EoR imaging with the SKA: the challenge of foreground removal

Anna Bonaldi

SKA Organization, Jodrell Bank, Lower Withington, Macclesfield,
Cheshire, SK11 9DL, United Kingdom
email: `a.bonaldi@skatelescope.org`

Abstract. 21-cm observations of the Cosmic dawn (CD) and Epoch of Reionization (EoR) are one of the high priority science objectives for SKA Low. One of the most difficult aspects of the 21-cm measurement is the presence of foreground emission, due to our Galaxy and extragalactic sources, which is about four orders of magnitude brighter than the cosmological signal. While end-to-end simulations are being produced to investigate in details the foreground subtraction strategy, it is useful to complement this thorough but time-consuming approach with simpler, quicker ways to evaluate performance and identify possible critical steps. In this work, I present a forecast method, based on Bonaldi *et al.* (2015), Bonaldi & Ricciardi (2011), to understand the level of residual contamination after a component separation step, and its impact on our ability to investigate CD and EoR.

Keywords. instrumentation: interferometers, methods: data analysis, stars: formation, cosmology: large-scale structure of universe

1. Introduction

The Cosmic Dawn (CD) and the Epoch of Reionization (EoR) are two of the last unobserved eras of our Universe. The CD signal ($z \sim 30$–15) probes the formation of the very first stars, whose heat and radiation affected the coupling between the spin temperature of neutral hydrogen and the temperature of neutral gas. During EoR ($z \sim 15$–6), with star and structure formation progressing, and more energetic radiation (UV, X rays) being emitted, the Universe goes from fully neutral to fully ionised. The study of CD and EoR thus provides invaluable information on the early stages of structure formation.

The next generation instruments, like the Square Kilometre Array (SKA, †) will be able to accurately map this radiation over a wide redshifts range. The cosmological signal is however dwarfed by the presence of foreground contamination, both diffuse and point-like, coming from our Galaxy as well as extragalactic objects. The removal of such contaminants is referred to as "foreground removal" or "component separation", and often relies on the spectral smoothness of the foreground components, as opposed to the CD/EoR signal, which is essentially uncorrelated with frequency.

The removal of discrete point-sources and that of diffuse foregrounds typically requires different approaches, and therefore is achieved in sequence. In this work, we address the second step, the removal of diffuse foregrounds from our Galaxy, which are synchrotron and free-free emission.

Synchrotron radiation is due to cosmic-ray electrons spiralling in the Galactic magnetic field. Its frequency behaviour reflects the spectrum of the electrons and can be described

† `https://www.skatelescope.org`

to first order by a power-law model with spectral index β_{s}. At frequencies below a few GHz β_{s} ranges from -2.5 to -2.7 as a function of the position in the sky.

Free-free radiation is due to Brehmstrahlung emission. Its spectrum is quite uniform and can be predicted quite accurately; in the optically-thin regime, which holds at high Galactic latitudes, it is well approximated between 100 and 200 MHz by a power-law with spectral index $\beta_{\mathrm{ff}} = -2.08$.

We construct a simple forecasting approach by modelling foreground subtraction within the framework of the linear mixture data model (§2). We then use this model to assess the accuracy required for foreground removal to study CD and EoR with the SKA (§3).

2. Foreground residual forecaster

2.1. *The linear Mixture data model*

For component separation purposes, it is convenient to model the data as a linear mixture of the components. For each direction in the sky we write

$$\mathbf{x} = \mathbf{Hs} + \mathbf{n}. \tag{2.1}$$

The vectors $\mathbf{x}$ and $\mathbf{n}$ have dimension equal to the number of frequency channels, N_{c}, and contain the data and instrumental noise, respectively; $\mathbf{s}$ is the vector containing the astrophysical sources and has dimension equal to the number of components, N_{c}; $\mathbf{H}$ is the $N_{\mathrm{d}} \times N_{\mathrm{c}}$ mixing matrix, containing the frequency scaling of the components.

In order to be able to write the data model as in Eq. (2.1) we made some simplifying assumptions. The most important of these is that the instrumental resolution does not depend on frequency. This is in general not true, and requires a pre-processing step in which the resolution is equalised by suitably smoothing the data. In our case, because we focus on large scales, such a loss of resolution is not particularly problematic. Alternatively, this assumption is not required when applying the linear mixture data model of Eq. (2.1) directly to the uv plane. In fact, in this domain the convolution for the instrumental beam becomes a multiplication, and a frequency-dependent beam can be accounted for without breaking the simple linear combination model.

The mixing matrix $\mathbf{H}$ is in general different for different pixels, due to changes in the properties of the inter-Galactic medium responsible for the emission. It is in general reasonable to assume that these properties vary smoothly and not significantly over limited regions of the sky. This means that the components have very similar spatial distribution at all frequencies and vary mainly in intensity.

This assumption holds true for the Galactic synchrotron and free-free emission; however it cannot be applied to the 21 cm signal. In this case frequency corresponds to redshift, therefore we observe a significant change in the morphology of the component, due to the combined effect of position along the line of sight and evolution. When applying Eq. (2.1) to our case, we interpret the source vector $\mathbf{s}$ as containing the foreground components only, and we model the noise vector as $\mathbf{n} = \mathbf{n}_{\mathrm{inst}} + \mathbf{n}_{\mathrm{HI}}$.

2.2. *Foreground reconstruction and subtraction*

Starting from the linear mixture model of Eq. (2.1), we can obtain an estimate $\hat{\mathbf{s}}$ of the components $\mathbf{s}$ through a suitable linear mixture of the data:

$$\hat{\mathbf{s}} = \mathbf{Wx}, \tag{2.2}$$

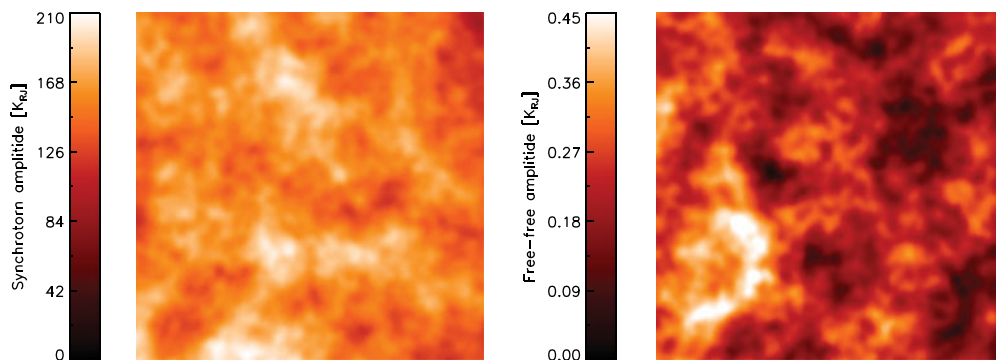where the matrix $\mathbf{W}$ is called reconstruction matrix. In this work we use a reconstruction matrix given by

Figure 1. Foreground realization at 150 MHz in K_{RJ}. The sky area is $3.25° \times 3.25°$ wide. From Bonaldi & Brown (2015).

$$\mathbf{W} = [\hat{\mathbf{H}}^T \mathbf{C}_n^{-1} \hat{\mathbf{H}}]^{-1} \hat{\mathbf{H}}^T \mathbf{C}_n^{-1}, \qquad (2.3)$$

where $\mathbf{C}_n$ is the noise covariance matrix and $\hat{\mathbf{H}}$ is an estimate of the mixing matrix. This linear combination is called the generalised least square solution (GLS). It is easy to show that, if $\hat{\mathbf{H}}$ is the true mixing matrix $\mathbf{H}$, Eq. (2.3) is an unbiased solution for the components.

We finally clean the frequency maps by subtracting the reconstructed foreground components scaled by means of the estimated mixing matrix

$$\mathbf{n}_{HI} + \mathbf{n}_{inst} = \mathbf{x} - \hat{\mathbf{H}}\hat{\mathbf{s}}. \qquad (2.4)$$

2.3. *Forecast of foreground residuals*

As soon as the estimated mixing matrix, $\hat{\mathbf{H}}$, is not the true mixing matrix, $\mathbf{H}$, the foreground subtraction is not perfect and leaves residuals in the frequency maps. The map of residuals, $\mathbf{s} - \hat{\mathbf{s}}$, for a linear mixture source reconstruction can be estimated as:

$$\mathbf{s} - \hat{\mathbf{s}} = (\mathbf{WH} - \mathbf{I})\tilde{\mathbf{s}}, \qquad (2.5)$$

where $\mathbf{I}$ is the identity matrix and $\tilde{\mathbf{s}}$ is a set of simulated components (Bonaldi & Ricciardi 2011). We used the simulated sky presented in Bonaldi & Brown (2015) to model the diffuse synchrotron and free-free emission. The templates are shown in Figure 1. A foreground realization at 150 MHz is shown in Fig. 3. The total foreground contamination on the CD/EoR signal is finally given by the sum of the synchrotron and free-free residuals at all frequencies.

3. Results

We ran the forecast method described above to predict foreground residual levels given a different uncertainty on the mixing matrix. In our case, this error is dominated by that on the synchrotron spectral index, because the free-free spectral index is known to much higher accuracy. We modelled the true spectral indices as $\beta_s = -2.6$ and $\beta_{ff} = -2.08$; we assumed errors ranging from 0.1% to 5% on the former and no error on the latter. For each error level, we generated 10 "estimated" spectral indices $\hat{\beta}_s$ drawn from a Gaussian distribution centred on the true value and with an RMS given by the error level. We averaged foreground residuals over those 10 realisations.
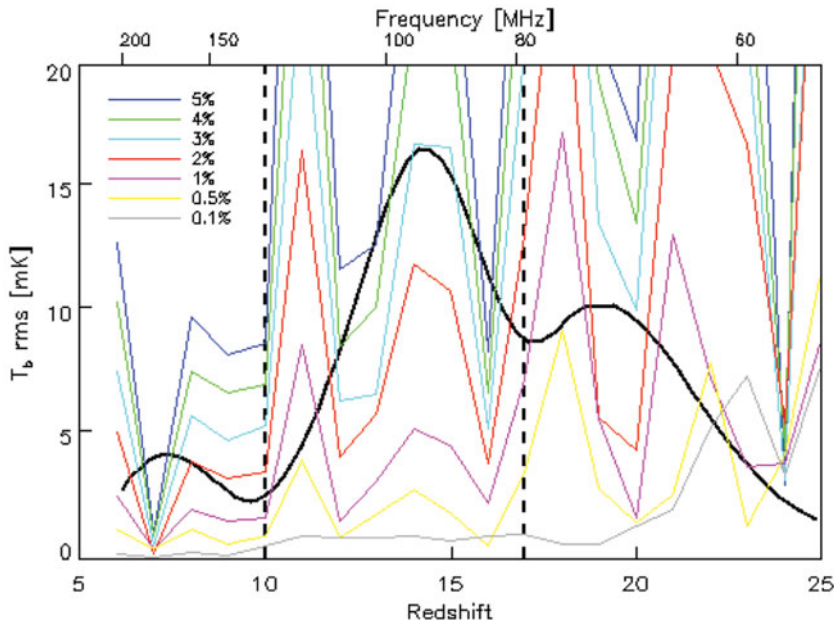
Figure 2. RMS on the scale of $0.204\,\mathrm{Mpc}^{-1}$ of the CD/EoR signal as a function of redshift and frequency (black solid line), from a theoretical model from J. Pritchard. We identified three redshift ranges, separated by the vertical dashed lines, corresponding to local minimal of the signal at $z = 10$ and $z = 17$. These ranges have been labelled *EoR*, *low-z CD* and *high-z CD* in the analysis. Coloured lines: foreground residuals forecasted for different error levels on the synchrotron spectral index, as detailed in the caption.

The foreground residuals represent the contamination left after component separation, and should be compared to the CD/EoR signal. In general, they will have a different impact at different frequencies and angular scales. We considered the RMS of the fluctuations at the scale of $k = 0.204\,\mathrm{Mpc}^{-1}$, which is the typical one for CD/EoR. For the latter, we used theoretical model (courtesy of J. Pritchard) shown as the black solid line in Figure 2. We split the analysis into the $z = 5$–10, 10–17 and 17–25 redshift ranges, in order to isolate the three main features in the theoretical model, which we labelled *EoR*, *low z CD* and and *high z CD*, respectively. We therefore processed the foregrounds separately in the corresponding frequency ranges, 54–79 MHz, 79–129 MHz and 129–236 MHz. Specifically, we used each range to compute the residuals, depending on the true mixing matrix and the estimated mixing matrix, with eqns. (2.5) and (2.3), and added the synchrotron and free-free contribution at all frequencies.

The results for the different error levels are shown as the coloured lines in Figure 2. The contamination oscillates widely with frequency. This is because the systematic errors on the two components add together in different ways at different frequencies, with the fluctuations either adding up or subtracting partially. If we disregard the fluctuations and look at the overall trends, the contamination gets worse at low frequency, where the foregrounds are intrinsically stronger. Both the contamination and the oscillations are reduced by reducing the error on the synchrotron spectral index.

In order to determine whether the CD/EoR signal would be detected in the various cases, we computed a "signal-to-noise" (S/N) figure of merit by simply averaging the ratio between the cosmological signal RMS and the foreground residual one for each of the three frequency ranges considered. This is shown in Figure 3.
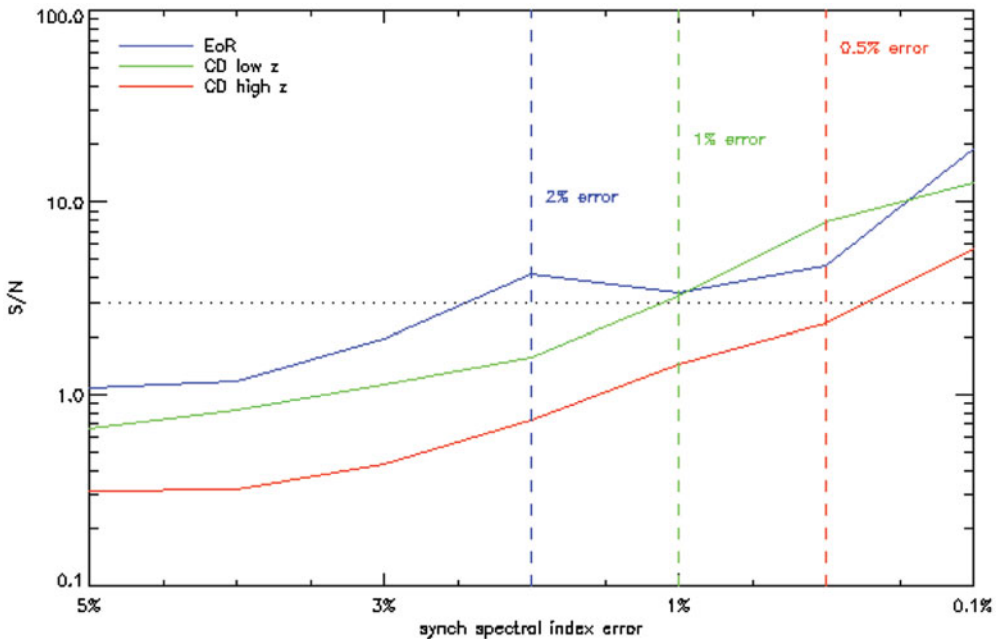
Figure 3. Signa-to-noise of CD/EoR detection over the residual foregrounds, as a function of the error on the synchrotron spectral index, for the three redshift ranges considered. The horizontal dotted line represent a detection threshold of S/N = 3, which would require a precision of 2, 1 and 0.5 & on the synchrotron spectral index for EoR, low-z CD and high-z CD, respectively.

If we set the threshold for a satisfactory detection to a S/N of 3 (dotted line in Figure 3), we see that this would require a precision of 2%, 1% and 0.5% on β_s respectively for EoR, CD at low redshift and CD at high redshift.

These results stress the point of the critical importance of accurate foreground removal for CD/EoR measurement. They are not general, since we considered just one CD/EoR model, one foreground realization and one foreground-removal approach (the GLS solution). However, the forecast is easy and quick to run and can be used to explore in more details all these parameters and provide useful indications, before a full end-to-end simulation is performed.

References

Bonaldi A. & Brown M. L., 2015, *MNRAS*, 447, 1973
Bonaldi A., Ricciardi S., & Brown M. L., 2014, *MNRAS*, 444, 1034
Bonaldi A. & Ricciardi S., 2011, *MNRAS*, 414, 615