# Remote Sensing

*Remote sensing across the spectrum*

Many of the topics in the previous parts of this course are relevant to the satellite and space probe technology that underpins remote sensing from space. For the rest of the course I want to concentrate on techniques used directly in remote sensing. It is a huge topic, and a whole course introducing remote sensing wouldn't cover a fraction of the material that a specialist would know. You'd expect a physics course to concentrate on how it works and we'll look at the ideas behind selected topics. Many of the class will meet remote sensing in practice in Geology and Geography courses in later years.

The electromagnetic spectrum is the key to remote sensing from space. You've seen its extent in earlier slides and found out how to calculate the frequency for any given wavelength, and vice versa. Make sure you appreciate the order of the electromagnetic spectrum regions. Remote sensing really does stretch from radio wavelengths to $\gamma$-rays, though $\gamma$-rays generated on Earth won't penetrate the Earth's atmosphere. $\gamma$-ray sensing is done around planets and moons that have no atmosphere.

Broadly speaking, remote sensing can be divided into **active** and **passive** techniques. Passive techniques detect what is given out 'naturally' by the target. IR and light are the obvious parts of the spectrum to use, even though the light is reflected sunlight. **Active** techniques are techniques in which the probe itself emits radiation and measures how much is returned by reflection.

*Microwaves*

I'm going to concentrate on a part of the spectrum you are less familiar with – microwaves. Everyone has met the microwave oven, though hopefully you haven't experienced any microwaves leaking out. Microwaves are completely invisible so how would we know if there are any in this room? We'd need a receiver. Unfortunately microwaves cover a big slice of the spectrum and man-made receivers are all highly tuned to individual parts of the spectrum. This is unlike our eyes, which detect the whole of the visible spectrum. If a blind person asked you if there was any light in the room you wouldn't need to ask first which colour of light he was talking about. It's almost the same with our heat sensing skin. Hold the back of your hand towards a bonfire and you'll feel the IR radiation even though it's spread over a wide range of wavelengths.

With microwave radiation, its frequency and wavelength are important for its use and there are distinctive microwave bands denoted by various letters. Thus you might see receivers or transmitters described as C-band or X-band. The table on the slide gives the approximate frequencies of the bands. You'll find some variation of the band limits depending on who's using the symbols. Mobile phones use the lowest frequency microwaves, below those in this table. Police radars use ~24.1 GHz; microwave ovens are in the S-band; satellite TV in Europe is generally in the X and $K_u$ bands.

*Radar*

Radar (**Ra**dio **d**etection **a**nd **r**anging) was the remote sensing technique that turned the tide of the Second World War for Britain. Radar is generally considered to have been invented by

Robert Watson-Watt, whose name I mentioned in connection with the ionosphere. John Logie Baird and others had similar ideas earlier but Watson-Watt was certainly responsible for building a successful radar system that was replicated along the coast of Britain in the second half of the 1930s as part of the country's air defences. The system was known as 'Chain Home'. The slide shows him pointing to one of the original radar screens he devised. Watson-Watt came from the small NE of Scotland town of Brechin.

In many radar systems, the aerial that both emits pulses and receives their reflection rotates around and the beam of the display tube sweeps round in sympathy. The traditional display tube was circular and a thin green line representing the beam came from the centre of the tube and stretched out in proportion to the time it toll for the radar reflection to be returned. When a reflection arrives, the beam brightens and a dot appears on the screen. By using a long-persistence screen, one where the image takes a good few seconds to fade, the dot is still there by the time the beam sweeps around again to give the reflection a few seconds later. If the object is moving, then the dot will move. How far away the object is can be calculated from the time it takes to pick up the reflected pulse. The speed of whatever is reflecting the pulse can be measured both across the field of view and towards or away from the radar. It is a powerful technique that works even in the presence of cloud. Let's look at it in a bit more detail.

*Distance detection*

For radar to work, you need a directional aerial and electronics to measure accurately the time it takes a pulse to travel out and be reflected. If that time is $t$, then the distance away of the reflecting object is v$t$/2, where v is the speed of travel of the pulse.

Microwaves are used for three reasons.

- First, because to get directionality in the radar beam, you need an aerial that is many wavelengths wide, as we'll see shortly. A rotating aerial has to be a modest size and hence the wavelength has to be a small fraction of a modest size, which takes you into microwave frequencies. There are two other relevant details.
- Objects only reflect back an appreciable amount of radiation if it is much shorter in wavelength than the size of the object. Otherwise the radiation effectively bends around the object and keeps going. You can see the same sort of behaviour happening with water waves striking an obstacle such as an isolated rock or a vessel in the water. If the obstacle is smaller than the wavelength of the waves then just beyond the obstacle you'll see the wave pattern almost the same as it would have been without the obstacle. If the obstacle is big, you'll see a substantial area that has been shielded from the waves and in front of the obstacle you'll see conspicuous reflected waves.
- Finally, if you want to locate objects to within a few metres, then you need a very sharply defined radio pulse, sharply defined in space to a fraction of a metre and sharply defined in time so your electronics can precisely measure the reflection time.

  [You can get a rough idea of what's going on by thinking how well you must time your pulses to locate your object to within 10 metres. Radio waves travel about a metre every 3 nanoseconds. Hence you need timing to about 30 nanoseconds if you want this accuracy, which is a 0.03 μs. To define a pulse sharply in time to within a time interval $\Delta t$ you need to include in it a range of frequencies $\Delta f$, where $\Delta f.\Delta t \approx 1$. Hence if $\Delta t$ = 0.03 μs, then $\Delta f \approx 30\times10^6$ = 30 MHz. This must be the spread in frequencies of the signal. The spread sits on top of the basic sending frequencies in what is termed *modulation*. It's the same when you broadcast from a radio station. The spread in frequencies from an audio broadcast station may be a few tens of kHz but the basic frequency is MHz. For radar, the spread is MHz and the basic frequency is GHz.]

There is a second way of detecting motion of the object that is used in the likes of rain radar, where the individual drops are not detected, and police speed radar (which is remote sensing!). This method uses the Doppler effect. The Doppler effect is the change in frequency when the source of waves is moving relative to the detector. If the source is moving away, then the frequency is lowered. If the source is moving towards you, then the frequency is raised. The classic audible demonstration of the Doppler effect (which works for sound waves too) quoted in textbooks is the express train whistling as it rushes past you as you are standing on the platform of a suburban station. That was in the days when trains had steam activated whistles that made a long continuous note. Nowadays trains make he-hawing noises and don't rush through suburban stations with siren sounding. Doppler himself used a train to demonstrate the effect in the very early days of the railway. He hired an open carriage and a band of trumpet players. They stood out in the open and played loudly as the train went past him and his observers. The observers estimated the change in pitch of the sound from a higher pitch as the musicians came towards them to a lower pitch as they receded. Interesting history but back to radar!

Nowadays the Doppler effect is well established, not only by experiment but also from first principles. It works for electromagnetic radiation as well as sound. The slide shows how. The effect depends only on the ratio of the speed of motion to the speed of light (or radio waves), which ratio is terribly small. However, for GHz waves, the frequency is very high so it's not hard to detect a frequency shift of say 1 kHz, which corresponds to a speed of 30 ms$^{-1}$.

### *Returning signal falls off as 1/d$^4$*

We've met the 'inverse square law' that governs how quickly radiation intensity falls off as you recede from a source. There were examples of this at work in two homework sheets. A serious problem with radar is that the echo signal falls off as the inverse 4th power of the distance of the object from the transmitter. You can see that this must be so on the slide. What makes radar reflection different is that it is not mirror reflection but diffuse reflection. The object sensed re-radiates the signal over a wide range of angles, acting like a new source. Basically, the inverse square law acts on both the signal reaching the reflecting object, the signal decreasing as the inverse square of the distance of the reflecting object, and the also on the signal returning from the object to the receiver. The examples on the slide shows that merely making the object 3 times further away will decrease the signal almost a hundred-fold. If the range of objects varies by a factor of 10, then the strengths of returning signals varies by a factor of ten thousand.

The inverse 4th power holds when you are looking at the return signal from a specific object that is a varying distance away. When you are reflecting from surfaces at different distances, the situation is different. As your radar beam travels further from the sending aerial, it illuminates a greater surface area. Indeed, for a fixed width of beam, the surface area increases as the square of the distance the beam reaches. If the whole of this area reflects a signal back, then you can see that we'll get much more signal that if a single object is taken further away. I'm not going to go into the detail of this.

*The basis of electromagnetic radiation*

Radio and microwaves have never played a greater role in society than they do today. Just think of radio, TV, mobile phones, wifi networks, radio mikes and so on. We should all know a bit about radio generation and reception as part of our general education these days, never minding its particular importance in satellite communication and remote sensing.

The first important fact to know about electromagnetic radiation is that it is produced by oscillating electric charges. That is the basis of all electromagnetic radiation. An electric current is nothing more than a moving charge. Hence an oscillating current radiates. [Strictly speaking I should say an accelerating charge radiates but I'm thinking in this context of the deliberate generation of radiation of known frequencies]. Our electric mains supplies an alternating current to all the devices in our house. Does this mean that all these devices are radiating? "Yes", they are. The energy radiated out per second varies as the 4th power of the frequency involved and in the case of domestic mains in Europe, the frequency is 50 Hz, which is very low. One consequence is that the radiation from mains wiring is much too little to be a hazard to health. It's also too little to cost you any money in electricity bills for energy your house has radiated and not used. It would be a different story if the frequency were 1 kHz, for radiation at 20 times the frequency would then be $20^4 = 1.6\times10^5$ times as great. The 50 Hz radiation can be picked up by sensitive electronics and can be a plague, for example, when you are trying to measure small electrical signals on an oscilloscope and the leads of your oscilloscope keep picking up this 50 HZ radiation.

*Transmitting & receiving*

The purpose of a transmitting aerial is to take the power generated in the transmitter and convert it into electromagnetic waves. If the aerial is doing its job well, no power is reflected back from the aerial into the transmitter. The power flows one way. The purpose of a receiving aerial is to absorb electromagnetic energy without reflecting power back the way it came. The retina at the back of your eye is almost black because it is so good at absorbing radiation and not reflecting any. In the language of radio reception, it makes a first class receiving aerial. You won't be surprised at the first law of aerials – *a good transmitting aerial makes an equally good receiving aerial*.

*Aerial considerations*

Considerations of wavelength are important in the microwave part of the spectrum in a way they are not in the IR and visible parts of the spectrum. I want to spend a few slides talking about radio aerials.

The simplest effective radio aerial is called a *half-wave dipole*. It consists of two separate wires or rods each a quarter of a wavelength long, with a gap in the middle. It is fed in the middle by two wires, one of which goes to one half of the aerial and one to the other. Think of it first as a transmitting aerial. The feed induces an alternating current in the two halves of the dipole, a current that is greatest in the centre and goes to zero at the tips.

Imagine your dipole aerial is vertical. The radiation does not come out equally in all directions. The sketch on the right in the slide shows the variation in a vertical plane. There is NO radiation parallel to the direction of the dipole, i.e. vertically upwards. There is the maximum in the horizontal plane. In fact the radiation varies as $\sin^2\theta$, where θ is the angle above or below the horizontal plane. This variation is plotted via Excel on the slide. The red line shows the strength of the radiation at different angles (θ) in a vertical plane.

In any horizontal plane the aerial looks the same from all directions and you should not be surprised that the radiation is the same all around the aerial. This is shown in the polar diagram at the bottom left of the slide. The red line shows the strength of the radiation for 360º all round. The radiation is equally strong in all directions. Put another way, a vertical dipole has no directionality in the horizontal plane. It's not going to be much use for radar, where you need a directed signal. It's good for picking up signals when you don't know in advance from which direction the signals might be coming from, or when signals might be coming from any direction, such as wifi signals to your portable PC.

*Making the aerial more directive*

The sight of directive aerials on the chimneys of houses is still quite common. They work by adding aerial elements side by side in a line. For the sake of taking specific examples, the slide shows a 2-element aerial and a 4-element aerial with a signal feed to all the elements. All the elements are half-a-wavelength long and the elements have been placed half-a-wavelength apart. I've imagined them as transmitting elements where successive elements are fed with signals in anti-phase, i.e. the radiation from each element is 180º out-of-phase with the radiation from neighbouring elements. You can make other choices and get different results.

The Excel plots show the variation of intensity in a horizontal plane, the plane that with a single dipole produced an equal amount of radiation in all directions. You can see now that the aerials have directive properties. The orange and blue coloured dots show what the aerials look like when looking down on them (the colours representing the two phases of the dipoles). The maximum amount of radiation is in the direction of the elements. The 4-element aerial is more directive than the 2-element aerial.

The general shape of the plots should make sense to you, especially if you've understood the diffraction grating story in our optics course. Look at the twin dipole example. Along a line at right angles to the two dipoles every point receives radiation from two sources that are exactly out-of-phase. Therefore the result is no signal. 'Destructive interference' is the optical phrase. Now look in the direction of the line of dipoles. The radiation from the further dipole is half a wavelength out of phase because the dipole is half a wavelength further away. That's a phase difference of 180º. An additional 180º is added on because it was the given condition that each element was fed with a signal 180º out-of-phase with the radiation from neighbouring elements. Hence in the direction in line with the two dipoles the total phase difference between the radiation from the two elements is 360º and the radiation adds together in phase and gives a maximum intensity. Hopefully what's happening makes sense if you know about adding together separate waves that may or may not be in step.

*Many elements improve an aerial's directionality*

The message from the previous slide can be taken further. In this picture I've calculated the radiation pattern from a 12-element array of our design. If you compare the directionality now with that of the versions on the previous slide, you'll see it's a lot better. Such devices used as transmitting aerials are called *phased arrays* and we'll meet them again shortly in satellite radars.

*Changing the phase of a phased array*

One very useful feature of a phased array transmitting aerial is that by altering the relative phase of signal at each element you can change the angle at which the radiation is directed. The slide shows what happens if instead of having each successive element of our array in anti-phase, we put them all in-phase. The direction of maximum transmission swings around by 90° and is now broadside on to our line of elements. Those who are taking our parallel *Light Science* course would expect this. The array behaves like 12 optical slits in a line. By changing the phase less drastically, you can alter the direction of propagation at will.

The plot at the lower left shows what happens when the phase change between elements is made 45°. You can see how the directionality swings round a bit in space.

We'll meet phased-arrays shortly, when looking at radar imaging from space, where the phased array is carried by the satellite. Another example of a phased array in action is at the impressive Fylingdales monitoring site run by the RAF. The space operations section at Fylingdales in deepest Yorkshire countryside monitors every piece of space junk bigger than 100 mm orbiting the Earth, some 10,000 objects, as well as some 4000 genuine payload carrying objects, mainly satellites. The phased-array, with no moving parts, is housed in an 8-storey high building described as being reminiscent of a Mayan temple. It replaced 3 large conventional moving radar systems. Commercial success, safety and even lives depend on knowing where the junk is in space. Fylingdales obliges.

*Yagis and parasitic elements*

The aerials on our chimneys are not quite of the same design although they use similar principles. They have only a single, active dipole connected to the receiver. The other dipoles are called *parasitic elements*. The incident radiation induces alternating currents in the parasitic elements. These then re-radiate. This re-radiation is received by the dipole and at the design wavelength it is all in-phase with the dipole's own signal. The combined effect of the dipole's own signal and the signals from the parasitic elements is a much stronger received signal, but only at the design wavelength and for signals arriving from a restricted range of directions. The directionality comes from the fact that radio waves coming in sideways to them have wavefronts that seem further apart and hence the phase of the re-radiated signals by the parasitic elements is not right to produce a strong signal at the active element.

*Parabolic reflectors*

The yagi aerials on our chimneys operate in the region of the radio spectrum a bit below microwaves. Microwaves have a sufficiently small wavelength that the even more directional spherical or parabolic dishes can be used without creating a device that's too cumbersome. This is the reason that sat dishes on the walls of our houses are this shape. At 10 GHz, the wavelength is 3 cm. A dish only 75 cm in diameter is therefore 25 wavelengths wide. Compare that with the 12 element phased array I calculated two slides back, which had only a length of 5.5 wavelengths.

The slide shows the radiation pattern from a spherical dish 25 wavelengths in diameter. It makes little difference to the overall shape of the radiation pattern if the dish is parabolic. You can see that almost all the radiation is given out within about 2° either side of the central direction. What you are experiencing here is what optical science calls the *diffraction pattern* of the dish. It follows a well-known shape and the first minimum after the central maximum occurs at an angle of $1.22\lambda/d$, where $d$ is the diameter of the dish. A little of the power goes into what are called *side-lobes* but very little. $1.22\lambda/d$ therefore gives the angle within which almost all the power of such a dish is spread into. Remembering the maxim that a good transmitting aerial is a good receiving aerial, then this figure applies equally well to the directionality of the dish as a receiving aerial. If you ever try to set up a satellite dish on the side of the house you'll find that even if you know the bearing of the satellite, it's a tricky operation because of the directionality of the dish. A sensitive signal meter is a great help.

*The bigger the dish, the more directive it is*

As far as directionality is concerned, size is almost everything. How big is a dish 100 wavelengths wide at 10 GHz? The slide tells you. What does its directional pattern look like? The slide shows the calculation. If you look closely, you'll see it's just the same as the pattern of the previous dish that was 25 wavelengths in diameter, except that the pattern is now compressed into a quarter of the angular size, so that almost all of the radiation transmitted or received is with 0.5° degrees and not 2°. I hope I'm not labouring the point. You can work out how big a circle this dish will 'see' at the distance of a geostationary satellite and it comes out to about 500 km in radius or nearly 1000 km across. Even this dish 'sees' quite a lot. On the negative side, if such a dish shakes in the wind, a vibration of half a degree only is enough to lose the signal. Also, it's important that the satellite doesn't drift too far from its intended position otherwise a fixed uplink dish will lose it.

The 76 m wide Jodrell Bank radiotelescope was *a really big dish* when it was built over half-a-century ago and it is still one of the largest in the world. It looks very impressive from over the hedgerows of the nearby Cheshire country lanes. It was built big for two reasons: directional accuracy and ability to collect weak signals. It was designed as a radio-telescope before the first satellite was ever put in orbit but it is equally good at picking up signals from distant space probes.

*Satellite coverage works the same way*

Now you've got the idea, you can see how to estimate the coverage over the ground that a satellite operator's dish will have. It depends on their satellite aerial. The slide gives an example calculation.

I'll give an example pertinent to remote sensing. Global sea surface temperature measurements are directly relevant to weather forecasting and climate modelling. They have been measured from satellites world-wide since 1979 by detecting the IR coming from the sea. If the sensor is on a polar orbiting satellite at a height 700 km then 1 km on Earth subtends an angle of 1/700 radians or about 5 minutes of arc. At a wavelength of 5 microns, an aerial only 50 mm across therefore has adequate resolution; a larger one even finer resolution. Sea surface salinity can be deduced from the microwave brightness of the sea (and some other input) at a frequency of about 1.5 GHz, wavelength 0.2 m. To obtain comparable resolution one would need a satellite aerial some 140 m wide. It took 3 decades from 1979 before sea surface salinity was measured globally from space and even then, into the 21st century, the resolution is a lot worse that for temperatures. In fact the technique of aperture synthesis (to be discussed shortly) is now applied to these measurements, a technique that effectively makes a small aerial appear like a big one, though at some cost.

In the communications arena, satellite operators who have a specific, well-located audience for their signals don’t want to be spreading their signals across thousands of square km where the audience either doesn’t exist, doesn’t pay or, for reasons of privacy, shouldn’t be listening. Other operators such as GPS systems are looking for a very wide coverage. There are related concerns. If a satellite or probe has a very directional dish, say radiating within 1°, then it has to maintain its orientation within that angle. If it loses orientation then it may be neither able to receive signals from Earth or send them back to the control centre. Many probes will have a very simple low-directional aerial just for the purpose of being able to pick up command signals should the satellite go into a spin due to failure of the orientation procedures and circuitry.

Beamspread is significant in the context of communication with remote probes. Microwaves have been used to communicate with probes partly because the uncertainty in the location of the probe in space has to be within the beam width of microwave transmitters. However, think about the problem of communicating with a probe at only the distance of Mars, say $10^8$ km. A microwave beam of width as narrow as 1 minute of arc has a spread of about 30,000 km at this distance and hence energy is being transmitted over a hugely greater area than is needed. If communications could be established by a more directed laser beam, then the signal would be confined to a smaller area at the distance of the probe and hence less power would be wasted. As we’ll see later, other things being equal, a stronger signal enables communication at faster speeds and hence a greater capacity for the link. Why should laser beams be more directive? Simply because the wavelength of an optical laser is about 20,000 times less than microwaves of wavelength 1 cm. Hence in terms of directionality a laser beam say 100 mm in diameter is as directive as a microwave beam from a dish 2 km in diameter, much larger than any steerable microwave aerial. Laser communication was tested successfully with the Messenger spacecraft that explored Mercury with a total beamspread of about a quarter of a minute of arc. Expect to hear more about direct communication with space probes using lasers in the future.

*A modern communications satellite*

This photograph of an Inmarsat-4 under construction in Britain shows the scale of a communications satellite. The Inmarsat geostationary group of 3 satellites were brought into operation to provide broad-band commercial services almost anywhere in the world. Each satellite cost £200 million over 10 years ago. They were mainly built in the UK by EADS Astrium, the UK’s principal satellite manufacturer. Each satellite is over 7 m long and when the solar panels are extended they spread over 45 m. The power output of the satellite is a good 10 KW, more at the start of its life, a bit less towards the end.

*Relative performance is measured in dB*

You’ll see signal strengths and other related quantities measured in *decibels* and it’s worth knowing what a decibel is. Decibels are abbreviated to dB. You might guess that ‘deci’ has something to do with one tenth, which it does because decibels are tenths of a bel, a unit seldom met with, and perhaps bels have something to do with sound. Decibels are indeed met with in sound and any sound level meter you get hold of will be calibrated in dB. The unit is named after that famous Scotsman Alexander Graham Bell, whose invention allowed people to talk at a remote distance. Thank him for what has led to today's mobile phones. A decibel is a measure of the ratio of two power levels, or in this context two signal intensities. Decibels are not just used for sound but are used a lot by signal engineers for measuring the ratios of power in signals. In brief:

$$\mathrm{dB} = 10 \log_{10}(P_1/P_2) \ .$$

They measure how much more power there is in one signal compared with another. Sound engineers like them because the human ear can respond to a range of sounds from the quietest to the loudest without getting deafened of about 12 orders of magnitude, that's $10^{12}$ in the ratio of loudest sound power to quietest sound power. By taking logarithms, this huge ratio is reduced to a manageable number. Each power of 10 covers 10 dB so sounds range from 1 to about 120 dB, if the quietest audible sound is used as a reference.

Electrical engineers also use decibels because the range of powers in various electrical signals is also pretty big. They sometimes use 1 watt as the standard power and call the result of comparing their signal with this signal level as its power in dBW.

The signal strength map on the slide shows varying signal strengths from a communication satellite beamed over Europe. A drop in signal strength by 10dB is a drop in signal power by a factor of 10; 20 dB is a drop by a factor of 100, and so on.

If you're rusty on logs, remember that the log of a number is the power of 10 that you need to equal that number. Thus the $\log_{10}$ of 10 is 1 since $10^1$ is 10; the $\log_{10}$ of 100 is 2 since $10^2$ = 100, etc. The log of 71 is 1.851 since $10^{1.851}$ = 71. Every factor of 10 adds 1 to the logs. That's where 'bels' come in. Increasing the power from 100 to 1000 for example changes the $\log_{10}$ from 2 to 3. It's not a big change so to make the numbers bigger, decibels are used and the factor of 10 is included in the definition.

*Sending from GEO and LEO*

**Geo** is the sat person's abbreviation for geostationary satellites. **Leo** is the abbreviation for low Earth orbit, which can be anything less than geostationary.

If you take an aerial further from the Earth then the signal becomes weaker by the inverse square law we've already met. How big an effect does this have? Geostationary satellites are 36,600 km above the equator and hence may well be 40,000 km from the target area. A typical polar orbiting meteorological or Earth observing satellite is at 800 km altitude. That's a big difference in height between these two kinds of satellite. The slide shows that the ratio of signals sent by the same dish at these 2 heights is a factor of 2500, or equivalently 34 dB. Another way of look at the situation is that for every increase in height of a satellite by a factor of 2, the signal power goes down by about 6 dB. This is because $\log_{10}2$ = 3.01 and the inverse square law doubles the effect.

In reality the difference between these two satellite orbits isn't as great as the figure suggests. The geostationary satellite can constantly beam its power to the same area. The LEO satellite is skimming over the Earth in comparison. If it wants to communicate with one particular place, then it either has to turn its aerial as it passes over, not that easy to do and if it involved physically turning a dish that would use fuel, or broadcasting over such a broad sweep that the signal is much weaker than that given by a well directed beam. An LEO satellite 800 km high is only above the horizon for about 7.6 minutes at the most. If it needs to spread its signal over 90º to be in contact with the ground, it looses quite a bit of the advantage of being lower. Remember that the geostationary satellite needs only a beam a few degrees wide.

### *Radar Imaging*

I've spent a while on aerials because they are the key to communicating with satellites and probes. It has distracted us from radar and I want to return to radar because it is the basis of one of the newer remote sensing techniques that is growing in application – radar imaging.

Imaging in IR is done by looking at thermal radiation emitted by the ground or by clouds solely on account of their temperature. Visible waveband imaging is done by looking at the ground as it appears in the illumination of the Sun. If you look at the pictures from meteorological satellites, you'll see very clearly that those taken at night in the visible part of the spectrum are completely black whereas the IR images are just as good at night as by day because the Earth is still emitting IR at night-time since its temperature in K is not much different from that during the day. The oceans hardly differ at all. The land may cool from say 290 K to 275 K, but that's not much.

How big, then is the Earth's natural radiation at microwave frequencies? Can we image the Earth at microwave wavelengths as well as at IR wavelengths? To answer that question you need to look at the blackbody radiation formula that Planck deduced. Actually, the long wavelength radiation from a hot body had been investigated even before Planck by Rayleigh, a famous British physicist. Planck's results for microwave radiation were the same as those of Rayleigh. The energy density of blackbody radiation $\propto T/\lambda^4$. Microwaves may have short wavelengths for a radio wave but they are $10^4$ times longer in wavelength than the IR wavelengths that allow meteorological satellites to take IR cloud pictures at night. $(10^4)^4$ is a huge factor and means in essence that the density of microwave radiation coming naturally from the Earth is very small indeed, too small to use for imaging. A microwave imager would see the Earth as black as night at all times of the day. Even if we could collect all the microwaves over a broad band, we'd not get much energy. Even the microwave illumination by the Sun is not very much. The conclusion is that microwave imaging must be done by active not passive sensors, ones that generate their own microwaves.

### *Atmospheric transmission*

One important issue to look at first is whether the atmosphere is transparent to microwaves. The slide shows that it is, over most of the microwave spectrum. This is one very good reason for pursuing microwave imagery. Even better in a way is that the usual obstructions to clear seeing in the atmosphere such as cloud droplets, fog and dust are all pretty transparent to microwaves, simply on account of the size of the obscuring particles. Particles much smaller than the wavelength of any given radiation have little effect on that radiation. The most commonly used radiation for remote sensing is in the range 1 – 10 cm, large enough to pass around all atmospheric impediments but small enough to be reflected off all obstacles of interest. Images can be made whatever the weather and whatever the time of day or night. There is no way to screen a large area from microwave imaging. The military, to be sure, are very interested in this. So are many remote sensing agencies.

### *Real aperture radar imaging (RAR)*

A radar image is built up from an array of picture elements each of which is a measured distance away. The radar beam measures the distance away. In real aperture radar, the size of the picture elements is determined by the radar beam size, which in turn is determined by the size of the sending aerial. As mentioned, with a suitably chosen wavelength, the radar can obtain images through cloud, fog, haze and atmospheric dust that restricts imaging at visible wavelengths.

Real aperture radar is imaging as you might expect it. It's the kind of imaging that rain radar gives from a single transmitter with a swept beam. The spatial separation across the image is given by the width of the beam. Our familiar expression of $1.22\lambda/d$ gives the width of the microwave beam for a dish aerial. The example on the slide shows a modest resolution. Resolution in distance away is given by the ability to time the return pulse.

*How big a dish for a resolution of 25 m from 800 km height?*

Landscape images with detail shown at 10 km resolution would be next to useless. Christopher Columbus might have appreciated one showing the Caribbean and the outline of North America before he set sail but even in his day, more than 500 years ago, maps of much of Europe were better than this. How big a dish will you need to make a landscape map at 25 m resolution from a polar orbiting satellite 800 km high? The slide shows what the expression $1.22\lambda/d$ tells you for dishes of increasing size using 5 cm radar signals, as is actually done. 25 m is off-scale. A dish 1 km wide will 'only' give a resolution of 50 m. Microwave imaging seemed like a great idea but there's no way that you can launch a dish that needs to be a km in diameter, never minding the problem of making sure it is kept exactly in shape to one mm tolerance or so across this distance. Is that the end of the story?

*A high resolution radar picture*

This slide shows a radar picture with 25 m resolution. Indeed, the section on the right of the slide shows an area imaged to a resolution of 8 m per pixel. How is it done?

*Synthetic aperture radar (SAR)*

The answer is synthetic aperture radar (SAR for short). It's a clever trick. ESA use it, NASA uses it and the very successful Canadian RADARSAT uses it. RADARSAT-1 was launched in 1995 and came out of operation in 2013. RADARSAT-2 was launched in December 2007. I'll use RADARSAT to illustrate the principles.

*Combine together 2 observations*

Synthetic aperture radar sweeps a radar beam sideways and uses the motion of the source on an airplane or satellite to sweep the beam parallel to the track of the craft. As in ordinary radar, the signal is coded so that the time of flight of the reflected signal can be deduced.

The synthetic aperture method combines together the signal from two separate receivers to give the *resolution* appropriate to a large dish but not the signal gathering capacity of a large dish. The whole reason that a big dish gives a better resolution than a small dish is that the signals from either side of the dish add together with their appropriate amplitude and phase in the central detector in the same way that was discussed in earlier slides for our phased array. Therefore if you are going to combine together signals from two separate receivers, then you must record not only the amplitude of the signals but also their phase. It makes a huge difference if the signals are in phase or out-of-phase. You must also combine together the signals received at the very same time, just as they would come together in a big dish. You need, then, to keep a very accurate note of the time signals are recorded.

The time requirement would seem to make aperture synthesis from a single satellite a non-starter. You would seem to need to combine 2 signals from 2 satellites a km or two apart both picking up simultaneously the reflected radar pulses. Ideally you want to receive the reflected signals at A and B at just the times that a large dish would do so. However, **2 satellites are not necessary**. The second part of the synthetic aperture radar trick is to do the whole thing from one satellite that moves between the two points A and B. A satellite 800 km high moves 2 km in 270 ms. It will receive the signal at B over 250 ms (for example) later than a big dish but this time is a precisely known amount and can be allowed for.

In reality, the satellite samples the received signal from a track stretching from A to B in this accurately known time. From this signal it is possible to work out the detail in the ground below to an accuracy comparable to that which would be have been shown by an aerial that stretched from A to B. That is, you can effectively make the correction for the motion from A to B and reconstruct an image that has a resolution detail corresponding to the full aperture from A to B. Reconstructing the image is a task for modern high power digital signal processors.

One example to look forward to as I'm updating these notes is ESA's Biomass Earth Explorer satellite, now set to fly in 2023, with its 12-m diameter radar antenna. The p-band radar will pierce through woodland canopies to perform a global survey of Earth's forests from an altitude of 600 km. The p-band has a long wavelength by radar standards, about 1 m, which is why it readily penetrates leaves, 'seeing' the bulk of trees.

*Sending & receiving*

In summary, the satellite sends down a stream of pulses (shown schematically in blue). This stream is highly directional and is swept across the area you want to image using a *phased array aerial* beneath the satellite. The direction that the beam comes out is varied by varying the phase of the signals sent to the different elements of the aerial. Hopefully you can see how this is done. Between the outgoing pulses, the incoming echoes are picked up, amplified and the signal stored with very accurate timing information.

*Data collection modes*

RADARSAT has 5 modes of data collection. Don't remember the details. The slide shows that different areas can be scanned with different resolutions, namely how much of the ground one pixel represents.

*Information in the image*

Each pixel records the strength of the radar scatter from an area of ground that depends on the *resolution* of the instrument. 25 m is pretty good from space but you can do better. 8 m is the resolution of RADARSAT-2 in fine resolution swath mode. Bright areas are where you get the most reflection, dark areas the least. The brightness depends on

- what is in the pixel area
- how rough or smooth the area is
- the moisture content
- the viewing angle
- the polarisation of the pulses
- the wavelength

Objects much smaller than the wavelength appear dark because they don't scatter much. Objects much larger may reflect a lot, depending on their composition and large scale texture. Roughness is more important than composition. Flat surfaces reflect little. Hence stretches of water, roads and so on appear dark. Vegetation is fairly rough and usually reflects well. Surfaces tilted towards the radar reflect back more strongly than surfaces tilted away. Thus buildings are of variable radar brightness depending on how their large sides are oriented with respect to the radar beam. Some surfaces show black because they are hidden from the radar beam by the surfaces in front. This obviously becomes more important the more sideways on to the landscape that the radar looks. Side views can hide the back-slopes of hills and mountains and even large buildings. The slide picture tries to summarise this.

The electrical properties of the reflecting surface show up in two ways. Water generally reflects well (unless it is a flat surface). Thus snow appears very bright and wet land appears brighter than dry land. Secondly, microwaves of different polarisations reflect differently. The polarisation is the direction of the electric field in the microwave pulse. You get the same polarisations with microwaves as you get with light. Four combinations of linear polarisation are used. HH means horizontally polarised microwaves are sent out and horizontally polarised microwaves are received back. You can also get VV(vertically polarised sent and received), HV and VH configurations. Some systems collect all 4 sets of data. Others, such as RADARSAT, collect just one (HH in this case). Interpreting the results is a specialist activity.

*Features of radar images*

I've mentioned above some aspects of radar images. RADARSAT is in a sun synchronous orbit along the dawn dusk line. It sees the Earth at the same solar time all the time. Choosing the dawn/dusk terminator means that the effects of possible daily variations in the appearance of crops is minimised. (E.g. flowers standing out proud on sunny days but not on cloudy days). It looks down from the same angle, shining its radar beam in the same way. Remember that it is illuminating with its own microwaves an Earth that is essentially dark all the time in the frequency range used by the radar. This constancy of illumination makes interpreting the pictures much easier than say interpreting visual pictures taken under sunlight conditions that vary throughout the day. RADARSAT includes 14 orbits per 24 hour day and can collect data for 28 minutes per orbit, transmitting it directly to the downlink site if that is in view or storing the information in the on-board recorder. The orbit is 'polar', actually inclined at 98.6º to the equator.

Images collected at one wavelength are essential *monochromatic*, i.e. are in one colour – black and white in old-fashioned language. Colour is added in what is called *false colour*, coding for other features. It could be height, it could be land use, it could be to make the picture seem realistic by super-imposing a visual image with less overall detail. Microwaves are more like a giant laser than a giant torch. The difference is that laser light is *coherent* and induces a speckle pattern over the illuminated area. This happens with microwave images, too, resulting in very small scale variations in brightness that are not a real reflection of changing conditions at small scale. This has to be factored out by the image processing routines.

*Example 1 SE Tibet*

Information courtesy NASA. This space-borne radar image covers a rugged mountainous area of southeast Tibet, about 90 km east of the city of Lhasa. In the lower right corner is a wide valley of the Lhasa River, which is populated with Tibetan farmers and yak herders. Mountains in this area reach about 5800 metres above sea level, while the valley floors lie about 4300 metres above sea level. The Lhasa River is part of the Brahmaputra River system, one of the larger rivers in Southeast Asia eroding the Tibetan plateau. The rugged relief in this area reflects the recent erosion of this part of the plateau. Most of the rocks exposed outside of the river valleys are granites, which have a brown-orange colour on the image. In the upper left centre of the image and in a few other patches, there are some older sedimentary and volcanic rocks that appear more bluish in the radar image. Geologists are using radar images like this one to map the distribution of different rock types and try to understand the history of the formation and erosion of the Tibetan plateau. This image was acquired by the Space-borne Imaging Radar-C/X-band Synthetic Aperture Radar (SIR-C/X-SAR) on board the space shuttle Endeavour. North is toward the upper left. The image is 49.8 km by 33.6 km. The colours assigned to the radar frequencies and polarizations are as follows: red is L-band (24 cm), HV; green is C-band (6 cm), HV; and blue is the ratio of C-band to L-band, HH. SIR-C/X-SAR, a joint mission of the German, Italian and United States space agencies, is part of NASA’s Earth Science Enterprise.

*Dublin*

This radar image of Dublin, Ireland, shows how the radar distinguishes between densely populated urban areas and nearby areas that are relatively unsettled. In the centre of the image is the city's natural harbour along the Irish Sea. The pinkish areas in the centre are the densely populated parts of the city and the blue/green areas are the suburbs. The two ends of the Dublin Bay are Howth Point, the circular peninsula near the upper right side of the image, and Dun Laoghaire, the point to the south. The small island just north of Howth is called "Ireland's Eye," and the larger island, near the upper right corner of the image is Lambay Island. The yellow/green mountains in the lower left of the image (south) are the Wicklow Mountains. The large lake in the lower left, nestled within these mountains, is the Poulaphouca Reservoir along River Liffey. The River Liffey, the River Dodden and the Tolka River are the three rivers that flow into Dublin. The straight features west of the city are the Grand Canal and the three rivers are the faint lines above and below these structures. The dark X-shaped feature just to the north of the city is the Dublin International Airport. The image was acquired by the Space-borne Imaging Radar-C/X-band Synthetic Aperture (SIR-C/X-SAR) when it flew aboard the space shuttle Endeavour. The area shown is approximately 55 km by 42 km. The colours are assigned to different frequencies and polarizations of the radar as follows: Red is L-band (24 cm) HH; green is L-band, VV; and blue is C-band (6 cm) VV. SIR-C/X-SAR, a joint mission of the German, Italian, and the United States space agencies, is part of NASA's Mission to Planet Earth.

Another synthetic aperture radar imaging satellite that was in the news was ESA's CryoSat-2, a satellite specifically targeted at measuring sea ice thickness in the polar regions and changes in the height of the Greenland and Antarctic ice sheets. These issues are key factors indicative of the rate of global warming as well as indicators of environmental changes in the Earth's polar regions. CryoSat was intended to be operational during International Polar Year (2008) but the original satellite was destroyed in a launch failure and a replacement was launched in April 2010. The mission has local interested too, for the Department of Geography at the University of Aberdeen were involved in the development project. As of 2015, Cryostat-2 is still operating successfully.

*Communications*

The place to make money in the space business is in communications. Customers are counted in the hundreds of millions. Satellites are used for TV dissemination to the public, for TV network companies to acquire their wide choice of channels, for mobile phone links, for sat phone linkage and of course for the internet communications including the world-wide-web. With all this activity it's not surprising that that satellite communications is a highly advanced technology. Other satellite applications such as remote sensing, which is carried on by government agencies (direct and indirect) and increasingly by private sector companies, don't have to invent this technology. I want to give an idea of what is involved in satellite communications, since we are all affected.

*The digital age*

When you hear the phrase that 'we live in a digital age', just what exactly is implied? The sharp end of remote sensing is a device that produces an electrical signal when some physical quantity we want to measure changes. It could be temperature, light, magnetic field or a host of quantities. Very few of these devices are intrinsically digital. They all produce a signal that changes in proportion to the change in the quantity being measured. This behaviour is called analogue. For example, if the temperature increases by 3.59% of the range covered by the sensor, then the signal increases by 3.59%. There is nothing digital about that. The digital nature of the electronics has to be forced upon the circuitry by what is known as an *analogue-to-digital converter* or ADC as it's universally known.

Digitisation means dividing the signal into discrete steps. This division is done both in time (for time varying signals) and in amplitude. It may look as if some of the original information is thrown away in the process. You'd be right – it is. However, in practice we can afford to lose what is thrown away, which concerns very fine changes in the signal, and the price paid has advantages that are worth it, as we'll see. The net result of this sampling is to associate each signal level with a binary number, one that can be represented by a moderate number of binary digits. 8 binary digits can represent 256 different signal levels and an 8-bit binary number has a special place in modern computer technology, giving it the name of a *byte*, as I'm sure everyone here knows. The slide shows 1 byte and its interpretation as a simple integer. Bytes can be coded to represent numbers in other ways but the details would take us away from our subject. 12-bit numbers represent 4096 signal levels and this is more than enough for many applications.

*ADCs*

We've already seen what ADCs do. They usually come on chips with surrounding circuitry to activate them correctly. A reasonably decent ADC will sample a varying signal and convert it into a 12-bit number in 10μs, which is pretty good going. That's $10^5$ conversions per second. If the last bit of the conversion is accurate, that means accuracy to about one part in 4000 of the range of the ADC, which is 0.025%. That's also impressive. Not many measurements in the pre-digital age were made to this accuracy. The slide shows a commercial ADC board in the background.

*Pictures are divided into pixels*

Pictures are digitised too, first dividing them spatially into pixels and then digitising both the colour information and the grey levels. CCD detectors have a surface structure that is intrinsically pixellated, being divided into a geometric pattern of sensing elements. The sensitive area of the CCD shown is about 7 mm by 9 mm. Some CCDs have a pattern of sensors for separate pixels sensitive to red, green and blue, such as the one shown diagrammatically. CCDs also have on-board circuitry that delivers the signal directly in digital form and hence they are a natural choice for visible imaging applications. Those taking our *Light Science* course will hear a little about how CCD detectors work.

The slide shows one way that grey levels can be coded into 256 steps from black to white. 256 steps is pretty well all you need in practice. Your eye can't really tell the difference between 2 neighbouring steps. Again, see the slide for an example chosen from the middle of the range, where you might expect it would be easier to notice changes.

Why is digital better than analogue? The key concept that makes digitisation worthwhile is *noise immunity*. Anyone who has heard a vinyl record playing will recognise the background hiss that used to accompany sound reproduction. For the first century of the telephone system, conversations were always accompanied by background hiss. This hiss is electronic noise. People got so used to it that it wasn't the irritation it would now be if similar noise were to suddenly appear. Radio programs could equally be accompanied by a background hiss, especially if the aerial was short and the signal was a bit weak. Noise and analogue signals go together. Nowadays, you put on the CD and the background is velvet black, not a hiss or a crackle or a rumble. You talk on the phone in Aberdeen and you can't tell if the speaker is outside your front door or in the USA. Noise immunity is the strength of digital signalling. How is it done?

*TTL signals*

At the risk of getting technical, I'll put in a slide showing how. Everyone with a science degree should know about the basic idea.

TTL is the logic employed in a great many digital circuits. It works by assigning a range of voltages to each of the two digital levels. In most applications, digital "1" is assigned from 2 to 5 V. Digital "0" from 0 to 0.8V. TTL devices are commonly powered from 5V sources but you can see that 3V sources will do the job. You can see how noise immunity works. Provided the electronic noise is modest, variations of signal voltage don't cause the device to change its state from a "1" to a "0" or vice-versa. Digital devices aren't completely immune from noise but the proof of the pudding, as the saying goes, is there for all to hear in modern sound reproduction and in modern digital copying of sound, pictures and textual information.

*What happens if the noise is too big?*

The digital signal is corrupted. The most likely place for this to happen is when the signal is in transit, for then the signal is more likely to ‘pick up’ noise, either by induction in the wires involved in sending the signal or because a radio detector can simultaneously pick up other signals along with the intended one. If a signal is corrupted, then a “1” is received as a “0”, or vice-versa. The result may a wrong character in a text, a wrong data value or a wrong pixel in a picture. Wrong characters do not just appear as spelling errors. Individual letters in a message are sent as 8-bit digital words and there are 256 possible combinations of 8 bits in an 8-bit word. There are only 26 letters of the English alphabet, 52 if you count upper and lower case, and hence the chances are that a digital bit mis-received is going to result in an 8-bit word that doesn’t represent a letter at all. It may appear as a non-alphabetic symbol or perhaps even a non-printable character. Clearly it’s important to spot errors when they occur because they mess up the message.

To spot an error you need to transmit some kind of extra information that allows you to make a check on the accuracy of the result. The simplest strategy involves the *parity* of the digital word, or perhaps a *packet* of words sent. Parity checks allow single bit transmission errors to be spotted, as the next slide shows.

*Parity*

The parity of a digital word, or set of words, is simply a count of the number of “1”s it contains. If it contains an even number of “1”s, then it has *even parity*. If it contains an odd number of “1”s, then it is *odd parity*. The slide shows examples.

The parity idea is used to help error checking by adding a *parity bit* to the information sent. Sometimes this bit is the last bit, sometimes the first bit. If you are sending 8-bit words with parity, then the expanded words will be 9 bits long. Obviously both sender and receiver must know whether the parity bit is the first bit or the last bit in each word, so that the appropriate 8-bit value can be identified. It is usually the last bit. Thus 001100110 is an 8-bit word (00110011) with even parity, sent as a 9-bit word with the last bit as the parity bit. The final ‘0’ tells you that the 8 bits sent have even parity. If the 8 bits sent had odd parity, then the final bit would be a ‘1’.

All that parity checks can do is spot that one of the bits reaching the receiver is wrong. (Strictly speaking, it can spot that an odd number of bits are wrong). The receiver then has to have a strategy for dealing with this, the most obvious being to request a re-send.

*Long distance signals are noisy*

Long distance signals, such as from space probes and planetary landers, are often weak by the time they reach the Earth and conspicuously noisy. The sending craft is too far away to request a re-send each time a wrong word is received. You need a technique that not only detects the presence of errors but allows them to be corrected. The simplest technique you can think of is possibly to send every message 3 times and hope that at any one point in the message at least 2 of the versions agree with each other. This does work but is not very efficient, obviously slowing down the overall transmission rate by a factor of 3 and sending a vast amount of repeated information. Can we do better?

*What is the problem?*

The problem is a corrupt word being interpreted as another valid word. What you need is distance between valid symbols so that if an error occurs then you can still associate the erroneous character with the intended symbol. What is meant by *distance* in this context? It was Richard Hamming, a man well known for his contributions to digital signal processing, who first defined the relevant distance. The distance between two digital words (or strings of words) is simply the number of bits that differ between them. The slide shows an example. Hamming went on to show that if the distance between all code words in a message is *n*, then you can detect **and correct** all words that arrive with (*n*-1)/2 errors. The minimum Hamming distance, as it is called, to detect and correct one error is clearly 3. The next slide shows how this works.

*How can error correction work?*

4-bit words can code integers from 0 to 15. However if a single bit is wrongly received, then the erroneous word will be interpreted as the wrong number. This is just what we don't want.

Now add 3 more bits onto each word as in the list shown on the slide. The result is a code in which each code word differs by at least 3 bits from each other code word. Now, if you send the code for the digit 3 as 0011010 but it arrives in error with a wrong bit, such as 0111010, then that wrong word is still nearer 3 than any other number in the set of 16 numbers and hence is taken as 3. Without the extra bits, the erroneous character would have been interpreted as a 7. The error correction has been much more efficient than sending the digit 3 times and taking the result if two versions agree. This technique isn't just used in exotic circumstances of sending information collected by Mars rovers but is used in everyday life. See the next slide.

*Examples for you to investigate*

I'm leaving you to investigate some of these if you are interested. (There won't be exam questions on them). Libraries, booksellers and others use the ISBN number to specify books instead of tediously typing out the book title, author and so on. Don't they get a lot of faulty information and perhaps errors in ordering books from afar, because it just needs one digit of the ISBN to be wrong before the wrong book is referred to? Well, "no", it doesn't often go wrong because the ISBN code has built in error correction. The receiving check can even determine if two digits have been typed in the wrong order, which is easily done. Parity alone couldn't tell this.

The Universal Product Code bar-coded on every single item a supermarket sells has to be accurately transmitted from the point-of-sales till to the accounting and stock-taking computer. Have you ever wondered how the bar-code is related to the digits shown? Well, each digit is represented by a 7-bit number. An even greater sophistication is that the 7 digit combination differs for a digit in the first half of the UPC and the second half. This allows the reader to know if the bar-code has been scanned backwards or not. There is also an internal parity check on each digit. You can read more about it on the web.

*Some error correction strategies*

*Forward Error Correction*. This is the one we've just been talking about. Extra information is sent with the signal that allows errors to be detected and corrected. There is a limit to what can be done. The more errors you want corrected the more extra information you must send. You therefore need to know the characteristics of your transmission channel. For example, the sort of errors produced by scratches on a CD are different in nature to the errors you get on a weak signal transmitted by radio. A different forward error correction strategy is employed in these two circumstances.

*Automatic repeat request*. The signal is divided into packets and when a packet arrives with an error, a request for a repeat send is initiated. The sending device must have some way of responding to this and knowing what it has sent without 'forgetting' the message once it has been despatched. This is just the strategy we use in everyday oral communication. You listen to someone speaking but when the words don't make sense because of background interruption, poor articulation or simply that your attention has wandered, you say 'pardon' or 'please say that again' or words to that effect.

*Stop and wait*. Acknowledgement of accurate receipt is needed before the next packet is sent. This kind of communication can commonly be found in administrative systems. In a multi-stage process, each form has to be received and acknowledged before the next form can be processed. The University's finance system is a good example of such a system.

*The communications limit*

There is a very important law of nature that isn't explained in any other physics course we teach. It's not an obscure law at all, but one that is central to the way society operates and to the information technology age we are now in. It is a law that limits how quickly information can be sent from one place to another. That information can be electronic, the obvious way of encoding information these days, or embodied in speech, light, ultrasound or any other physical medium. The law applies to every form of communication and was discovered by Claude Shannon, who was truly the founding father of information theory. It is a law that will hold whatever technology is used and hence a law that will govern not only our present day communications but that of our descendants far into the future. The law is known as **Shannon's law**, or sometimes Shannon's noisy channel law because Shannon also discovered another basic truth about information, namely how to define and measure it. That is another story, for you to read about in your own time. Shannon's noisy channel law is the one I want to talk about.

*What determines communication rates?*

If I talk rather quietly, some in the audience will begin to miss words. The reason is simple. Background noise from outside, or even from within your head, will obliterate enough of a word to make you miss the meaning. The effect is worse the faster I speak. If I slow down and speak quietly but very deliberately, you will probably pick out every word. Another acoustic example of where it can be hard to hear the words is in a vaulted railway station. The ceiling and platforms transmit quite a narrow range of fairly low frequencies very well. The bandwidth of the communications channel is the range of frequencies that it transmits. In a large railway station the bandwidth is narrow and this contributes to the difficulty of error-free transmission of speech.

These examples hint at the two effects that limit the ability to transmit information from one place to another. One effect is **bandwidth**, which is simply the range of frequencies available in the communications channel. The wider the bandwidth the more information per second can be transmitted. For example, a 625 line TV picture of the kind that was common in Britain required about 5 MHz of bandwidth to be transmitted. (HD TV has even more lines). You can never transmit such a signal on the medium waveband because that waveband runs from about 500 kHz to 1.5 MHz in round numbers, a bandwidth range of 1 MHz. The entire band is insufficient for a single TV picture, even though it is good for some 100 audio channels each occupying 10 kHz.

The second limitation is **signal-to-noise ratio**. This is measured as the ratio of the power of the received signal to the power of the noise at the detector. The noise is what's there when the signal isn't, and you can't get completely noise-free channels. It is this effect, for example, that limits the Mars Odyssey orbiter's communication rates to Earth to 128 K bits per second, slower than your home broadband connection. Noise always limits the speed at which you can transmit information. Shannon's law tells us by how much, as the next slide shows.

*Shannon's law*

The law tells us the maximum capacity, C, of an information channel in bits per second. Shannon's law is a natural limit. Many communications systems only attempt to transmit information much below this limit.

$$C = B \times \log_2(1+ S/N).$$

To recap:

- C is the communications channel capacity in bits $s^{-1}$
- B is the bandwidth in Hz
- S/N is the signal-to-noise ratio, a dimensionless number

Shannon's arguments that led him to this law don't tell how you can code a signal to reach this limit. What they say is that if you are ingenious enough you can reduce the transmitted errors in the message to as low as you want, provided you transmit the message at less than the limiting rate C. If you transmit it faster than C, then an unlimited number of errors may build up. The slide shows a couple of examples. A standard telephone transmission system used for much of the 20th century could transmit a bandwidth of no more than 5 kHz. It was developed and used to transmit voices and the resulting modest quality reproduction was lucky if it succeeded in providing a frequency range of 5 kHz. However, supposing the signal-to-noise ratio was 100 (20 dB in the more usual units used in the business), you can see that the Shannon limit for such a line is actually 33.3 kbit $s^{-1}$. How can you achieve that? I'll give a clue in the next section when I talk about modulating the signal. If engineers had the technology to approach the Shannon limit, then they could make a fair go at transmitting concert-hall quality sound down a telephone line but they aren't going to manage a full-sized TV video picture, for that comfortably exceeds the Shannon limit.

The reason that you can now get much more down a phone line than ever before is that bandwidth limiting equipment has been stripped out of telephone exchanges, leaving the bandwidth limitation provided by the transmission lines themselves. That limit is much higher than 5 kHz and hence down the very same wires that I was struggling to receive digital signals at home at 2400 bits $s^{-1}$ 30 years ago, we now get 1 Mbit $s^{-1}$, and more with luck.

The final example shows how even with a lot of noise on a line, provided you have suitable error detection and correction codes then you can recover a clean signal with all the original information if it is sent at a moderate rate.

I should add that logs to the base 2 appear naturally in Shannon's law because the law is about bits per second and bits are binary digits, with just two possibilities 0 and 1. You won't find logs to the base 2 on your calculator but they are related to logs to any other base by a constant multiplier. The slide shows that they are related to logs to the base 10 by a multiplicative factor of 3.322. This factor is just $1/(\log_{10}2)$, as you can check on your calculator.

*Pushing the limit*

The take-home message from this slide is that Shannon's law tells people what to aim for. If your technology can get near the limit, then you can receive mobile phone-calls further from the mast because the error correction is more tolerant of noise. The phone companies need erect fewer masts, which is good for them and good for us, or they may use less power, which will cheapen the cost of calls and that's good for everyone too. The systems in use today have a much larger signal to noise ratio to transmit their information than Shannon's law says is necessary. Likewise if we can get closer to the Shannon limit, wireless networks will work with less power over a larger range. Spacecraft will be able to make do with less power, so they will have longer lifetimes or be able to transmit from further away. There are huge commercial advantages for pushing the technology as close to the Shannon limit as possible.

*Approaching the limit*

The final part of this story is that new techniques in the field are solving the bit of the puzzle that Shannon left unanswered. How can error correcting codes be designed and implemented that will allow information transmission at rates close to the Shannon limit? Today's buzz words are LDPC – **low density parity check** coding. The technique is an elaboration of the basic idea of parity checking, which you've met not long ago. The simple parity check worked out the sum of all the digits in a packet and compared it with the expected sum. The elaboration requires coding the sums of several different selections of bits in the packet. It is pretty easy to encode. The really hard bit is deducing which bit is in error when the packet arrives corrupted. To do so turns out to take a lot of computing power that wasn't available in earlier decades. It now is. The development of ever faster microprocessors may have been substantially developed by competition to keep selling customers new computers but the new technology is now finding quite different uses. Multi-processor chips are succeeding in decoding LDPC signals in 'real time', as the messages arrive at high speed. That is what is needed for the mass market of mobile phones, wifi and other areas. Space technology is in the forefront of these developments because the commercial gains from improved data transfer rates are enormous for satellite-based systems. 'Watch this space', or rather read in the scientific press about developments that are bound to happen in the coming years.

*Communications are important*

Why am I talking so much about the detail of communications? Simply because it's central to space science in general, remote sensing in particular and of course daily life in the 21st century. A 200 million dollar probe or satellite with 10 instruments on board all fully functioning is worse than scrap if the comms link is down. At least you can get some money for scrap but a dead probe or satellite is at best worth nothing and at worst may even be a liability if it could return to Earth in an uncontrolled manner. Looking at the two probes on the slide, the size of the aerials gives away the fact that communications is a very big factor in space-probe design. The New Horizons mission was launched in early 2006 on a path for Pluto and the Kuiper belt beyond. The resulting images that have been sent back now it has passed Pluto are stunning, particular so for being acquired and sent back from over 4 billion km away. 'Remote sensing' at its most remote. The probe is almost a once-in-our-lifetime shot because Pluto will move significantly out of the plane of the rest of the planets in the solar system in coming decades. To get the benefit of slingshots, a probe's orbit must be in the plane of the solar system to meet the assisting planet.

*[Aside*

By way of introduction to the final lecture I'll say that when I offered to give lectures on remote sensing, I'd no idea that I'd spend much of the time talking about communications. I'd imagined covering a series of more specific topics such as how satellites measure ozone levels, sea heights, the vegetation index and other interesting things. Some of the class would probably have preferred that, especially if my choice of subjects happened to correspond with their interests. On reflection, however, the topic of 'communications' is not only at the heart of all remote sensing but it underpins the electronic age of the 21st century, from applications like mobile phones that are now mundane to the latest n-th generation games consoles that use 'remote sensing' across a room. Some appreciation of 'communications' will demystify jargon used not only by the remote sensing community but now found in many walks of life. I'll begin the final lecture talking about how information is impressed on radio signals.]

*Transmitting the signal*

So, you have a signal you want to send, or perhaps receive. Morse code signalling is binary signalling. There's either a signal or not a signal. Signals used to be sent along a telegraph line by switching the current on and off. Since the advent of radio, signals are not sent by switching the current on and off. They are sent on *a carrier*. Signals may even be sent on a carrier over a copper wire. The information to be sent is impressed on the carrier by a process called *modulation*. The signal is extracted from the carrier by the process of *demodulation*. A *modem* is a modulator for the uplink and demodulator for the downlink.

*The carrier wave*

A carrier wave consists of a sine wave of definite frequency. The slide shows 50 periods but the carrier keeps on going for as long as the signal is transmitted. The carrier itself contains no information, other than that the transmitter is switched on.

Modulation involves changing one of the parameters of the carrier wave slowly in comparison to the frequency of the wave.

*Amplitude modulation*

*Amplitude modulation* (AM) is the easiest form of modulation to describe. It's the one used on medium-wave broadcast signals. It's very simple. The amplitude of the carrier is changed in proportion to the signal you want to send. The example shows an analogue signal providing the modulation.

*Digital amplitude modulation*

This slide shows a digital signal modulating a carrier. It's very obvious how the signal contains the information.

The reason that amplitude modulation isn't used on high frequency signals is that it's quite susceptible to interference. If the carrier fades or receives some interference then the change in its amplitude is interpreted as a change in the transmitted information, which it isn't. The result is 'noise' of some kind.

*Digital frequency modulation*

The method of modulation that took over from AM is *frequency modulation* (FM). In this scheme the frequency is changed according to the size of the signal to be sent. A digital signal can be sent with just two frequencies, what used to be known as *frequency shift keying* or FSK. The slide shows the appearance of a carrier frequency modulated with a digital signal. In practice the shift in frequency is quite small and you would hardly see the difference on a trace. FM is used quite a lot because it is less susceptible than AM to the effects of interference such as fading. The amplitude of the signal contains no information so it doesn't matter if it changes.

*Phase modulation*

This story is leading on to the favoured technique for digital communication: *phase modulation*. In this situation, neither the frequency nor the amplitude of the signal changes. In the simple case a "1" is represented by one phase of the signal and a "0" is represented by another phase. The obvious other phase to choose is one that is 180º out-of-phase, such as is shown on the slide. There is a jump in the phase of the signal every time the digital signal changes.

*QPSK - Quadrature phase shift keying*

*QPSK* is the technique used in many digital communications networks, such as the one that connects your PC to the network when you login from home. In this technique 2 bits are sent at a time, called a *dibit*. 4 possible phases are used, as you can see from the slide. The signal keeps jumping in phase. You need to keep the carrier for longer on each phase than I've shown on the slide but I hope the idea is clear. The signal jumps in phase every time a different pair of bits is sent.

*Demodulation*

Demodulation is the process of extraction of the information from the carrier. Demodulation of QPSK sounds difficult but in fact it's not. For those who'd like to see how it's done I'll show you on the next slide.

*Demodulating QPSK*

The incoming signal of frequency ω, with its phase shift of ϕ, is multiplied by sinωt, a steady signal of the same frequency generated within the demodulating circuitry. The result is a signal with two components. One component is at the same frequency as the incoming signal and is proportional to cosϕ and hence contains information about the phase of the incoming signal. A second component of the product appears at twice the carrier frequency and this is thrown away, or "filtered out" in the jargon. Another multiplication of the original signal is performed at the same time but using the multiplying signal cosωt. This produces a second output at the same frequency that in this case is proportional to sinϕ. Now it's a well known mathematical result that if you know both sinϕ and cosϕ then you can find ϕ. Hence you can find the 2 bits that have been sent. This may sound a bit elaborate but it can be done at great speed by modern electronics.

*Aside on Broadband*

ADSL – *asymmetric digital subscriber line*. This uses QPSK or variations of it. As you'll see from the adverts that are sometimes on TV these days, public access ADSL runs to typically 2 Mbits per second (not megabytes per second) on standard copper wire. The asymmetric part in the name is because the downlink is faster than the uplink. There is built-in forward error correction and you now know what this means!

*A final word or two on graphics*

I would like to come back to the images that all this communication is often about. On a full-colour screen each pixel is represented by the intensity of 3 colour components, red, green and blue. These are all you need in varying proportions to produce any colour of choice, as has been discussed in our *Light Science* course. This means that each pixel is represented by 3 bytes or 24 bits, since 1 byte is used for each colour. The next 2 slides show an image gradually magnified.

The image is a jpeg satellite photograph of an area in Assynt in NW Scotland – the Coigach peninsular just north of the Summer Isles. Ullapool is just to the south and Lochinver just to the North. The actual values of the colour pixels in the range 0 to 255 are shown on the last two magnifications.

*Jpeg compression*

A full screen picture on typical computer display takes over 2 Mbytes, remembering that it takes 3 bytes for each pixel. The satellite photo of NW Scotland shown 2 slides back was stored as a jpeg file taking only 145 kbytes and you could hardly tell the difference between that and the original. Jpeg is a compression system that is good for transmitting pictures of the photographic kind. The jpeg file is smaller than the original and therefore transmits more quickly. Jpeg was developed by the **J**oint **P**hotographic **E**xperts **G**roup, an international body, as a standard for compressing still pictures. The development for moving pictures is known as Mpeg.

The compression is based on exploiting the properties of the human visual system, which is sensitive to brightness levels but not nearly so sensitive to colour changes. The jpeg algorithm, namely the recipe for the compression, is *lossy* in the sense that some information is indeed lost in the compression. Keep repeating the jpeg process on a single picture and you will gradually lose more and more of the detail. When making a jpeg version, there is a user selectable compression factor, usually expressed as an integer from 1 to 10 (or sometimes 1 to 100), that you can choose. 2 is very highly compressed and loses lots of information but creates a very small file. It's good for making images for previewing or indexing, or such like. Integers 8 and 9 on the 1 to 10 scale ensure that you can hardly tell the difference between the reconstructed jpeg file and the original. I usually use these options and still find the compression can be in the range ×10 to ×20, meaning the compressed file is 10 or 20 times smaller than the original. These days, jpeg format is the default storage for digital cameras.

Because of the loss of fine detail, jpeg doesn't preserve very sharp contrast edges. For example small print on a picture isn't preserved as sharply as you'd like. Because averaging is involved in the jpeg algorithm, two pixels in the original that are the same colour may not appear the same colour in the re-constructed jpeg image. Never use jpeg for black and white drawings. Jpeg needs at least 16 grey-scale levels and will not only spoil the simple black or white nature of such drawings but will also introduce some loss. Gif, on the other hand, will compress such drawings without any loss and preserve the 2-level nature of the drawings. Finally, the jpeg compression system does **not** specify a unique file format and some programs will not be able to open jpeg files produced by other programs. This is a pain to users!

*Gif compression*

Gif stands for **G**raphics **I**nterchange **F**ormat. Gif is a lossless technique for files that have one byte per pixel. Grey-scale files with 256 levels are the obvious good example where Gif compression reduces the files size without any loss of information. Colour files are reduced to a 256 palette. This means that Gif files are not good for photographs and such images. They are, though, good for block diagrams, logos and all sorts of schematic diagrams where 256 colours is perfectly adequate. The fact that Gif compression is lossless ensures that sharp edges are preserved.

Gif compression is based on the fact that many images contain repeating sets of information. It could simply be blocks of the same colour, wide lines, and so on. Because of the need to keep all the information in the original, Gif compression isn't as compact as jpeg, though the starting files with only 8 bits per pixel are one third of the size. Two additional advantages of Gif compression is that it can handle transparent backgrounds (basically by reserving one of the 256 levels for a character that is interpreted as a transparent pixel) and by piecing together the image area from overlaid patterns, animated Gifs can be generated. Finally, Gif does specify a file format and hence Gif files written by different software should be mutually intelligible.

Gif was originally patented (the patents have now expired) and to avoid the risk of prosecution the patent-free variant png (**P**ortable **N**etwork **G**raphics) was developed as an alternative lossless compression regime. Png files code for a greater range of colours, but don't support animated images.

*Digital signal processing*

Once you have your signal in digital form, then a huge range of powerful processing techniques known as *digital signal processing* (commonly shortened to DSP) become available. They can be applied to all types of audio, static and moving pictures, raw data and so on. Over the next years and decades, DSP is set to become much bigger than it already is. Custom chips that perform DSP in specific circumstances are already a significant fraction of the whole chip market. Every mobile phone includes DSP chips for cleaning up the electronically noisy signal received.

DSP can significantly enhance what you can do and it's relatively cheap, judged by the expenses typical of space science budgets. It's therefore cost effective, which is a major driver for investment.

In the context of diagrams, digital signal processing becomes *image processing*. One specialised but widely used form of digital signal processing is the GPU – graphics processing unit – that can parallel process large numbers of pixels simultaneously, unlike the normal CPU – central processing unit – in desktop computers that processes instructions serially. The final slide shows two simple examples of image processing acting on the satellite picture that was used a few slides back. A simple inverse colour transformation highlights the hills and shows some details more clearly than on the original, such as the A835 snaking up the valley to Elphin. An edge highlight transformation picks out changes in the landscape, in this case highlighting the boundaries of the surface waters and the topographic features. Those taking courses in Geography and Geology will encounter much more sophisticated digital image manipulation to bring out of the picture features of land and land use that it is even quite hard to see by eye. What I wanted to show in this course is that there is a lot of science behind the techniques of acquiring such pictures but that once acquired, the digital format has huge potential that a photograph on a piece of paper does not have.

*Course summary*

This course has taken a wide view of space science with a particular eye on remote sensing as one application. Many of the subjects in this course are not dealt with in any other BSc courses at Aberdeen, yet they are very relevant to 21st century life. I hope you'll find that the course is relevant to your own future.

I've put on the course web page a detailed course summary that should help with revision. Good luck with the revision and future courses!

*JSR*