 'acq' is starting afresh, the deviation from the previous equation is not checked as it would normally be. Similarly for the STA equation it requires that least half of 64 intervals in a stretch of 64 STA cycles must be good before deriving a fresh STA equation. When the PC time jumps a time accuracy of less than 100μsec is not guaranteed for the time interval between the jump and the time when a new set of equations are established. However, there are a limited number of environmental factors which cause the PC time to jump. These are known and are avoided during observations.

26.5 Further desirable features

A few of the features that would be desirable are:

- Monitoring the health of the digital system (correlator) is possibly the most important in this list. One should identify the parameters that are to be monitored and define actions to be taken on specific conditions. The conditions should also get flagged in the data.
- The communication between the 'acq30' and the 'corr-config' is not good yet. It causes variety of problems. For example, the 'acq30' has no way of figuring whether the control values that it wanted set at a given time have been set, and if so, whether at the exact time specified.

  Therefore, a better connection with a well defined protocol, must be created between 'acq30' and correlator configurator in order to achieve the following goals.

  - To reduce the time delays between the time when the request is made and the time when the control values are set in the correlator.
  - To have better timing control in setting the values.
  - To be able to report back to 'acq30' the errors encountered by the configurator, so that the corrective measures are initiated to counter the error, and also to allow 'acq30' to flag the data appropriately.
- Communication between the program for 'online control of antenna' and 'data acquisition system' is truly minimal. A better ties with the 'online system' is required, so that the problems in antenna pointing, or with the analog systems can be reported to the 'acq30' for appropriate flagging of the data. At the moment an user has to take the log file generated by 'online system' and associate such conditions to the data manually.

The critical low level tools and routines already exist for the fulfillment of the items in the list.