15

Mapping the universe with 21 cm observations

Xuelei Chen
National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China

Hydrogen is the most abundant element in the universe, and its 21 cm line is a very important observational probe for radio astronomy. The 21 cm cosmology is at the forefront of current research. I first briefly introduce the physics of the 21 line emission and absorption, and its observational signature during the cosmic evolution. I then describe some ongoing 21 cm tomography experimental efforts, and illustrate the typical experimental setup with the Tianlai pathfinder array; the global spectrum experiment is also briefly introduced. The basics of synthesis imaging are briefly introduced and the challenge posed by foreground presented, with an example of foreground subtraction using the PCA method. The foreground wedge in the 21 cm power spectrum is explained. Finally, I conclude with an outlook to future development.

1 The neutral hydrogen and 21 cm line

The ground state of the neutral hydrogen atom is split into two states by the interaction between the magnetic moments of the proton and electron. Transitions between these hyperfine split states produce or absorb radio waves with a frequency of 1420 MHz corresponding to a wavelength of 21 cm; this spectral line is usually called the 21 cm line. Historically, this was the first spectral line predicted and detected in radio astronomy, and it has been widely used in the observation of neutral hydrogen (often referred to as HI) in the galactic interstellar medium as well as nearby galaxies.

Big Data in Astronomy. https://doi.org/10.1016/B978-0-12-819084-5.00009-2

With the expansion of the universe, a radio wave emitted in an earlier epoch is redshifted to a longer wavelength, but customarily it is still referred to as the (redshifted) 21 cm radiation for its origin. As the space is filled with photons from the cosmic microwave background (CMB) or more generally some continuum radio background, when we speak of observing the 21 cm radiation, we are actually speaking of observing the excess (emission) or dearth (absorption) features produced by the neutral hydrogen on top of such background radiation.

The neutral hydrogen atoms first appeared right after the epoch of recombination, which marked the end of the Big Bang. As the baryonic matter in the early universe was almost entirely made up of hydrogen and helium atoms with very little heavy elements, the 21 cm line provides perhaps the only means to probe the ensuing dark age. As structures formed, the ionizing photons from stars and accreting black holes began to ionize the gas of the universe, and through this epoch of reionization (EoR), the 21 cm line can be used to map neutral hydrogen distribution, thus revealing the process of reionization. After the EoR, the neutral hydrogen atoms are found primarily within galaxies where the gas density is sufficiently high, such that the recombination rate (proportional to density squared) can compensate the ionization rate from the extragalactic ultraviolet background.

Although the redshifted 21 cm radiation is in principle observable up to the Dark Ages, at present it is mostly detected at low redshifts ($z<0.2$) because its signal is much smaller than some foreground radiations, such as the synchrotron emission of cosmic ray electrons moving in the galactic magnetic field, the Bremsstrahlung radiation from electrons in the interstellar medium, and the radio galaxies and quasars. The galactic synchrotron radiation is particularly severe, as its strength is about 10^5 that of the EoR 21 cm signal. At least in principle, the foreground can be identified and removed as their spectra are smooth, whereas the 21 cm spectrum varies as it follows the underlying HI density, which on large scales traces the matter density and fluctuates accordingly. The potential of 21 cm observation is great as it can be used to probe a large portion of the comoving volume of the observable universe. Once a precision observation technique is mastered and the signal extracted, this will supply an unprecedented amount of precious information for cosmological studies.

1.1 The physics of spin temperature

The 21 cm line is produced by the transition between the singlet ($F=0$) and triplet ($F=1$) hyperfine levels of the hydrogen atom at the electronic ground state (1s) (Fig. 15.1). These two

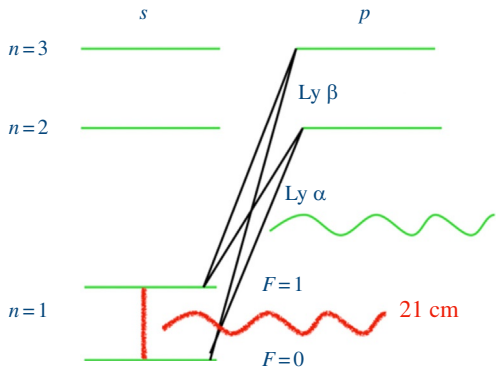

Fig. 15.1 The neutral hydrogen energy levels.

states are distinguished by the different spin of the electron with respect to the proton. A spin temperature can be defined according to the occupation numbers in the two states,

$$\frac{n_1}{n_0} = 3e^{-h\nu/kT_s}$$

where h is the Planck constant, ν the frequency, so hν gives the energy of the 21 cm photon, k is the Boltzmann constant, and T_s the spin temperature. If the spin temperature is much higher than $h\nu$, then the difference in energy level is negligible and the two states are nearly equally occupied except for the degeneracy factor of 3. But if the spin temperature is low, then the lower state is much more occupied than the higher one.

In a cosmological setting, the line can be observed in absorption or emission against the cosmic microwave background (CMB) with the brightness temperature given by [1].

$$\begin{aligned}\delta T &= \frac{T_s - T_r}{1+z}(1-e^{-\tau}) \\ &\simeq (0.025K)\left(\frac{\Omega_b h_0}{0.03}\right)\left(\frac{0.3}{\Omega_{m0}}\right)^{1/2}\left(\frac{1+z}{10}\right)^{1/2}\frac{\rho_{HI}}{\overline{\rho}_H}\frac{T_s - T_r}{T_s}.\end{aligned}$$

where T_r is the radio background temperature at the frequency 1420 MHz at that particular redshift. In the absence of other strong backgrounds, it would be the CMB temperature, and τ is the optical depth. The second line is the approximation at the epoch of reionization. ρ_{HI} is the density of neutral hydrogen and $\overline{\rho}_H$ is the mean hydrogen density of the universe (we have assumed the optically thin case and approximated $\Omega_{m0}(1+z)^3 \gg 1$.

The hydrogen spin temperature plays a key role in determining the sign and amplitude of the 21 cm emission or absorption signal against the radio background. To study the evolution of the spin

temperature, we may regard the two hyperfine split energy levels of the ground state hydrogen atoms as a thermal system in contact with other thermal systems (Fig. 15.2). The systems in contact include the radio background photons (below we shall take this as the CMB, the radio emission from galaxies and stars may also contribute some part but it is just a small fraction in the conventional scenario), the kinetic motion of the atoms, and the Lyman series photons. According to the zeroth law of thermodynamics, two contacting thermal systems will approach a common equilibrium temperature. As the spin system has a very small heat capacity, it has little impact on either of the other two systems, and if it is in contact with only one of them, then the spin temperature would approach the temperature of that system. As it is in contact with several systems, the spin temperature is a weighted average of the temperature of these, with the weights given by the corresponding coupling strength.

The coupling between the radio background and the spin system depends only on the transition oscillator strength and the flux of the radio background, and in the case of a CMB-dominated background, it is entirely predictable. The kinetic motion is directly coupled to the spin by atomic collisions. In an atomic collision, both the velocity and the spin can be changed simultaneously. As the collision rate is proportional to the density squared, it is ineffective in thin gas.

The atom on the ground state 1s can jump to a higher excited state by absorbing a photon, ending up on the $2p$, $3p$, … states for the Lyman alpha (Ly α), Lyman beta (Ly β), … Lyman series photon. The most important among these are the Ly α photons. In this case, after a short time the atom will fall back to the ground state

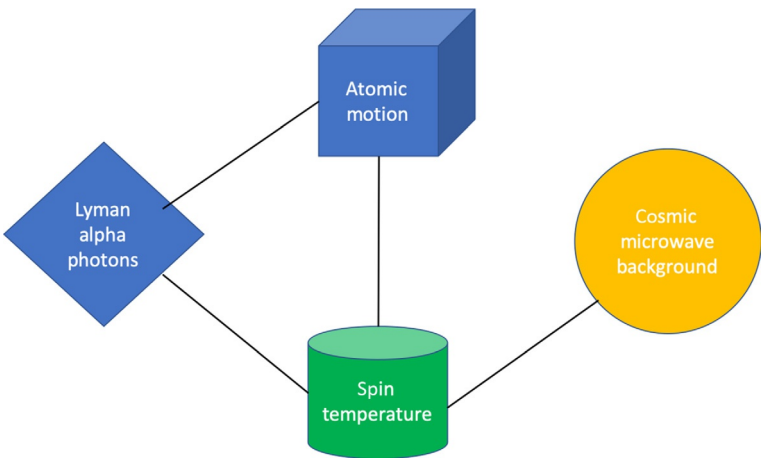

Fig. 15.2 The atomic motion, Lyman alpha photons, CMB, and spin as interacting thermal systems.

with the spontaneous emission of a Ly α photon. We may also regard this as the resonant scattering of the Ly α photon. The higher Lyman series are less important because they do not always fall back to the ground state, but can instead cascade down the energy levels. Due to the large scattering cross-section, each Ly α photon will be resonantly scattered repeatedly. This whole process can also be seen as the atom spin system exchanging energy with the Ly α photons, so that the spin temperature should approach the color temperature of the Ly α photons. The color temperature is defined by the slope of the radiation spectrum at the line center. Normalized to the predictable radio background temperature, the spin temperature is then given by

$$T_S = \frac{T_r + y_c T_k + y_\alpha T_\alpha}{1 + y_c + y_\alpha}$$

where T_r, T_k, T_α are the radio background temperature, kinetic temperature of the gas, and color temperature of the Ly α photons, respectively, and the collisional and Wouthuysen-Field coupling weights are given by

$$y_\alpha \equiv \frac{P_{10} T_*}{A_{10} T_k}; y_c \equiv \frac{C_{10} T_*}{A_{10} T_k}$$

where $T_* = 0.0682\,\mathrm{K}$ is the hyperfine energy splitting, $A_{10} = 2.87 \times 10^{-15} s^{-1}$ is the spontaneous emission coefficient of the 21 cm line, $C_{10} = \kappa_{10} n_H$ is the collisional deexcitation rate, κ_{10} ranges from $2 \times 10^{-14} \sim 2.5 \times 10^{-10} \mathrm{cm^3 s^{-1}}$ and has been tabulated.

The indirect deexcitation rate P_{10} of the hyperfine structure levels is related to the total Ly α scattering rate $P_{10} = 4P_\alpha/27$, with

$$P_\alpha = \int d\nu c n_\nu \sigma(\nu)$$

where n_ν is the number density of photons per unit frequency, and $\sigma(\nu)$ is the cross-section for Ly α scattering. The interaction between the Ly α photons and the hydrogen gas, on the other hand, tends to make the color temperature to that of the gas kinetic temperature. Thus, to a good approximation,

$$T_\alpha = T_k.$$

1.2 The evolution of 21 cm signal through cosmic history

We now consider the evolution of the spin temperature and 21 cm brightness temperature throughout the history of the universe. The neutral hydrogen first appeared after recombination. Initially,

its temperature would be nearly equal to that of the CMB. In adiabatic expansion, the nonrelativistic gas temperature drops as $T \propto a^{-2}$ while the CMB photons drops as $T \propto a^{-1}$ where a is the scale factor, so the gas temperature would soon fall below that of the CMB if there is no heat change between the photons and the gas. In fact, there is indeed some heat exchange—there are still residue-free electrons that can scatter with the CMB photons, during which energy is transferred from the photons to gas. However, as the universe expands, this heat transfer rate falls quickly, such that during the late Dark Ages ($z < 200$), the gas temperature fell well below that of the CMB. The gas density is sufficiently high, so the collisional coupling is important, and the gas spin temperature approaches that of the kinetic temperature. As a result, we will see a broad negative 21 cm brightness temperature or absorption feature in the average spectrum corresponding to this redshift. On the other hand, the primordial fluctuations began to grow during the dark age, forming nonlinear structures such as halos and filaments. In such systems, the gas is compressed and shock heated and has a significantly higher temperature, though as such the structures are still very rare and occupy a very small fraction. However, as the density drops further, the collisional coupling becomes less effective, the spin temperature tends to that of the CMB temperature, and the 21 cm brightness temperature approaches 0 (Fig. 15.3).

Then, as sufficiently large halos formed, the gas within can collapse to form stars. These first stars supply Ly α photons in two ways: the continuum radiation can be redshifted to the Ly α wavelength, and ionization and deexcitation in the gas

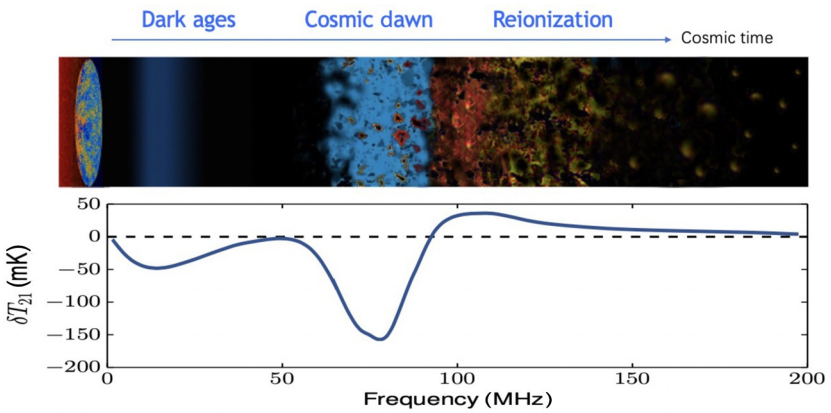

Fig. 15.3 The cosmic evolution history and the 21 cm brightness temperature.

surrounding the star can produce the Ly α photons. This first form a Ly α halo, in which the spin temperature couples to the kinetic temperature by the Wouthuysen-Field effect of the Ly α photons. Then, a Ly α background is built up in the whole space as more emitters are formed. If the majority of the gas is not heated by x-ray photons, its kinetic temperature will remain below the CMB temperature, so that a 21 cm absorption trough will be produced again [3,4].

$T_S \to T_K$

Eventually, the gas would be heated by the ionizing photons, so some sort of 21 cm emission would be seen. However, as more and more stars, galaxies, and accreting blackholes (quasars) are formed, the gas would all be ionized, except that within the galaxies or dense clumps inside the dark matter halos where the recombination rate is higher, which can keep a self-screened neutral region in the presence of the ionizing background. The 21 cm signal from the intergalactic gas would again drop to zero, but of course there are still the 21 cm signals from the neutral hydrogen in the galaxies [2].

2 The 21 cm experiments

There are generally three types of 21 cm experiments: 21 cm tomography, 21 cm forest observation, and 21 cm global spectrum measurement.

In the 21 cm tomography experiment, one surveys the 21 cm variation in both the spatial direction and frequency, thus yielding a three-dimensional data cube. As different frequency corresponds to different redshift, this gives a three-dimensional map of the universe along the light cone, revealing spatial distribution as well as time evolution. This is the primary mode of observation. It can either be carried out in the traditional galaxy survey mode, where one tries to observe the distribution of individual objects (galaxies), or it can be carried out in the intensity mapping mode, where the focus is on obtaining the large-scale distribution.

The 21 cm forest is analogous to the well-known Lyman alpha forest in the quasar absorption line. One tries to observe the spectra of high redshift bright continuum spectra radio sources, such as quasar, gamma ray burst radio afterglow, or fast radio burst. The neutral hydrogen absorbers at different redshifts along the line of sight produce absorption lines at different part of the spectrum. A great advantage of the 21 cm forest is that it is a good probe of the spin temperature. Finding the usable bright radio source is the challenge for this kind of observation [5] (Fig. 15.4).

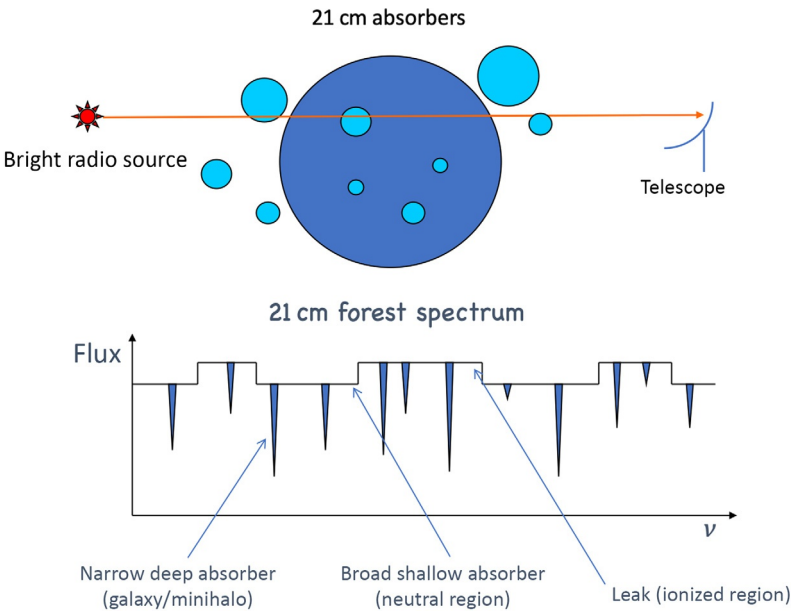

Fig. 15.4 Illustration of the 21 cm forest concept *(top)* and the absorption spectrum *(bottom)*.

The global averaged 21 cm spectrum gives the mean brightness temperature of the 21 cm emission as a function of redshift. In such an experiment, the flux of the whole sky or a large portion of it is integrated, so it can be done with a single dipole-type antenna that has a very wide beam. However, on top of the 21 cm signal, there is the much stronger foreground radiation. Nevertheless, the cosmology theory predicts some features in the global spectrum, such as the absorption trough at the cosmic dawn and Dark Age, and the sharp drop of the mean brightness temperature at the completion of the reionization. The global spectrum experiments can look for such features.

Below, we shall discuss mostly the 21 cm tomography experiments, as these are the primary efforts in 21 cm cosmological observation. This is more relevant for the big data problem, although we will also briefly review the progress in the 21 cm global spectrum experiment.

2.1 HI galaxy survey

The traditional way of cosmological observation with the 21 cm line is the HI galaxy survey, where one detects the 21 cm line of individual galaxies, just as in the typical optical galaxy redshift survey. Although the 21 cm line has long been detected in the

Milky Way and nearby galaxies, it is not easy to detect in the distant universe. For an interferometer array, the sensitivity is

$$\Delta S = \frac{2kT_{sys}}{\eta A\sqrt{2N_p\Delta\nu t}}$$

where T_{sys} is the system temperature, $\eta < 1$ is a factor describing quantization loss, A the effective area of one antenna, $N_p = n_a(n_a-1)/2$ is the number of antenna pairs, $\Delta\nu$ the bandwidth, and t the integration time. For large n_a, we may approximate

$$\Delta S = \frac{2kT_{sys}}{\eta A_{eff}\sqrt{\Delta\nu t}} \approx 2.74\left(\frac{1000}{\eta A_{eff}/T_{sys}}\right)\left(\frac{\Delta\nu t}{MHzs}\right)^{-1/2} mJy$$

As a reference, the typical A_{eff}/T_{sys} values for currently operating or upcoming arrays are, for example, eMERLIN(60), ASKAP(65), GMRT(250), JVLA(265), MeerKAT(321), and SKA1-mid (1650) [6].

The 21 cm line flux from the galaxy is given by

$$S_\nu = 4.25\times10^{-2}\frac{(1+z)^2}{(d_L/Mpc)^2}\left(\frac{M_{HI}}{10^6 M_\odot}\right)\left(\frac{\Delta V}{100 kms^{-1}}\right)^{-1} Jy$$

where d_L is the luminosity distance, M_{HI} is the mass in HI in the volume pixel (voxel), and ΔV is the velocity spread in the rest frame of the galaxy. For optimal detection, the voxel should be just the size of the galaxy. The limiting mass of the galaxy in terms of the flux limit is then

$$M_{HI} = 2.36\times10^{10}\left(\frac{D_L}{Gpc}\right)^2\frac{S_i}{mJy}\frac{\Delta\nu}{100 kms^{-1}}(1+z)^{-2}M_\odot$$

We see a massive galaxy in cosmological distance, sub-mJy sensitivity is required, and even for the SKA-1mid array, an integration time of several hours is needed.

Furthermore, in order to distinguish the different galaxies, one needs to achieve an angular resolution of a few arcseconds. This can be achieved with an array distributed over several tens of kilometers. Many interferometer arrays such as JVLA, ASKAP, MeerKAT, and SKA1-mid have baselines on such scales. However, if one wants to survey a sizable fraction of the sky (a few thousand square degrees or more) within a plausible time frame (a few years), and at each pointing integrate for a few hours as discussed above, then a large field of view is also required. For SKA1-mid, which is made up of antennas of 15 m aperture plus the 13 m MeerKAT antennas, the field of view is 0.8 degrees at 1420 MHz. This means that within each field of view, there would be at least

10^6 pixels, and then combined with $10^{3–4}$ frequency bins, each image cube would have at least 10^{10} voxels. For the whole survey, this would be 10^{14} voxels, posing a daunting task for imaging synthesis computation and data storage.

2.2 Intensity map observations

Another mode of observation is the intensity mapping observation, in which one maps the radiation intensity of the sky at a low resolution without attempting to resolve individual objects but to try to obtain the distribution on the larger scale (Fig. 15.5). This mode of observation was first considered for the epoch of reionization, where the main aim has been observing the reionization of the intergalactic medium, which is expected to happen on scales much larger than individual galaxies. Later, it is also considered for the HI survey at low- or mid-redshift surveys, which can be used to measure the baryon acoustic oscillation (BAO) features and to detect the dark energy. It is noted that such a survey can be conducted with a dedicated array of ∼100 m size [7,8].

The flux is related to the brightness temperature by

$$S = \frac{2k\Omega_b}{\lambda^2} T_b$$

where Ω_b is the solid angle of the synthesized beam while λ is the wavelength, so the temperature noise per pixel in intensity mapping is given by

$$\Delta T_b = \frac{\lambda^2 T_{sys}}{\Omega_b \eta A \sqrt{2N_p \Delta\nu t}}$$

For the same number of antennas and integration time, a smaller error on the temperature can be achieved with larger pixels.

Intensity mapping is carried out with a compact, almost aperture-filled array. This is necessary for the observation in the presence of a strong foreground. The foreground, especially the galactic synchrotron radiation, is always much larger (by 4–5 orders of magnitude) than the 21 cm signal. In order to extract the 21 cm signal, one has to use the property that along the line of sight, the foreground radiation has a relatively smooth spectrum while the 21 cm flux fluctuates randomly to separate them. However, at a given wavelength, the interferometry of each baseline probes a particular spatial mode of the sky radiation intensity [9], and as the wavelength varies, the interferometer array output gives different spatial modes. It is necessary to have a compact array for which the coverage in the baseline (u,v,w) space is dense to subtract the foreground modes.

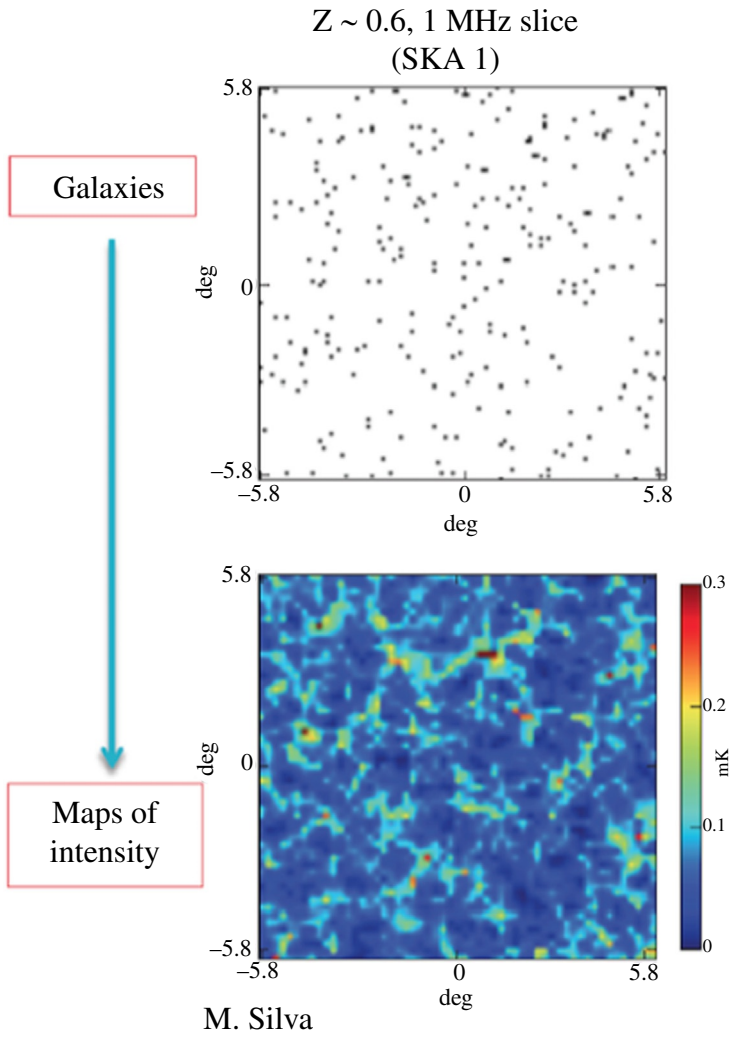

Fig. 15.5 An illustration of intensity mapping. Made by M. Silva.

A potential problem in the intensity mapping observation is the confusion of other spectral lines. In the galaxy survey, it is rare for two galaxies at two different redshifts to happen to be on the same line of sight and get confused with each other. In the intensity mapping observation, two spectral lines with wavelengths λ_1, λ_2 will be confused if they are located at redshifts z_1, z_2 if $(1+z_1)\ \lambda_1 = (1+z_2)\lambda_2$. Fortunately, at the low frequency of the 21 cm line, there are not many other lines of comparable strength, though for higher frequency molecule lines this may be a problem [10].

2.3 The 21 cm tomography experiments

The intensity mapping technique has been tried on existing telescopes, such as GBT (United States) and Parkes (Australia), and the results show that there is a correlation between the signal and the density of optical galaxies, though so far the signal has proved to be too weak to be detected positively in the autopower spectrum [11–14]. It has also been suggested to apply this technique to the new large multipurpose telescopes, such as the FAST (China) [15] and the SKA1-mid (South Africa) [16]. The layout of the SKA1-mid does not have a complete covering in the (u,v) plane, but it has been proposed that the many dishes in the array may be used individually as single dishes for this task.

All the EOR or low-frequency radio experiments such as the 21 cmA (China), LOFAR (Netherland and European countries), MWA (Australia), PAPER (United States-South Africa), HERA (United States-South Africa), and the SKA1-low (Australia) can make intensity mapping mode observations, though some of them also have longer baselines for other imaging observations. These experiments make observations in the meter-wave band, as the waves are not very sensitive to the details of the antenna much smaller than the wavelength, so most of them use some form of dipole antennas. HERA, however, uses dishes made of cheap structural material. As single dipoles are still relatively small, multiple dipoles are put together to form a station. The signals received by the dipoles are combined after amplification. The station serves as interferometer units. In some cases, multiple beams can also be formed for each station (Figs. 15.6 and 15.7).

Here, as a concrete example, we discuss the Tianlai experiment [17,18] in some detail. This experiment aims to detect the baryon acoustic oscillations of the Big Bang, so it is named Tianlai or "heavenly sound," a phrase that appeared in the classical works of the ancient Taoist philosopher Zhuangzi (Chuang-Tzu). The goal of the Tianlai experiment is to observe the 21 cm signal at redshift 0–3. At present, it is a pathfinder experiment to verify the basic principles (e.g., array design, calibration method, foreground subtraction techniques, etc.) and key technologies (e.g., the analog receiver, signal transportation system, and digital correlators). The pathfinder consists of three $15\,\mathrm{m} \times 40\,\mathrm{m}$ cylinders lined in the north-south direction and adjacent to each other as well as 16 dishes of 6 m aperture. Cylinder reflectors offer a low-cost way to build a telescope array with a large effective area. For the Tianlai cylinder array, the feeds are put on the focus line that lies along the N-S line, so that a narrow strip crossing the zenith could be observed at any time. The reflector is fixed on ground, as Earth rotates, the array maps out the whole observable

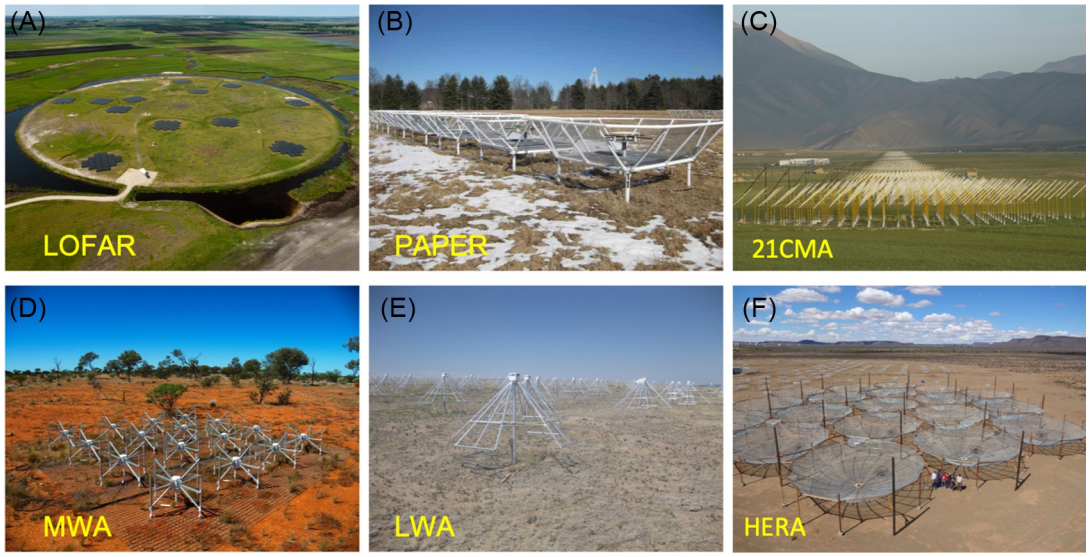

Fig. 15.6 Some EOR experiments. (A) From https://en.wikipedia.org/wiki/Low-Frequency_Array_(LOFAR)#/media/File:LOFAR_Superterp.jpg. (D) From https://commons.wikimedia.org/wiki/File:MWA_32T_Tile.jpg. (F) HERA Collaboration, from http://reionization.org/.

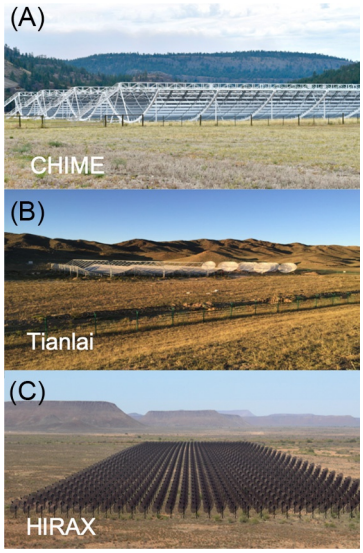

Fig. 15.7 Some mid-redshift 21 cm experiments. (A) From https://commons.wikimedia.org/wiki/File:Canadian_Hydrogen_Intensity_Mapping_Experiment_-_overall.jpg. (C) Used with permission from Cynthia Chiang.

part of the sky (see Fig. 15.8). As there is no movable part, the array is stable and can be constructed at low cost.

In the dish array, the instant field of view is a circular area defined by the primary beam of the dish. It can also make

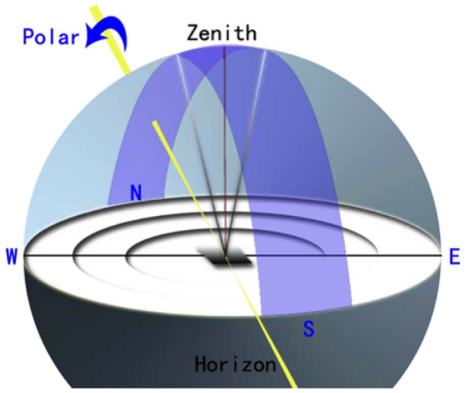

Fig. 15.8 The instant view of the sky by the cylinder array; the whole northern sky is surveyed as Earth rotates.

observations with the drift scan mode, that is, use the rotation of the Earth to scan over the area to be surveyed. Because of the finite size of the primary beam, one needs to adjust the declination angle of the dish in order to scan over the whole observable sky [19]. The dishes, however, also have certain advantages. For each receiver, the effective antenna area is large even for small dishes, so it is more sensitive. With a driver mechanism, the dishes can be pointed to different directions and track targets, making it easier for test and calibration.

The signals received by the feeds on the cylinder reflectors or dishes are first amplified with a low noise amplifier (LNA), then sent via a coaxial cable to an optical transmitter below the antenna. Here, the radio signal is used to modulate the optical signal, which is sent via a multifiber optical cable to the station house that is located in the nearby village about 7 km away. The optical signal is converted back to radio there, then downconverted to an intermediate frequency (IF), then further amplified and digitized. The digitized time series signal is fast Fourier transformed (FFTed), then cross-correlated in the FX correlator. The integrated cross-correlations (interferometry visibilities) are stored on hard drives for offline processing (Fig. 15.9).

2.4 The 21 cm global spectrum experiments

Besides tomographic observations, the global average 21 cm brightness temperature can also be measured. As the measurement is for the whole sky, angular resolution is not required, and a single antenna can be used for such a measurement. The EDGES (Arizona State and MIT) experiment is the first to make

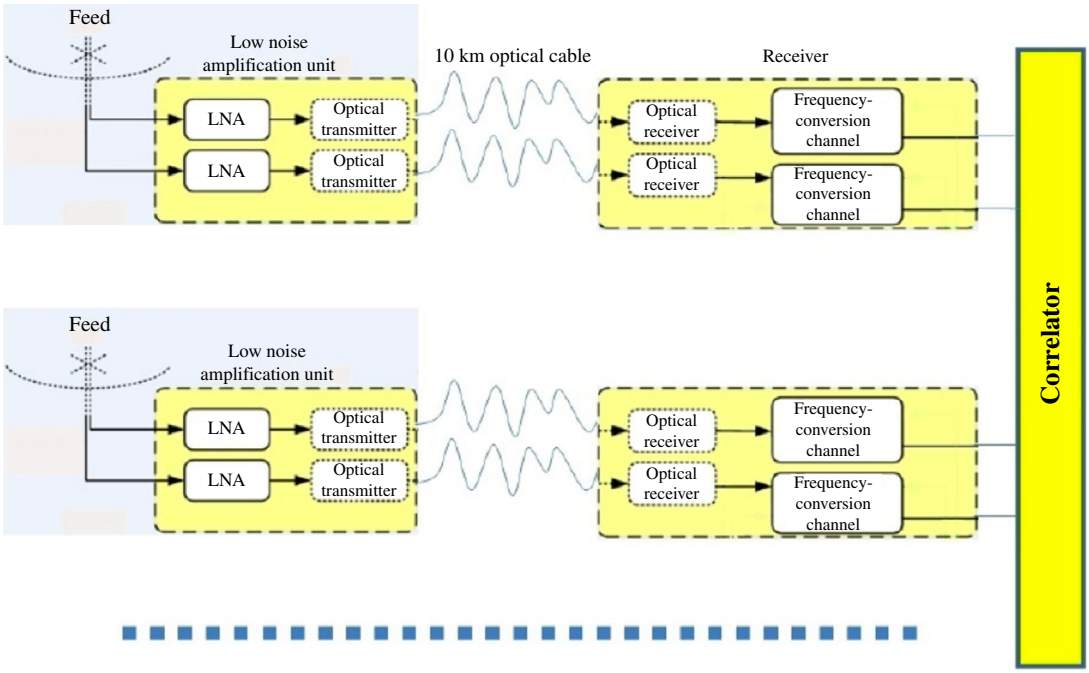

Fig. 15.9 Schematic view of the analog signal path of the Tianlai pathfinder array.

such experiments [20]. Subsequently, similar experiments have also been carried out by the SCI-HI (United States-Mexico) experiment, the PRIZM (South Africa and Canada) experiment, the SARAS (India) experiment, the Big Horn experiment (Australia), and LEDA (Harvard-Berkeley). There are also upcoming ones such as the REACH (Cambridge University) and Tianlai-Global (China) (Fig. 15.10). In addition, several lunar orbit mission concepts have been studied, including the DARE, DAPPER, and DSL [21].

In the global spectrum experiment, the whole sky average is to be measured. It does not require high angular resolution. In low frequency, the sky is quite bright, but the 21 cm signal is small so the experiment requires very high precision and dynamic range. Some usually ignored effects, such as the drifting of amplifier gain and the reflection in the circuit due to impedance mismatch, must all be considered.

In order to achieve such precision, the EDGES experiment receiver includes internal calibration mechanisms. As shown in Fig. 15.11, there is a switch at the base of the antenna that is connected to the circuitry for the measurement of antenna reflection, which can be used to measure the impedance of the antenna. The

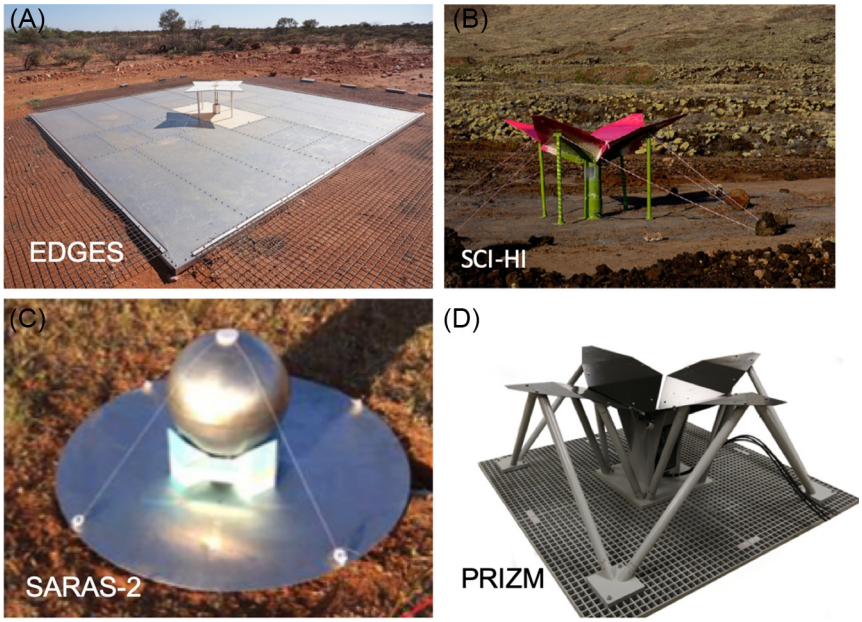

Fig. 15.10 Some global spectrum experiments. (A) From https://commons.wikimedia.org/wiki/File:EDGES_antenna.JPG. (B) From Tabitha C. Voytek et al. 2014, Probing the Dark Ages at z ~20: The SCI-HI 21 cm All-Sky Spectrum Experiment, ApJL 782 L9. (C) From Saurabh Singh et al. 2017, First results on the Epoch of Reionization from First Light with SARAS 2, ApJL 845 L12. (D) Used with permission from Cynthia Chiang.

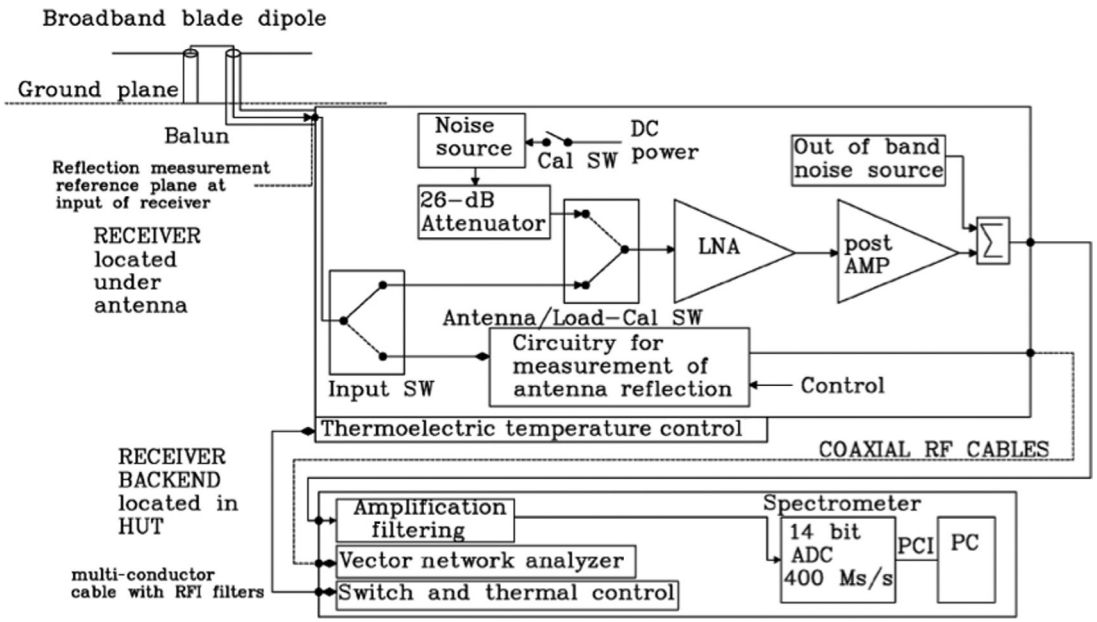

Fig. 15.11 Schematic of the EDGES receiver system [22].

input of the low noise amplifier (LNA) can also be switched between the antenna input and a noise source with an adjustable attenuator so that the gain of the LNA can be measured. The output is then digitized. The system will be automatically controlled to switch between measuring the sky signal and calibration with the internal noise source.

The other experiments have used principles similar to EDGES, though in detail the designs are each different. For example, different types of antennas have been employed. The antenna should be as independent of the frequency as possible because a frequency-dependent beam would mean that the antenna will see slightly different skies at the different frequencies, thus mixing the spatial pattern of the sky with the frequency feature to be measured. The SCI-HI and PRIZM experiments used a "flower petal" shaped antenna while the SARAS used a cone-sphere antenna. These experiments also designed the system to have a higher characteristic frequency so that in the band of observation, the system response is smoother. The price paid is that due to the imbalance of impedance, only a small fraction of the antenna power can be fed in to the amplifier, and hence the amplifier noise is more significant.

In 2018, the EDGES experiment announced that an absorption feature was detected around 78 MHz, corresponding to redshift $z\sim 18$, which is exactly the expected cosmic dawn time [22]. However, the detected absorption of 550 mK is much stronger than the prediction of most models—indeed, the trough is about twice as deep as the maximum allowed by the standard model. Various ideas have been proposed to explain such strong absorption, such as the gas could be cooler than usually assumed by an exotic interaction with dark matter that is generally colder than the gas due to earlier decoupling from the cosmic fluid. Alternatively, a strong radio background radiation other than the cosmic microwave background might be present in the early universe, so that the absorption signal is stronger. On the other hand, the required precision of the experiment is extremely high, so it might suffer from various contaminations or unexpected system errors. Further experiments are required to produce convincing results.

3 Data processing

3.1 Imaging and beam forming

The angular resolution of an observation is given by

$$\Theta \approx \frac{\lambda}{D}$$

where λ is the wavelength and D is the aperture of the instrument. Due to the long wavelength of the radio waves, in order to obtain a good angular resolution, large D is required. Even with the 500 m spherical telescope (FAST) that has a total size of 500 m and an aperture of 300 m, its angular resolution is merely 2.9 arcmin at the 21 cm wavelength.

In order to achieve higher angular resolution, an interferometer array is often used. The voltage ε_i induced on the antenna i can be written as

$$\varepsilon_i = \int d^2\hat{\mathbf{n}} A_i(\hat{\mathbf{n}}) E(\hat{\mathbf{n}}) e^{-i2\pi\vec{u_i}\cdot\hat{n}}$$

where $E(\mathbf{n})$ is the electric field induced by the radiation from direction $\mathbf{n}$ and $A_i(\mathbf{n})$ is the voltage response by antenna i. Under the assumption that the time-varying electric field is uncorrelated for any two different directions, as one would expect for astrophysical sources, the interferometer visibility, that is, the short time averaged correlation between the voltage signal of the two elements, is related to the sky temperature $T(\hat{n})$ by

$$V_{ij} = <\varepsilon_i{}^*\varepsilon_j> = \int d^2\hat{\mathbf{n}} A_{ij}(\hat{n}) T(\hat{\mathbf{n}}) e^{-i2\pi\vec{u}\cdot\hat{n}}$$

where $\mathbf{u}$ is the baseline vector between the elements i and j in wavelength units and A_{ij} is the beam response for the antenna pair (i,j). Roughly speaking, each pair of interferometry units (antennas) measures one Fourier component of the sky intensity [9]. The sky temperature can be recovered from the measured visibility data by inverting the above relation. If we define a reference point in the celestial sphere, and use it to set up a coordinate system (u,v,w), the vector pointing to an arbitrary point is defined in terms of the direction cosines (l, m, n) with respect to the (u,v,w) axes (see Fig. 15.12). Then, the above integration can be reduced to

$$V_{ij} = \int \frac{dl\,dm}{\sqrt{1-l^2-m^2}} A_{ij}(l,m)\, T(\mathbf{l},\mathbf{m})\, e^{-i2\pi\left(ul+vm+w\left(\sqrt{1-l^2-m^2}-1\right)\right)}$$

where we have used $l^2 + m^2 + n^2 = 1$.

With small angle approximation, the w-term can be neglected. The above is then reduced to a two-dimensional Fourier transform, and can be easily inverted as efficiently as an inverse two-dimensional Fourier transform. Of course, in reality the array baselines can only sample part of the (u,v) space, so one has to make interpolations from the measurement to obtain the visibility at regular grids. This procedure is called "gridding." In many interferometer arrays, the antenna distribution is sparse and the spacing between the antennas is far larger than the size of the antenna,

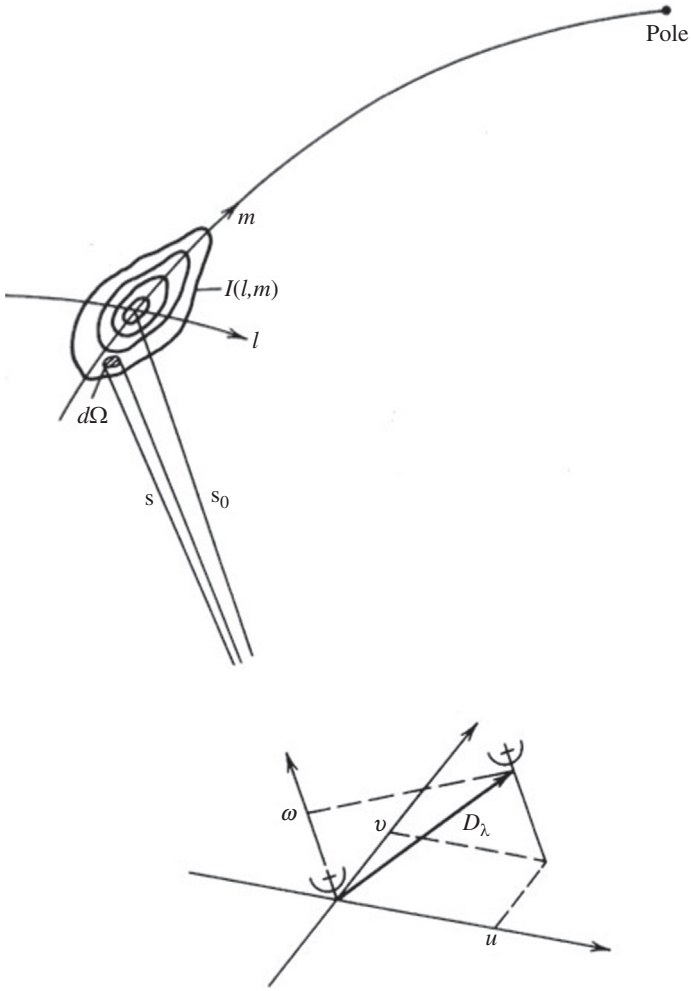

Fig. 15.12 A coordinate system often used in radio interferometry. From A.R. Thompson, J.M. Moran, G. Swenson Jr., Interferometry and Synthesis in Radio Astronomy, 3rd ed., Springer, Cham, Switzerland, 2017.

so that the geometric area of the array is far larger than the collecting area. In such an array, the sampling on the uv plane is also sparse. Because each (u,v) point corresponds to one Fourier mode of intensity distribution, the missing modes cannot be reconstructed accurately. However, usually a good image of the sky can still be obtained, thanks to the sparse nature of sky source distributions. Such an image often has strong side lobes. This can be improved by a deconvolution procedure, using either the CLEAN or maximum entropy method (MEM). The general procedures of such processing are reviewed in the usual radio astronomy textbooks [9].

An alternative to visibility-based image synthesis is beam forming. In this case, one can form a beam by

$$\mathrm{I}(\mathbf{n}) = |\sum_j w_j(\mathbf{n})\varepsilon_j|^2$$

where the weight for array element j is given by

$$w_j(\mathbf{n}) = \mathrm{e}^{2\pi i\,\mathbf{u}_j\cdot\mathbf{n}}$$

which compensates the phase difference between the element j and the array origin point for direction $\mathbf{n}$, so that for waves coming from direction $\mathbf{n}$, the voltages from the different array elements add constructively. Given sufficient computing power, multiple beams toward different directions can be formed simultaneously. Note that for an array with N elements, at most N independent beams can be formed simultaneously. More beams will be a linear combination of the first N independent beams. The beam forming is equivalent to the visibility measurement. From the definitions, the short time average of the beam output is

$$<\mathrm{I}(\mathbf{n})>=\sum_{ij} w_i^* w_j V_{ij}$$

It would be possible to form the time-averaged beam output by using the visibilities, as shown by the equation above. However, the beam-forming procedure allows using all the array elements to check for the radiation from a particular direction without time averaging, which is particularly suitable for transient source search or observation.

3.2 Foreground

In the intensity mapping observation mode, there is strong foreground radiation from other sources. The diffuse galactic synchrotron emission is the dominant foreground at low frequencies. It is approximately given by

$$T(\nu) = T_0\left(\frac{\nu}{\nu_0}\right)^{-\gamma}$$

with $\gamma \approx 2.7$, and for $\nu_0 = 150$ MHz, $\mathrm{T}_0 \approx 150$ K at high galactic latitude while greater near the galactic plane. There are also radio point sources, galactic free-free emissions, etc. Fortunately, these foreground signals are all smooth in the spectrum, so that the 21 cm signal can be extracted by removing the smooth components. This is illustrated in Fig. 15.13.

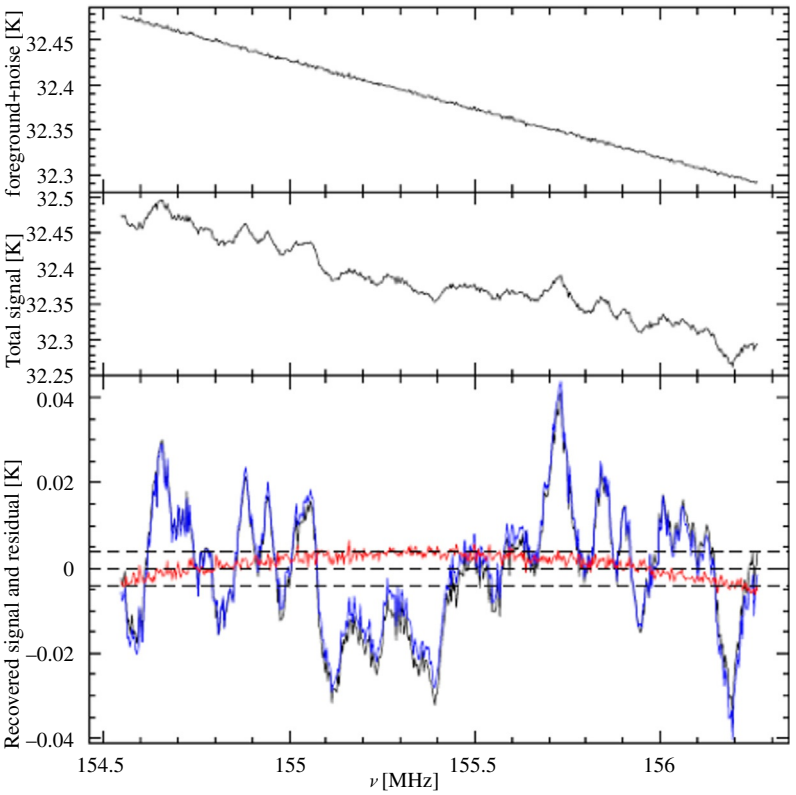

Fig. 15.13 The smooth foreground (top panel), foreground plus 21 cm signal *(middle panel)*, the original 21 cm signal *(black)*, the recovered signal after foreground removal (*blue*; dark gray in print version), and the residue difference (*red*; dark gray in print version) *(bottom panel)* [23].

The simplest method of foreground removal is polynomial fitting. We may assume the foreground brightness temperature is given in the form of

$$\log\frac{T}{T_0} = a_1 \log\frac{\nu}{\nu_0} + a_2\left(\log\frac{\nu}{\nu_0}\right)^2 + \ldots$$

then the smooth component can be removed.

However, in reality, the telescope response is neither known nor smooth, and the beam size generally depends on the frequency. This all makes the foreground removal task very difficult. Nevertheless, blind or semiblind methods have been proposed to remove the largely unknown foreground and extract the 21 cm signal.

As a concrete example, we shall consider the PCA/SVD method for foreground removal in the single dish survey with the GBT. The observation data are a mixture of the 21 cm signal and the foregrounds convolved with the unknown beam response function, plus the noise. The telescope response function is apparently

not smooth, for the data spectrum is far from smooth and the nonsmoothness is much higher than the estimated noise level.

Without knowing the precise response of the telescope, can we subtract out this nonsmooth "apparent foreground?" Suppose the data are related to the sky temperature linearly, so in discrete form, we can write down the matrix equation

$$\mathbf{y}=\mathbf{A}\,(\mathbf{sf}+\mathbf{s21})+\mathbf{n},$$

where $\mathbf{s_f}$ and $\mathbf{s_{21}}$ are the foreground and 21 cm components, respectively, $\mathbf{A}$ is the telescope response matrix, and $\mathbf{n}$ the noise. Here, we have omitted indices, but the data have indices of angular position and frequencies. Taking the autocorrelation, we have

$$<\mathbf{y}\,\mathbf{y}^{\dagger}>=\mathbf{A}\left(<\boldsymbol{s_f}\boldsymbol{s_f}^{\dagger}>+<\boldsymbol{s_{21}}\boldsymbol{s_{21}}^{\dagger}>+<\mathbf{n}\mathbf{n}^{\dagger}\right)\mathbf{A}^{\dagger}$$

or

$$\mathbf{C}_{\boldsymbol{y}}=\boldsymbol{A}(\mathbf{C_f}+\mathbf{C_{21}}+\mathbf{C_n})\,\mathbf{A}^{\dagger}$$

where we assume that the foreground, 21 cm signal, and noise are all uncorrelated. For different frequencies,

$$\mathbf{C_{21}}(\nu_1,\nu_2)\approx\mathbf{0}$$

while the foreground part is nonzero. This suggests that we subtract the components that have nonzero correlations.

This can be done by performing the principal component analysis (PCA). Solving the eigenvalue equation

$$\mathbf{C}\,\mathbf{x}=\lambda\,\mathbf{x}$$

one would obtain a series of eigenvalues and associate eigenvectors. Ordering them by the magnitude of the eigenvalue, we may subtract the largest ones (principal components), so then a large amount of foreground would be removed. In practice, the matrix involved is sometimes a nonsquare matrix. In such cases, the singular value decomposition (SVD) analysis can be performed instead.

The singular values for some GBT HI intensity mapping data are shown in Fig. 15.14. As one can see, the first few eigenvalues are quite high. The first few modes in frequency are shown in Fig. 15.15. Except for the first one, they are not smooth in frequency. With the subtraction of the first 10 or 20 modes, the variance of the data is significantly reduced.

The blind foreground subtraction procedure will also remove some 21 cm signal. This is unavoidable, but the amount of signal loss can be computed statistically. This is done by generating a mock 21 cm signal, which is generally much smaller than the data.

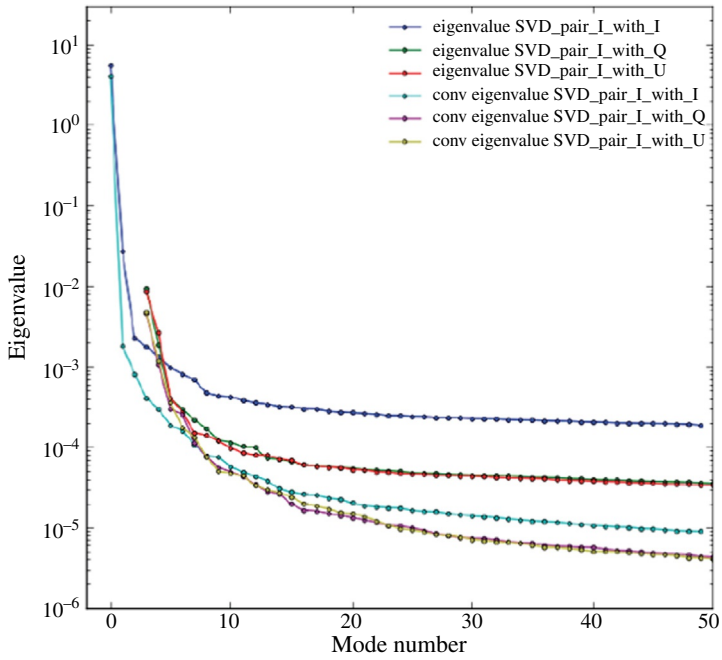

Fig. 15.14 The SVD eigenvalues in an analysis of GBT HI intensity mapping survey data.

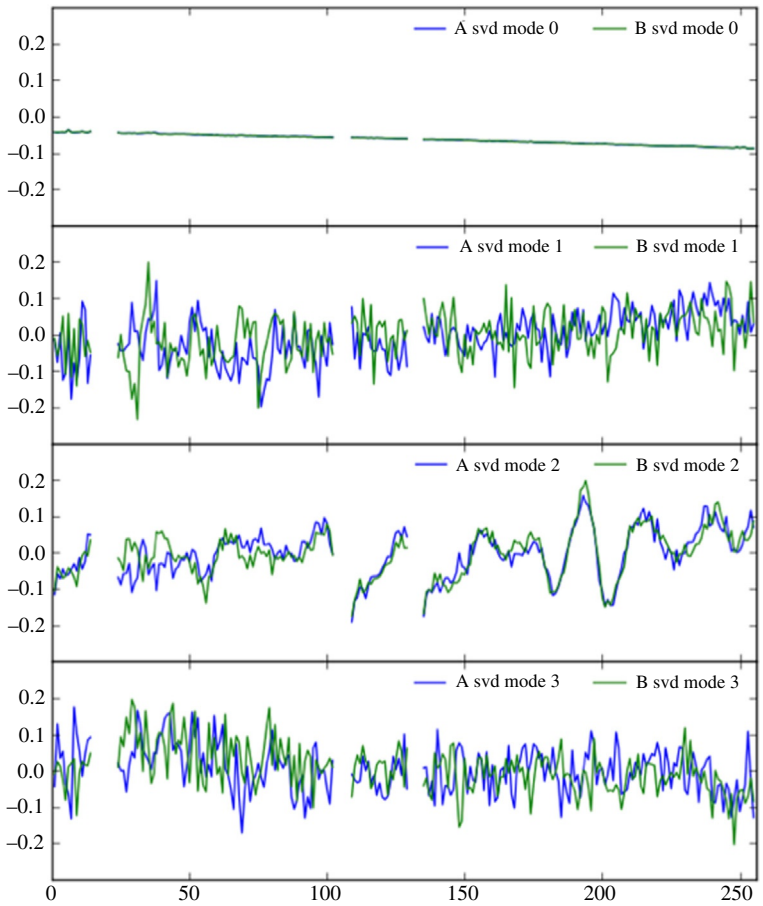

Fig. 15.15 The first few SVD modes of the GBT observation data.

We add such a mock signal into the dataset, perform the foreground subtraction procedure, and compute the amount of 21 cm signal loss. Repeat this simulation many times with different random numbers, and we will be able to estimate the signal loss transfer function [13].

Other semiblind foreground subtraction methods have also been proposed, including the independent component analysis (ICA), which uses the non-Gaussianity of the foregrounds to subtract them [24]. Generally speaking, the foreground subtraction can be put into a Bayesian framework to extract the 21 cm signal as much as possible by utilizing the known information.

3.3 The foreground wedge

For cosmology, the power spectrum is the observable to be measured. The cosmological 21 cm signal can be Fourier transformed,

$$\widetilde{T}\left(k_{\perp}, k_{\parallel}\right) = \int d^2 r_{\perp} dr_{\parallel}\, e^{-i\left(k_{\perp} r_{\perp} + k_{\parallel} r_{\parallel}\right)}\, T\left(r_{\perp}, r_{\parallel}\right).$$

Similarly, we can also Fourier transform the visibilities,

$$\widetilde{V}\left(\mathbf{b}, \eta\right) \approx \int d^2\mathbf{u}\ \widetilde{T}\left(\mathbf{u}, \boldsymbol{\eta}\right) \widetilde{\mathbf{A}}_{\boldsymbol{p}}\left(\frac{\mathbf{b}}{\lambda} - \mathbf{u}, \nu_0\right) \approx \widetilde{T}\left(\frac{\mathbf{b}}{\lambda}, \eta\right)$$

where η is the Fourier dual of frequency ν. In the last step of the above, we have made the approximation of ignoring the primary beam of the antenna, which ideally is smooth or even constant on the scale of interest. However, the same physical baseline probes different scales at different frequencies. The cosmological scales are related to the instrumental scales in the visibility by

$$k_{\perp} = \frac{2\pi\nu_0\, b}{c\, D_c}; k_{\parallel} = \frac{2\pi\nu_{21} H_0 E(z)\eta}{c(1+\mathrm{z})^2}$$

where D_c is the comoving distance, and

$$E(z) = \sqrt{\Omega_m (1+z)^3 + \Omega_{\Lambda}}.$$

The finest angular scales or maximum $k_{\perp}$ are determined by the angular resolution, which is determined by the longest baselines of the array. The largest angular scales (lowest $k_{\perp}$) are determined by the survey area. Along the radial direction, the finest resolution is limited by the spectral resolution of the instrument while the lowest $k_{\parallel}$ is limited by the frequency range of observation.

The foregrounds with a smooth or even flat spectrum will induce some foreground power that contaminates the 21 cm

signal. Naively, one might think this sets a lower limit in $k_{\parallel}$, above which the 21 cm signal can be measured. However, the fact that the same baseline measures different angular scales at different frequencies induces a mode-mixing effect, which tilts the limit in power spectrum scales. To see this, note that the measured power spectrum is given by

$$\hat{P}(\mathbf{k}) \propto \left\langle \left| \widetilde{V}_b(\tau) \right|^2 \right\rangle = \int \frac{d^3k}{(2\pi)^3} P(\mathbf{k}) \left| \int dy e^{ik_y y} A_p \left[\frac{cD_c}{2\pi b} (\alpha k_{\parallel} + 2\pi\tau), y \right] \right|^2 .$$

The squared term in the integrand corresponds to the window function of the power spectrum estimator. For unclustered diffuse point sources with flat spectra, the foreground power has the form $P(k) \propto \delta(k_{\parallel})$, that is, strong noise at low $k_{\parallel}$ as expected. Plugging this foreground power into the above expression, we would obtain

$$\hat{P}(\mathbf{k}) \propto \int dy A_p^2 \left[\frac{cD_c\tau}{b}, y \right] \equiv \overline{A_p^2} \left[\frac{k_{\parallel}}{k_{\perp}} \frac{c}{H_0} \frac{(1+z)}{E(z)} \right]$$

where $\overline{A_p^2}$ is the squared primary beam integrated over the direction perpendicular to the baseline. Even though the initial foreground is entirely at $k_{\parallel} = 0$, the power is leaked out, contaminating a region with $k_{\parallel} \propto k_{\perp}$. This is the so-called foreground wedge [25], as shown below in Fig. 15.16. The primary beam response falls off away from the pointing center, at an angle θ_0. In some cases, the primary beam is very wide, but it is still limited by the horizon. Thus, the function $\overline{A_p^2}$ drops to 0 at

$$k_{\parallel} = k_{\perp} \frac{H_0 D_c E(z) \theta_0}{c(1+z)},$$

This is the horizon wedge in the figure. Above the horizon wedge is the so-called EoR window where the 21 cm signal can be extracted without contamination.

4 Conclusion

The 21 cm line offers an observational probe that covers most of the observable cosmic volume, from the Dark Ages down to the modern universe. At present, a number of experiments are running or are being planned to measure the cosmological 21 cm signal at different redshifts, as shown in Fig. 15.17.

The low- and mid-redshift experiments are primarily designed to map the large-scale structure and measure the baryon acoustic

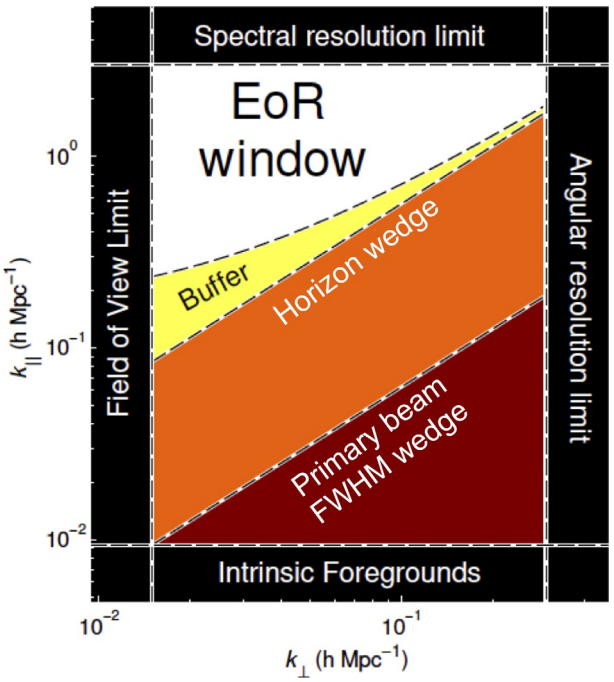

Fig. 15.16 The power spectrum in ($k_{\perp}$, $k_{||}$), showing the EoR wedge.

oscillation features to probe the dark energy equation of state. The existing 21 cm experiments, such as Tianlai, CHIME, HIRAX, and the intensity mapping survey of SKA-mid, are mostly dedicated to $z<3$. It has been proposed that a Stage 2 experiment could map the postreionization universe at $2<z<6$ [26]. The higher redshift experiments such as LOFAR, MWA, HERA, and the future SKA-low are looking into the epoch of reionization and cosmic dawns.

Ideas have also been advanced to go beyond the cosmic dawn into the cosmic Dark Ages. Due to the limitation of the Earth's ionosphere, such experiments may need to be conducted from the far side of the moon [27]. If fully explored, the Fourier modes that can be measured with the 21 cm line are about 10^6 that of the CMB modes [28]. Therefore, this contains the largest known wealth of information about the primordial fluctuations, which we can use to probe the cosmic origin. However, as the synchrotron foreground scales as $\nu^{-\gamma}$ where $\gamma \approx 2.5$, the foreground is stronger at lower frequencies, so it is also increasingly harder to make the measurement at the higher redshifts. This requires increasingly larger arrays to satisfy the basic sensitivity requirement.

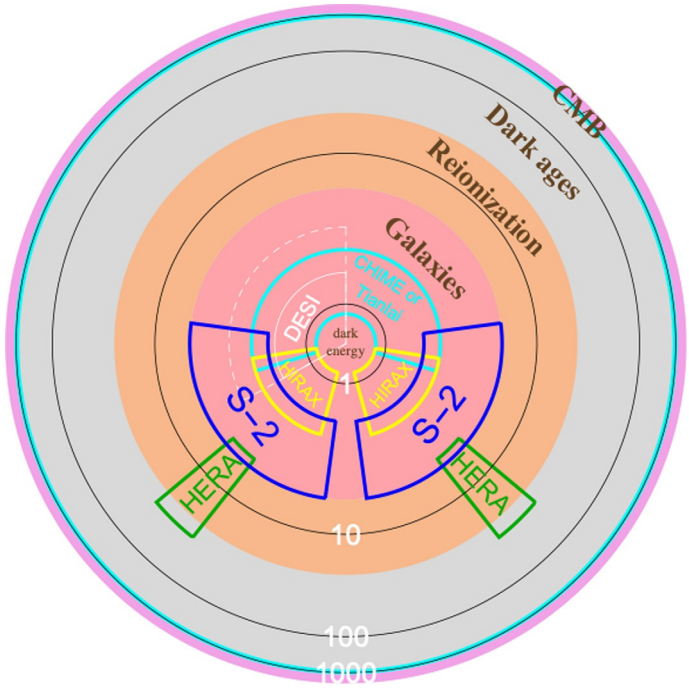

Fig. 15.17 A schematic two-dimensional representation of the observable universe where the area is proportional to the comoving volume, with the covered redshift range and survey field of the various experiments [26].

The real world 21 cm experiment also has to deal with many complications, especially the complicated telescope response and the contamination of the various foregrounds. Nevertheless, with the increase of sensitivity and growing computational capability, these problems will be overcome, and we expect 21 cm cosmology will come of age in the next decade.

References

[1] S.R. Furlanetto, S.P. Oh, F.H. Briggs, Cosmology at low frequencies: the 21cm transition and the high redshift universe, Phys. Rep. 433 (2006) 181.
[2] J.R. Pritchard, A. Loeb, 21 cm cosmology in the 21st century, Rep. Prog. Phys. 75 (2012) 086901.
[3] X. Chen, J. Miralda-Escude, The spin-kinetic temperature coupling and the heating rate due to Lyman alpha scattering before reionization: predictions for 21cm emission and absorption, Astrophys. J. 602 (2004) 1.
[4] X. Chen, J. Miralda-Escude, The 21cm signature of the first stars, Astrophys. J. 684 (2008) 18.
[5] B. Ciardi, S. Inoue, K.J. Mack, Y. Xu, G. Bernardi, 21 cm forest with the SKA, arxiv:1501.04425. (2015).
[6] P. Dewdney, SKA1 System Baseline Design V2, SKA-TEL-SKO-0000002, 26 February (2016).
[7] T. Chang, U.-L. Pen, J.B. Peterson, P. McDonald, Baryon acoustic oscillation intensity mapping as a test of dark energy, Phys. Rev. Lett. 100 (2008) 091303.

[8] H.J. Seo, et al., A ground-based 21cm baryon acoustic oscillation survey, Astrophys. J. 721 (2010) 164.
[9] A.R. Thompson, J.M. Moran, G. Swenson Jr., Interferometry and Synthesis in Radio Astronomy, third ed., Springer, Cham, Switzerland, 2017.
[10] Y. Gong, et al., The OH line contamination of 21 cm intensity fluctuation measurements for $z=1\sim4$, Astrophys. J. Lett. 740 (2011) 20.
[11] T. Chang, et al., Hydrogen 21-cm intensity mapping at redshift 0.8, Nature 466 (2010) 463.
[12] K.W. Masui, et al., Measurement of 21 cm brightness fluctuations at $z\sim0.8$ in cross-correlation, Astrophys. J. Lett. 763 (2013) 20.
[13] E.R. Switzer, et al., Determination of $z\sim0.8$ neutral hydrogen fluctuations using the 21 cm intensity mapping auto-correlation, Mon. Not. R. Astron. Soc. Lett. 434 (2013) 46.
[14] C.J. Anderson, et al., Low-amplitude clustering in low-redshift 21-cm intensity maps cross-correlated with 2dF galaxy densities, Mon. Not. R. Astron. Soc. 476 (2017) 3382.
[15] W. Hu, et al., Forecast for FAST: from galaxies survey to intensity mapping, (2019). arxiv:1909.10946.
[16] D.J. Bacon, et al., Cosmology with Phase 1 of the Square Kilometre Array, Red Book 2018: Technical specifications and performance forecasts, (2018) arxiv:1811.02743.
[17] X. Chen, The Tianlai project: a 21cm cosmology experiment, Int. J. Phys. Conf. Ser. 12 (2012) 256 Proceeding of the 2nd Galileo-Xu Guangqi Meeting, arxiv:1212.6278.
[18] Y. Xu, X. Wang, X. Chen, Forecasts on the dark energy and primordial non-Gaussianity observations with the Tianlai cylinder array, Astrophys. J. 798 (2015) 40.
[19] J. Zhang, et al., Sky reconstruction from transit visibilities: PAON-4 and Tianlai dish array, Mon. Not. R. Astron. Soc. 461 (2016) 1950.
[20] J.D. Bowman, A.E.E. Rogers, J.N. Hweitt, Toward empirical constraints on the global redshifted 21 cm brightness temperature during the epoch of reionization, Astrophys. J. 676 (2008) 1.
[21] X. Chen, et al., Discovering the sky at the longest wavelengths with small satellite constellations, (2019). arxiv:1907.10853.
[22] J.D. Bowman, et al., An absorption profile centred at 78 megahertz in the sky-averaged spectrum, Nature 555 (2018) 67.
[23] X. Wang, et al., Twenty-one centimeter tomography with foregrounds, Astrophys. J. 650 (2006) 529.
[24] E. Chapman, et al., Foreground removal using FASTICA: a showcase of LOFAR-EoR, Mon. Not. R. Astron. Soc. 423 (2012) 2518.
[25] M.F. Morales, et al., Four fundamental foreground power spectrum shapes for 21 cm cosmology observations, Astrophys. J. 752 (2012) 137.
[26] R. Ansari, et al., Cosmic visions dark energy: inflation and early dark energy with a stage II hydrogen intensity mapping experiment, (2018)arxiv:1810.09572.
[27] L. Koopmans, et al., Peering into the dark (ages) with low-frequency space interferometers, (2019). White paper submitted to ESA voyage 2050, arxiv:1908.04296.
[28] A. Loeb, M. Zaldarriaga, Measuring the small-scale power spectrum of cosmic density fluctuations through 21cm tomography prior to the epoch of structure formation, Phys. Rev. Lett. 92 (2004) 1301.

Further reading

A. Liu, R. Shaw, Data analysis for precision 21cm cosmology, (2019) arxiv:1907.08211.