Chapter 2

Data model and corruptions

> It is a capital mistake to theorize before one has data.
>
> Sir A. C. Doyle

ABSTRACT
A number of experiments are set to measure the 21-cm signal of neutral hydrogen from the Epoch of Reionization (EoR). The common denominator of these experiments are the large data sets produced, contaminated by various instrumental effects, ionospheric distortions, radio frequency interference (RFI) and strong Galactic and extragalactic foregrounds. In this Chapter we present the *Instrumental and Data Model* that will be the basis of the signal analysis for the LOFAR (Low Frequency Array) EoR Key Science Project (LOFAR EoR KSP). Unlike other observations that will be done using LOFAR, the detection of the EoR signal requires error reduction below one part in a million. Therefore, it is essential to have an accurate and realistic understanding of the causes of these errors, before we attempt to correct them. This Chapter will give a thorough overview of all probable causes of these errors. In the following Chapters, we will describe realistic simulations of such errors and correction for such errors to extract the EoR signal.

2.1 Introduction

Recent years have seen a marked increase in the study, both theoretical and observational, of the epoch in the history of our Universe after the cosmological recombination era: the so called "Dark Ages" and the *Epoch of Reionization* (EoR) (Madau et al., 1997a). A cold and dark Universe, after the recombination era, was illuminated by sources of radiation, be it stars, quasars or dark matter annihilation. These first objects ionized and heated their surrounding inter-galactic medium (IGM), carving out "bubbles" in the otherwise neutral hydrogen filled Universe. These bubbles grew rapidly, both in size

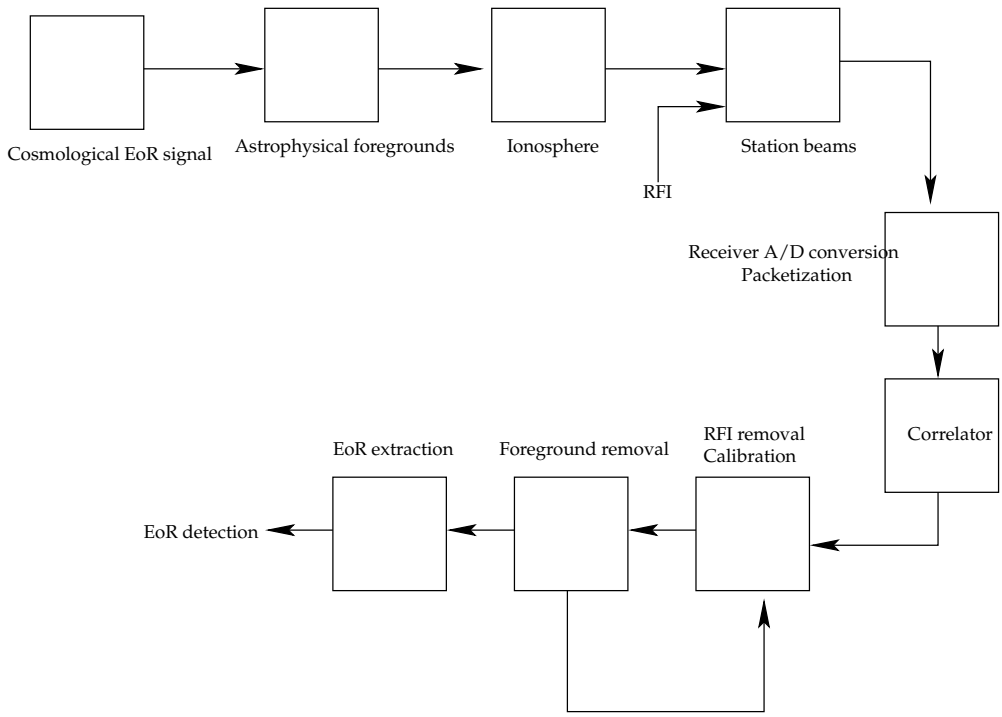

Figure 2.1: The signal path of the LOFAR EoR experiment.

and number, and caused a phase transition in the ionized hydrogen fraction of our Universe at redshifts between 6 and 20 (Sunyaev & Zeldovich, 1975). Although the EoR spanned a relatively small fraction of the Universe's age, its impact on subsequent structure formation (at least baryonic) is crucial. Hence, studying the EoR directly improves our understanding of issues in contemporary astrophysical research such as metal-poor stars, early galaxy formation, quasars and cosmology. For a detailed review of the EoR and the current efforts to detect it, we refer the reader to Furlanetto et al. (2006b) and the references therein.

Despite the astrophysical importance, current astronomical data, for instance the WMAP (Spergel et al., 2007) results and Gunn-Peterson troughs (Gunn & Peterson, 1965), is only able to provide us with crude boundaries, within which, reionization occurred. In order to properly characterize the onset, evolution and completion of the EoR and derive results on its impact on subsequent evolution of structures in the Universe, we need more direct measurements from the EoR.

Observations of the hydrogen 21-cm hyperfine spin-flip transition, using radio interferometry, provide just such a direct probe of the dark ages and the EoR over a wide spatial and redshift range. It is worth mentioning that the spatial range here implies the two dimensions on the sky, which is a function of the baseline lengths of an interferometer, and the third dimension along the redshift direction, which depends on the frequency resolution of the observation. The 21-cm emission line from the EoR is redshifted by a

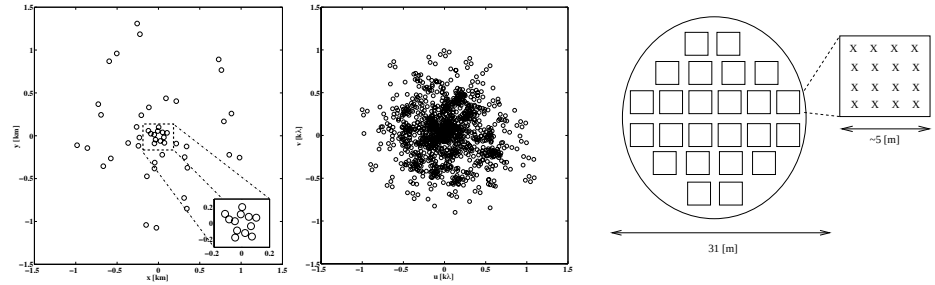

Figure 2.2: Left: The LOFAR compact-core layout. The open circle-size corresponds to the High Band Antenna (HBA) micro-station size of approximately 31 metres in diameter. The 6 inner stations ("six–pack") (12 micro-stations) are shown in the the inset figure. **Middle:** The snapshot uv-coverage of the LOFAR compact core at the zenith and at a frequency 150 MHz. **Right:** The layout of a LOFAR HBA station. Each square represents a tile that consists of four by four, orthogonal bowtie dipoles, as shown in the blown–up inset. Each dipole pair can be rotated individually during the installation of the tile.

factor of $1 + z$, because of the expansion of the Universe, to wavelengths in the few meters regime, which are within the capability of planned low-frequency arrays. Current constraints and simulations converge on reionization happening for a large part in the redshift range $z \simeq 11.38$ ($\sim$115 MHz) to $z \simeq 6$ ($\sim$203 MHz), which is the range probed by LOFAR[1], a radio interferometer currently being built in the Netherlands and other European countries. The 21-cm radiation can not only trace the matter power spectrum in the period after recombination, but can also constrain reionization scenarios (Thomas & Zaroubi, 2008; Barkana & Loeb, 2001).

Developments in radio-wave sensor technologies in recent years have enabled us to conceive of and design extremely large, high sensitivity and high resolution radio interferometers, a development which is essential to conduct a successful 21-cm experiment to characterize and image the EoR. A series of radio telescopes are being built similar to LOFAR, such as MWA[2], PAPER [3], 21CMA[4] and further in the future the SKA[5], all with one of their primary goals being the detection of the redshifted 21-cm signal from the EoR. The GMRT[6] has a programme already under way to detect the EoR or at least to constrain the foregrounds that may hamper the experiment (Pen et al., 2009; Paciga et al., 2010).

Calculations predict the cosmological 21-cm signal from the EoR to be extremely faint. Apart from the intrinsic low strength of the 21-cm signal, the experiment is challenged by a myriad of signal contaminants like man-made and natural (e.g. lightnings) interference, ionospheric distortions, Galactic free-free and synchrotron radiation, clusters and radio galaxies, along the path of the signal. Thus long integration times, exquisite cali-

[1]Low Frequency Array: http://www.lofar.org
[2]Murchison Widefield Array: http://www.mwatelescope.org
[3]Precision Array to Probe Epoch of Reionization: http://astro.berkeley.edu/~dbacker/eor
[4]21-cm Array: http://21cma.bao.ac.cn
[5]Square Kilometer Array: http://www.skatelescope.org
[6]Giant Meterwave Radio Telescope: http://www.gmrt.ncra.tifr.res.in

bration and well-designed RFI mitigation techniques are needed in order to ensure the detection of the underlying 21-cm signal. It is also imperative to properly model all these effects beforehand, in order to develop sophisticated schemes that will be needed to clean the data cubes from these contaminants (Shaver et al., 1999; Zaldarriaga et al., 2004). Due to the low signal-to-noise ratio per resolution element (of the order of 0.2 for LOFAR and even less for e.g. MWA), the initial aim of all current experiments is to obtain a statistical detection of the signal. By statistical, we mean a global change in a property of the signal, for example the variance, as a function of frequency and angular scale. Note that this task involves distinguishing these statistical properties from those of the calibration residuals and the thermal noise.

This Chapter describes the instrumental and propagation model used for the LOFAR EoR signal extraction. The block diagram of the EoR signal path is given in Fig. 2.1. This Chapter focuses on the overall description of the major components in Fig. 2.1. In order to thoroughly understand the interactions between these components and its effect on the EoR signal extraction, exhaustive simulations are essential. In Chapter 3, we focus on realistic simulations of the signal path. Finally, in Chapter 4, we focus on inversion of the data model and extraction of the EoR signal.

This paper is organized as follows: First, we describe the mathematical framework used in the description of the signal model, in particular, the full polarization measurement equation. In section 2.3, after introducing the LOFAR EoR project, we give a general description of the system components of LOFAR that are essential for the EoR experiment. Next, in section 2.4, we describe the propagation effects and corruption by strong foreground signals. Before concluding, we describe the measurement equation relevant to LOFAR in section 2.5.

2.2 Mathematical Framework of the LOFAR Data Model

The most important part of any physical experiment is to find a correspondence between the physical quantities and the measured quantities. In radio interferometry this is achieved through the so called measurement equations (Hamaker et al., 1996; Boonstra, 2005). The measurement equation (ME) describes the relationship between the measured visibilities and the polarized brightness distribution of the sky.

Astronomical radio signals appear as spatially wide-band random noise with superimposed features, such as polarization, emission and absorption lines. The physical quantity that underlies this kind of measurement is the electric field, but for convenience astronomers try to recover the intensity in the direction of interest. If the direction cosines on the unit sphere are given as l, m, the intensity $I(l,m)$ is given as $I(l,m) = \left\langle |\mathbf{E}(l,m)|^2 \right\rangle$, where $\mathbf{E}(l,m)$ is the time harmonic electric field. Correlation of the electric fields between two sensors at $\mathbf{r}_p$ and $\mathbf{r}_q$ position vectors is called the complex visibility, $V_{pq}(u,v,w)$. The distance between the two sensors are given by u, v, w (baseline length).

$$[u,v,w]^T = (\mathbf{r}_p - \mathbf{r}_q)/\lambda \tag{2.1}$$

For a perfect observation without any corruptions or noise, we have

$$V_{pq}(u,v,w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(l,m) \exp\left(-i2\pi\left(ul+vm+w(\sqrt{1-l^2-m^2}-1)\right)\right)\frac{dldm}{\sqrt{1-l^2-m^2}} \tag{2.2}$$

which is called the van Cittert–Zernike theorem (Thompson et al., 2001). The visibilities are sampled for all different sensor pairs p and q. Due to earth rotation, u, w, v will change with time. Moreover, they scale with frequency. This enables us to sample $V_{pq}(u,v,w)$ at many different values of u, v, w. The assumptions under equation 2.2 can be found in Thompson et al. (2001). In particular, we shall point out that the source is assumed to be far away from the receivers, i.e., the electric field is considered to be a plane wave.

2.2.1 The Measurement Equation

In practice, the received signal $V_{pq}(u,v,w)$ is corrupted by instrumental and propagation effects and (2.2) does not hold. In order to study these corruptions, we use the measurement equation as proposed in Hamaker et al. (1996). The major assumption made in converting (2.2) to the formalism given in Hamaker et al. (1996) is that the sources have compact support. This enables to convert the integral in (2.2) to a summation of point-sources.

We can thus consider a plane wave with arbitrary polarization approaching the receivers. This wave can be decomposed into two orthogonal components normal to the direction of propagation, which we call as x and y directions. As described later, each LOFAR receiver has X and Y dipoles sensitive to projections of this this wave. The time harmonic electric field at receiver p can then be written as a 2×1 vector (in polarization) as

$$\mathbf{e}_p = \begin{bmatrix} \mathbf{E}_{xp} \\ \mathbf{E}_{yp} \end{bmatrix} \tag{2.3}$$

Moreover, the coherency matrix (between stations p and q) is defined as

$$\mathbf{C}_{pq} \equiv \left\langle \mathbf{e}_p \mathbf{e}_q^H \right\rangle = \begin{bmatrix} \left\langle \mathbf{E}_{xp}\mathbf{E}_{xq}^* \right\rangle & \left\langle \mathbf{E}_{xp}\mathbf{E}_{yq}^* \right\rangle \\ \left\langle \mathbf{E}_{yp}\mathbf{E}_{xq}^* \right\rangle & \left\langle \mathbf{E}_{yp}\mathbf{E}_{yq}^* \right\rangle \end{bmatrix} \tag{2.4}$$

The elements of the coherency matrix can be related to the Stokes I, Q, U and V parameters (Born & Wolf, 1999) as

$$\mathbf{C}_{pq} = \frac{1}{2}\begin{bmatrix} I+Q & U-iV \\ U+iV & I-Q \end{bmatrix}. \tag{2.5}$$

Let us now consider the propagation of the electric field through a medium that distorts the original wave in power and polarization. We use the Jones formalism to represent the effects on the wave. If we describe the effects of the medium into a Jones matrix

$\mathbf{J}_p$, the received signal at sensor p is $\mathbf{e}'_p = \mathbf{J}_p\mathbf{e}_p$. The modified visibility matrix at baseline pq then becomes

$$\begin{aligned}\mathbf{C}'_{pq} &= \left\langle \mathbf{e}'_p(\mathbf{e}'_q)^H \right\rangle = \left\langle (\mathbf{J}_p\mathbf{e}_p)(\mathbf{J}_q\mathbf{e}_q)^H \right\rangle \\ &= \mathbf{J}_p \left\langle \mathbf{e}_p\mathbf{e}_q^H \right\rangle \mathbf{J}_q^H = \mathbf{J}_p\mathbf{C}_{pq}\mathbf{J}_q^H\end{aligned} \tag{2.6}$$

The important assumption made above is that the Jones matrices are unaffected by the passing electric field. This is taken for granted in radio interferometry. We shall consider possible situations where this might not hold in future work. For multiple receivers, where p and q could change, we see that the received visibility might change depending on the values of $\mathbf{J}_p$ and $\mathbf{J}_q$. This is the reason that we call $\mathbf{C}_{pq}$ the visibility matrix, instead of the coherency matrix. The coherency matrix is the term used to describe the coherency at the source, without any corruptions. The ultimate aim of calibration of LOFAR is to reconstruct this matrix.

We continue to use Jones matrices to describe any form of corruption in the signal path in Fig. 2.1. Note that multiple corruptions can be represented as a product of respective Jones matrices. In certain cases, it is also possible to interchange to order of application of these matrices, when they commute. One such Jones matrix that commutes is the Fourier kernel, given by

$$\mathbf{K} = e^{-i2\pi\left(ul+vm+w\left(\sqrt{1-l^2-m^2}-1\right)\right)} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.7}$$

which corresponds to the Fourier transform in (2.2) between direction (l, m) and the baseline of length (u, v, w).

2.3 Description of the Low-Frequency Array

For a typical interferometric observation, several instrumental components of LOFAR have to function in unison. In this section, we give an overview of all components needed for such an observation. This is the main observing mode for the LOFAR EoR-KSP.

2.3.1 The LOFAR EoR Key Science project

The immediate science goals of the LOFAR EoR KSP, that define some aspects of the design of LOFAR (Bregman, 2002), are: (1) extraction of the 21-cm neutral hydrogen signal averaged along lines of sight, i.e. the global signal (e.g. Shaver et al., 1999; Jelić et al., 2008), (2) determination of the spatial-frequency power spectrum of the brightness temperature fluctuations on angular scales of about 1 arcminute to 1 degree and frequency scales between 0.1 and 10 MHz in the redshift range of $\sim$6-11 and (3) search for Strömgren ionization bubbles around bright sources and (4) the 21-cm absorption-line forest (LOFAR EoR KSP project plan, 2007). In order to achieve these science goals, LOFAR requires a good uv-coverage, a good frequency coverage and a large collecting area, as well as a stable response to the received radiation.

Below, we give a comprehensive description of the aspects of LOFAR which are relevant to the EoR experiment. In particular, we describe possible sources of instrumental

errors, most of which can be ignored for a typical interferometric observation using LOFAR. However, for the EoR experiment, we have to foresee the possible occurrence of even the rarest errors. Whenever feasible, we also introduce relevant Jones matrices that shall be used in the measurement equation at a later stage.

2.3.2 Station configuration and uv-coverage

In its current layout, the LOFAR telescope (de Bruyn et al., 2007; Falcke et al., 2007) will consist of up to 44 stations, at least 20 of which will be located in the core region (Figure 2.2), near the village of Exloo in the Netherlands. The core marks an area of 1.7 by 2.3 kilometres. There are up to 16 remote stations with a maximum distance of about 125 km from the core. Furthermore, according to the present plans, there will be 8 international stations with a maximum distance of about 1300 km from the core. Each High Band Antenna station (HBA station; 110–240 MHz; see next section) in the core is further split into two micro-stations of half the collecting area ($\sim$31 metre diameter), separated by $\sim$130 metres. The two micro-stations of a LOFAR HBA station share the same electronics cabinet. This split further improves the uv-coverage, though at the cost of quadrupling the BlueGene/P correlator demands (i.e. there are four times more base-lines). The central region of the core consists of six closely-packed stations, the six-pack, to ensure improved coverage of the shortest baselines necessary to map out the largest scales on the sky, such as the Milky Way. The station-layout yields a snapshot uv-coverage at the zenith as shown in Figure 2.2. The uv-coverages for a typical synthesis time of 4 hours are shown in Figure 2.3.

Unlike some other instruments aimed at detecting the EoR signal, LOFAR is built to serve different science goals. Hence, optimization of the uv-coverage is done not only for the sole purpose of EoR detection (Bregman & Brentjens, 2008). A good uv-coverage is crucial for several reasons. First to improve sampling of the power spectrum of the EoR signal (Santos et al., 2005; Hobson & Maisinger, 2002; Bowman et al., 2006). Second, to obtain precise Local (Nijboer et al., 2006) and Global (Smirnov & Noordam, 2004) Sky models (LSM/GSM; i.e. catalogues of the brightest, mostly compact, sources in and all-over the sky, i.e. local versus global). A further complication is the extraction of the Galactic and extragalactic foregrounds. This is a vital step in the recovery of the 21-cm EoR signal and requires good sampling of the uv-plane at all frequencies (Bowman et al., 2009, 2006). On the other hand, unlike other EoR arrays, LOFAR will have much longer baselines (running up to hundreds of kilometres). Not only will this provide better resolution to build accurate sky models, but it also enables us to probe ionospheric structure at a broad range of scales (e.g. Koopmans (2010)).

Because LOFAR has no mechanically movable parts, we expect to have no errors in the positions of the stations or position drifts due to wind etc.

2.3.3 Station beams

LOFAR employs two sets of dipoles, the Low Band Antennas (LBA) and the High Band Antennas (HBA). For the EoR experiment we are mostly interested in the HBA dipoles which cover the 110 to 240 MHz frequency range. Each dipole is a crossed dipole which

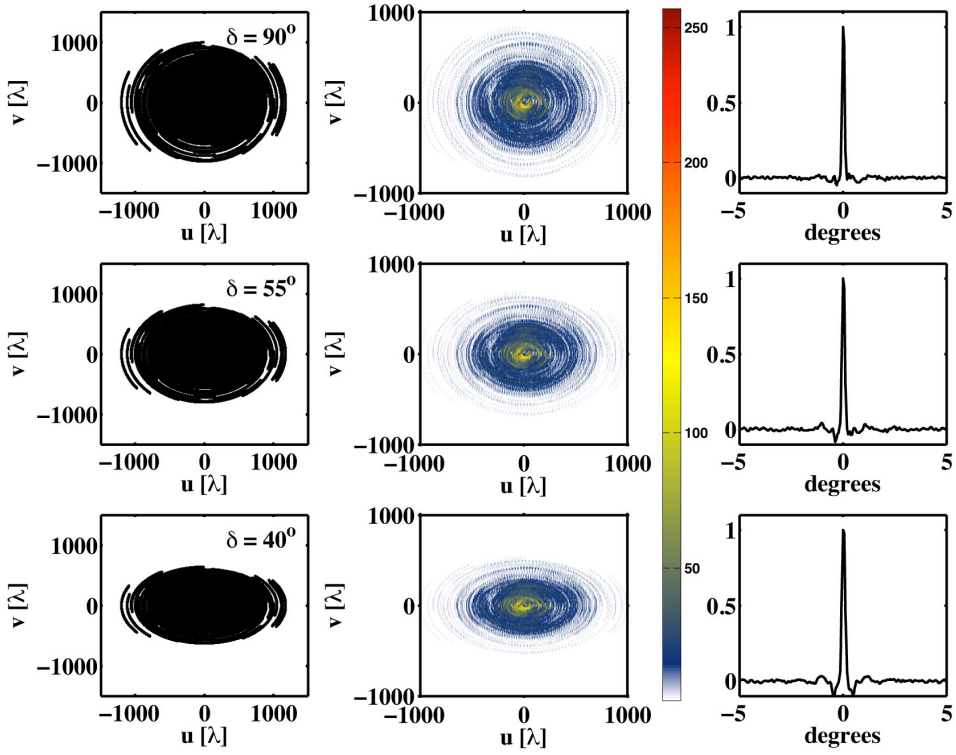

Figure 2.3: The first column of figures shows the uv-coverage of the LOFAR core, that will be used for the EOR experiment, for different values of the declination δ. The size of each point corresponds to the station diameter. shows the corresponding uv point density. The uv plane is gridded with a cell size of 8.5 wavelengths squared. In the last column a horizontal cut of the "dirty beam" is shown.

enables dual polarization observations. The signals from each dipole are coherently combined to form narrower FoV beams (beamforming). For the HBA system, the beamforming is done in two stages. A 4 by 4 rectangular arrangement of HBA dipoles is called a tile (see Fig. 2.2). Within a tile, the 16 signals from dipoles are combined using an analogue beamformer. A full HBA station consists of 24 tiles for each (split) core station and 48 tiles for each for a remote station. The international stations have 96 tiles each. At the second stage of beamforming, signals from each tile are combined using a digital beamformer.

For a given direction (l, m) on the sky, the sensitivity of the beam depends both on the dipole beam pattern and the beamforming gain given to each dipole. The cumulative effect can be represented as the E-Jones matrix. For station p

$$\mathbf{E}(\theta(l,m),\phi(l,m))_p = \begin{bmatrix} \mathbf{E}_{\hat{\theta}p}(\theta,\phi) & \mathbf{E}_{\hat{\phi}p}(\theta,\phi) \\ \mathbf{E}_{\hat{\theta}p}(\theta,\phi+\pi/2) & \mathbf{E}_{\hat{\phi}p}(\theta,\phi+\pi/2) \end{bmatrix} \tag{2.8}$$

where $\theta(l,m)$ and $\phi(l,m)$ are spherical polar coordinates corresponding to elevation and azimuth. The unit vectors on the sphere are given as $\hat{\theta}$ and $\hat{\phi}$, which correspond to the two orthogonal axes along which we have decomposed the electric field of the beam. Note that the beam in the y polarization is taken to be the x beam rotated by $\pi/2$ in azimuth. As the sky rotates, for a given l, m, both θ and ϕ will change. For a detailed calculation of an analytical approximation for (2.8), see Yatawatta (2007). It is also possible to use numerical electromagnetic simulations to obtain these values.

The beam shape or E-Jones matrix in (2.8) is different for each station due to the following reasons. Firstly, the effective sensitivity (and the width) of the beam depends on the number of dipoles (or tiles) used to form the beam. Thus, for a core station, we expect wider beams compared to a remote station (see Fig. 2.4). Secondly, beamforming results in undesirable sidelobes, i.e., increased sensitivity in directions other than the primary beamforming direction. In order to minimize this, each station has a predetermined rotation along the zenith direction. These rotations are almost random such that the product of the sidelobes of two stations that have different rotations cancels out. However, the dipoles in all stations are kept perfectly parallel so that this rotation does not introduce strong polarization. As a result of this rotation, the main lobes of station beams are also rotated. This effect is most significant towards the edge of the beam.

We shall now discuss some subtle errors that could arise from the station beams, regardless of how unlikely they are.

- Currents flowing in a dipole induce currents in the surrounding dipoles. This is termed mutual coupling and it is known to exist but at a very low level. This mainly results in warping of station beamshapes and an increase of noise at the receiver.

- Digital beamformer assume a narrowband system. The narrowband assumption makes it possible to use a phase shift instead of a time delay in beamforming. This is the reason to have an analog beamformer (that has time delay lines) in the first stage of beamforming for the HBA system. For a wideband observation, beamforming is done in sub-bands of 195 kHz bandwidth. The beamforming weights are calculated for the middle frequency of each band. Therefore, there is an error towards the edges of each band in the beamforming process.

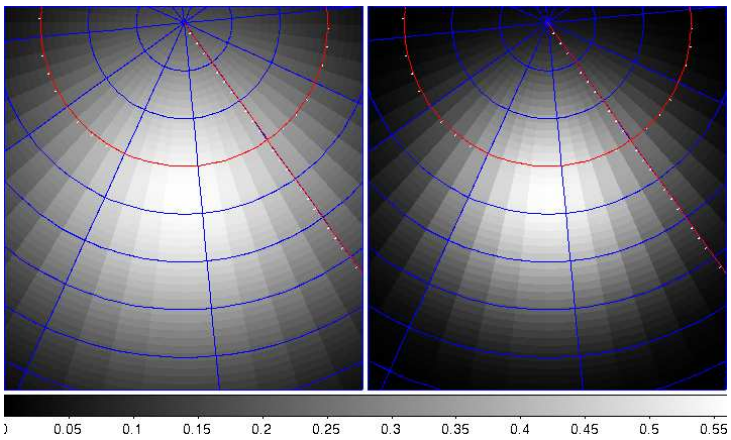

Figure 2.4: Average power beam for a 24 hour observation close to the north celestial pole on the sky at 115 MHz. The beam for a core station is given on left and for a remote station on right.

- To track a fixed direction in the sky, the beamshape and therefore the beamforming weights have to be updated with time. This update is only done at finite intervals of time due to practical reasons. Therefore, we should expect a slightly non-smooth variation of the beamshape with time. In order to minimize this for the HBA system, the analog tile beamformer for a given station is updated at random offsets from a few minute interval .
- As mentioned before, sidelobes of station beams also cause corruptions of the observations. Even after station rotation, we expect a very small leakage from the sidelobes. In the worst case, we will discard data that has too many corruptions due to sidelobes.
- Finally, beamforming assumes dipoles with perfect geometric delays. However, cable length mismatches etc. will violate this assumption. This is corrected by station calibration (Wijnholds & van der Veen, 2009) prior to beamforming. However we expect some residual errors even after station calibration.

2.3.4 Receiver

The receiver of each LOFAR station samples the radio frequency signal from the dipoles, performs digital beamforming (after filtering to get narrow sub bands), encodes this signal using a fixed number of bits, and transmits this to the correlator as a UDP (user datagram protocol) packet via a wide area fibre network. Each packet includes a timestamp so that the correlator performs the correlation in the correct order. This timestamp is obtained by a clock (oscillator) local to each station. Unlike many other radio telescopes, LOFAR does not share a common clock, except for the super-terp stations. The nominal sampling frequency of the clock can be chosen to be 200 MHz or 160 MHz. The sampled data at this frequency is sent through an FPGA (field programmable gate array) for division into different sub-bands. A typical observation is divided into 1024 sub-bands.

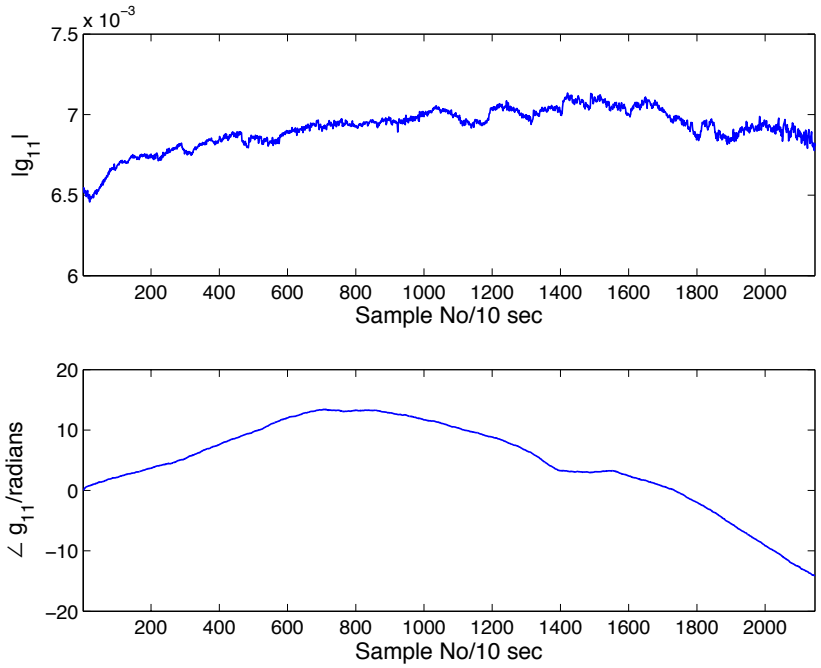

Figure 2.5: The amplitude (top) and phase (bottom) variation of a typical HBA receiver.

However, not all the sub-bands are sent to the correlator. This is to minimize bandwidth requirements and to accommodate multiple beams.

Due to lack of perfect synchronization before an observation and arbitrary drifts during an observation, each station will have a relative error in time. This gives a phase error that varies with time as well as with frequency.

Moreover, the gain (or amplitude) of the signal path starting from the low noise amplifier at a dipole to the A/D converter at the receiver will not be the same for all dipoles. In fact, this gain is also dependent on the impedance of the dipole, which can vary with frequency and external causes such as rain. Hence, we expect the receiver gain to be variable with time and frequency as well.

The cumulative amplitude and phase variation of a receiver p is represented by the G-Jones matrix, given as

$$\mathbf{G}_p = \begin{bmatrix} g_{11,p} & g_{12,p} \\ g_{21,p} & g_{22,p} \end{bmatrix} \tag{2.9}$$

Normally, the off–diagonal terms in (2.9) are assumed to be zero. However, we anticipate the case where they are not zero due to cross coupling between x and y dipoles. In Figure 2.5, we show a typical amplitude and phase variation for a HBA receiver at 130 MHz and in Figure 2.6, we show the frequency response for the same receiver, obtained by calibrating against Cassiopeia A (assumed to have a flat spectrum).

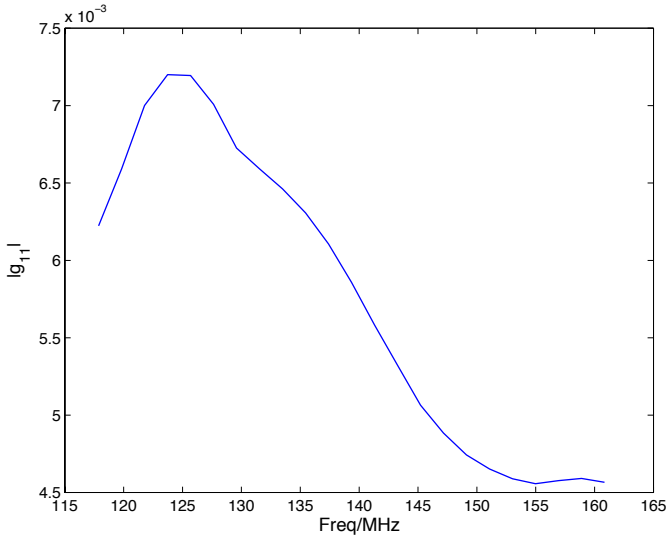

Figure 2.6: The frequency response of a typical HBA receiver, obtained by assuming Cassiopeia A has a flat spectrum.

2.3.5 Correlator

The LOFAR correlator is entirely software-based and operates in real time. For an extensive description, the reader is referred to Romein et al. (2010). We shall highlight some aspects of the correlator that are of interest to EoR detection.

Before correlation, the data from all stations which belong to a single sub-band are divided into narrow channels. For a sub-band of 195 kHz bandwidth, it is possible to obtain 256 channels of 763 Hz bandwidth each. This is done by using a polyphase filterbank channelizer (Harris et al., 2003). The same principle is used at each station for division of the baseband data into multiple sub-bands, using an FPGA. Thus, the whole operation of obtaining a narrowband channel from the baseband data can be seen as a cascade of two channelization operations, one at the station level and the other at the correlator level.

This operation also introduces additional frequency structure to the data. The channels are corrected at the correlator for the frequency behavior introduced by the station FPGA. However, there is still some frequency structure due to the channelization at the correlator. Therefore, we expect residual errors that have to be calibrated.

The correlation is performed by buffering only a few seconds of collected data. If some data packets from some stations are lost, or fail to arrive within this buffering time, that data is assumed to be missing (i.e., flagged). This process is highly random and could act as an additional source of error.

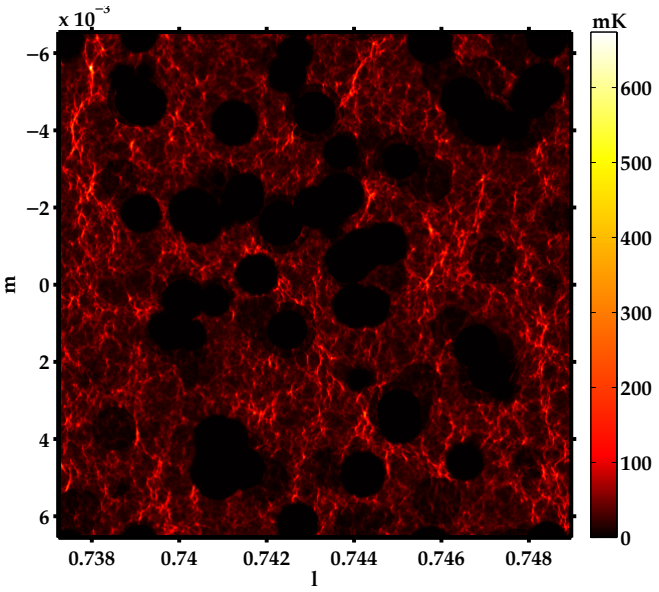

Figure 2.7: Simulated image corresponding to one possible scenario of reionization (from Thomas et al. 2009). The image size is $\sim 2 \times 2$ square degrees.

2.4 Celestial/Terrestrial signals and corruptions

In this section, we describe the signal path, starting from the EoR signal until it is received by a LOFAR dipole (or station).

2.4.1 Cosmological signal and foregrounds

The starting point of our data model is the redshifted 21-cm EoR signal. Until we detect this in our observations, we have to rely on simulated signals. A simulated image from Thomas et al. 2009 is given in Fig. 2.7. The power of this signal is at such a low level that we can ignore its contribution, except at the last stage of signal extraction (see Fig. 2.1).

On the other hand, diffuse foregrounds have significantly more power and considerable effect on our data processing. Of all components of the foregrounds, Galactic diffuse synchrotron emission (GDSE) is by far the most dominant and originates from the interaction between free electrons in the interstellar medium and the Galactic magnetic field. The intensity of the synchrotron emission can be expressed in terms of the brightness temperature T_b and its spectrum is close to a featureless power law $T_b \sim \nu^\beta$, where β is the brightness temperature spectral index. Apart from simulations (see Fig. 2.8, Jelić et al. (2008)), we also have considerable observational information on such foregrounds (Bernardi et al., 2009). For instance, at high Galactic latitudes the minimum brightness temperature of the GDSE is about 20 K at 325 MHz with variations of the order of 2 per cent on scales from 5–30 arcmin across the sky (Rengelink et al., 1997). At the same Galactic latitudes, the temperature spectral index β of the GDSE is about

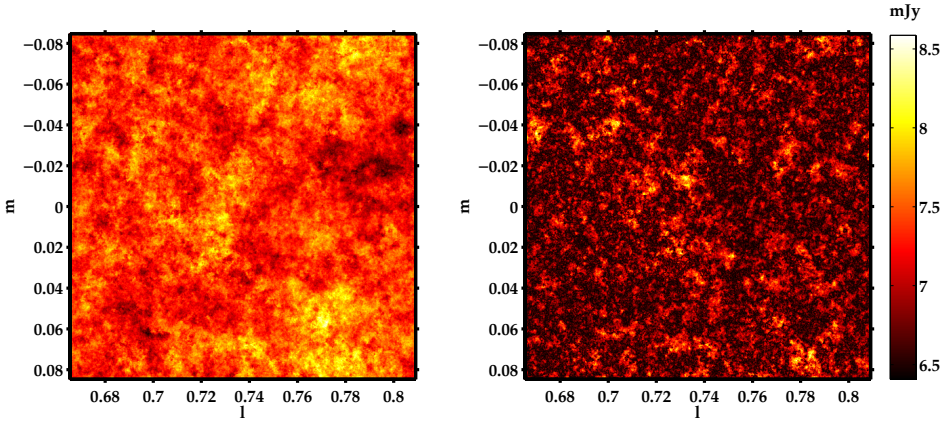

Figure 2.8: Simulated diffuse foreground from Jelić et al. 2010. Left panel shows total intensity while right panel shows polarization. The image size is $\sim 10 \times 10$ square degrees.

-2.55 at 100 MHz (Shaver et al., 1999) and steepens towards higher frequencies, but also gradually changes with position on the sky (e.g. Reich & Reich, 1988).

It should be emphasized that such diffuse structures cannot be incorporated into the measurement equation properly, because they violate the ME assumption of the sky being a collection of discrete point sources, with the most of the space in between them blank. We shall give alternative methods of handling them in Chapter 3. The fact that we are not able to model the effects of the diffuse structure within the measurement equation makes them appear to act as an additional source of noise. In this case ones needs to sample and integrate the data on a refined grid. The simplest method to eliminate this effect is to discard the short baselines during calibration. We also test the accuracy of this shortcut in the aforementioned work.

Apart from GDSE, there are compact radio sources. After repetitive observations of the same part of the sky including the long baselines of LOFAR, we expect to acquire very accurate models for all such sources within the EoR field. It is straightforward to incorporate compact sources (even extended ones) to the measurement equation. The coherency matrix of any such source, seen from a baseline u, v, w (corresponding to stations p and q can be given as

$$\mathbf{C}_{pq}(u,v,w) = \frac{1}{2}\begin{bmatrix} I_{u,v,w} + Q_{u,v,w} & U_{u,v,w} - iV_{u,v,w} \\ U_{u,v,w} + iV_{u,v,w} & I_{u,v,w} - Q_{u,v,w} \end{bmatrix}. \tag{2.10}$$

For a point source, the elements within $\mathbf{C}_{pq}(u,v,w)$ do not vary with u, v, w. However, for extended sources, we can use arbitrary parametrizations (see Yatawatta et al. 2010 for instance).

2.4.2 Ionosphere

Before the celestial signal reaches LOFAR, the first noticeable corruption to this signal is due to the ionosphere. The ionosphere is a birefringent medium, such that one handedness of circular polarization is delayed with respect to the other, introducing a dispersive phase shift. It rotates the linear polarization position angle and is more important at longer wavelengths, at times close to the solar maximum and at Sunrise or Sunset, when the ionosphere is most active and variable (Spoelstra, 1996a, 1995; Spoelstra & Yang, 1990).

For a direction l, m on the sky (and corresponding θ,ϕ), the ionospheric phase-delay introduced can be given as

$$\tau_{l,m} = -a\frac{TEC}{\nu^2} \tag{2.11}$$

where a is a constant, TEC is the total electron content along the line of sight between source at l, m and the receiver measure in TECUs (1 TECU = 10^{16} electrons per m^2), and ν is the frequency of the observation (van Velthoven & Spoelstra, 1992). This is incorporated into the measurement equation using the Z-Jones matrix, for receiver p, as

$$\mathbf{Z}(\theta,\phi)_p = e^{i2\pi\tau\nu}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{2.12}$$

The ionospheric Faraday rotation is defined as

$$\mathbf{F}(\theta,\phi)_p = \begin{bmatrix} \cos(RM/\nu^2) & -\sin(RM/\nu^2) \\ \sin(RM/\nu^2) & \cos(RM/\nu^2) \end{bmatrix} \tag{2.13}$$

where RM is the rotation measure, which is also proportional to TEC (see e.g. Brentjens & de Bruyn (2005)). The rotation measure in SI units is defined as

$$\mathrm{RM} = \frac{e^3}{8\pi^2\epsilon_0 m^2 c^4}\int_0^d n_e B_{||}\,\mathrm{d}s \tag{2.14}$$

where ϵ_0 is the vacuum permittivity, with $B_{||}$ the magnetic field, m the mass of the electron, e the charge of the electron and n_e the electron density. In Fig. 2.9, we have given a typical ionospheric TEC distribution expected and the corresponding delay phase fluctuations.

2.4.3 Radio frequency interference

A substantial amount of work has been done in studying and mitigation of radio frequency interference (RFI) affecting LOFAR, in particular because it is built in a non RFI-free zone. Mitigation techniques have been developed for both pre- and post- correlation data (see Fridman & Baan 2001; Offringa et al. 2010a,b. However, we expect some fraction of RFI to escape undetected by these techniques. This acts as an additional source of noise at later stages of processing and will be considered in the future. For LOFAR it affects less than five per cent of the data (Offringa et al., 2010a).

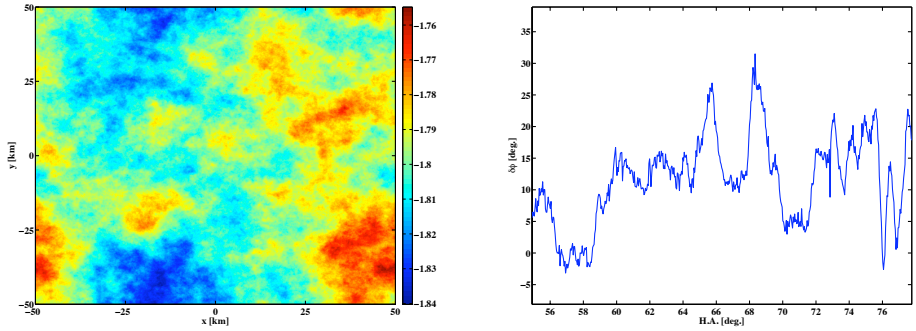

Figure 2.9: Left: A simulated map of the TEC of an simulation of the ionosphere at an arbitrary time, above the LOFAR core. Each LOFAR station sees approximately 50 kilometers squared of the ionosphere at an altitude of 250 km. The values are displayed in TECUs. **Right:** Example of phase fluctuations caused by the ionosphere at a baseline of 100 metres as a function of Hour Angle

2.4.4 Noise

In its broadest sense, noise in our data model corresponds to everything that is not included in the model. The major contributions to this noise are the power from the sky and the receiver. The theoretical root mean square noise level in terms of the real and imaginary part of the complex visibilities for an interferometer pair between stations p and q is given by :

$$\Delta V \left\{\mathrm{Re}, \mathrm{Im}\right\}_{p,q} = \frac{1}{\eta_s} \times \sqrt{\frac{\mathrm{SEFD}_p \times \mathrm{SEFD}_q}{2 \times \Delta\nu \times \tau_{\mathrm{avg}}}}$$

where η_s is the system efficiency that accounts for electronic digital losses, $\Delta\nu$ is the frequency bandwidth and τ_{avg} is the averaging time during which each station accumulates data. The System Equivalent Flux Density (SEFD) can be written as $\mathrm{SEFD} = T_{\mathrm{sys}}/K$, where $K = (\eta_{\mathrm{a}} \times A_{\mathrm{eff}})/(2 \times k_{\mathrm{B}})$ and depends on the station efficiency η_{a} and the effective collecting area A_{eff} of the station, with k_{B} being the Boltzmann constant. A typical value of K is 0.5 for a LOFAR HBA station.

There are widely used assumptions in radio astronomy about the properties of noise. First, given two stations p and q, we only perform one correlation (2.6). Thus, if the noise in both polarizations at receiver p is $\mathbf{n}_p$, it is assumed that $\left\langle \mathbf{n}_p \mathbf{n}_q^H \right\rangle = \left\langle \mathbf{n}_q \mathbf{n}_p^H \right\rangle^H$. This assumption is used in imaging to get a real valued image using complex visibilities. Secondly, the noise is assumed to be white and Gaussian. In self-calibration (Pearson & Readhead, 1984), the least squares estimate of calibration parameters is also the maximum likelihood estimate because of this property. In the measurement equation we represent the noise contribution to baseline pq as $\mathbf{N}_{pq}$, which is a 2 by 2 matrix.

2.5 LOFAR measurement equation

In this section, we assimilate all Jones matrices described previously to construct the measurement equation relevant to LOFAR. To recap, we described effects that can be split into three categories: image-plane effects, uv-plane effects and noise. Image plane effects are named as such because of their dependence on direction on the sky. On the other hand, uv-plane effects are direction independent, telescope dependent effects.

- Image–plane effects
  - Faraday rotation **F**
  - Ionospheric phase fluctuations **Z**
  - Antenna voltage pattern **E**
  - Polarization leakage and instrumental polarization **D**
  - Fourier kernel **K**
- uv–plane effects
  - Complex gains **G**
  - Frequency bandpasses **B**

Hence, finally, we get the visibility observed at baseline pq as

$$\mathbf{V}_{pq} = \mathbf{G}_p \left(\sum_k \mathbf{K}_p \mathbf{E}_p \mathbf{Z}_p \mathbf{F}_p \mathbf{C}_{kpq} \mathbf{F}_q^H \mathbf{Z}_q^H \mathbf{E}_q^H \mathbf{K}_q^H \right) \mathbf{G}_q^H + \mathbf{N}_{pq} \tag{2.15}$$

Note that the coherency matrix $\mathbf{C}_{kpq}$ corresponds to sources (k) that are compact (i.e., point sources and extended sources with finite support). As mentioned in section 4.1, diffuse structure with infinite support cannot be incorporated into the above measurement equation. We shall describe the addition of diffused foregrounds to the visibilities in Chapter 3 of this thesis. We also note that effects that are baseline dependent (closure errors) have not been included in the above formulation.

2.6 Summary

The main goal of this Chapter is to introduce a physics-based data model for the LOFAR EoR Key Science Project. In Chapter, we provided a brief description of the physical connections between the Hamaker–Bregman–Sault formalism Hamaker et al. (1996) and instrumental/propagation parameters. We shall use this formalism for extensive simulations of typical LOFAR EoR observations (Chapter 3) and, furthermore, calibration and signal extraction (Chapter 4).

Acknowledgments

We would also like to thank Johan Hamaker, Jan Noordam, Oleg Smirnov and Stephan Wijnholds for their useful suggestions and discussions during the various stages of this work. LOFAR is being funded by the European Union, European Regional Development Fund, and by “Samenwerkingsverband Noord-Nederland”, EZ/KOMPAS.