# *Modulation*

Signal communication, be it audio, video, data, garage door openers, whatever, generally involves some form of modulation. AM (amplitude modulation) radio is perhaps the oldest and simplest example. Begin with a sinewave,

$$V(t) = A\sin(\omega_C t)$$

where $\omega_C = 2\pi f_C$ called the *carrier frequency*. For AM radio in the USA, $0.52 \text{ MHz} < f_C < 1.71 \text{ MHz}$. The carrier doesn't convey any information. Instead, as the name implies, it carries the information in the form of modulation. For AM, the information is encoded in a time-varying amplitude A(t). So,

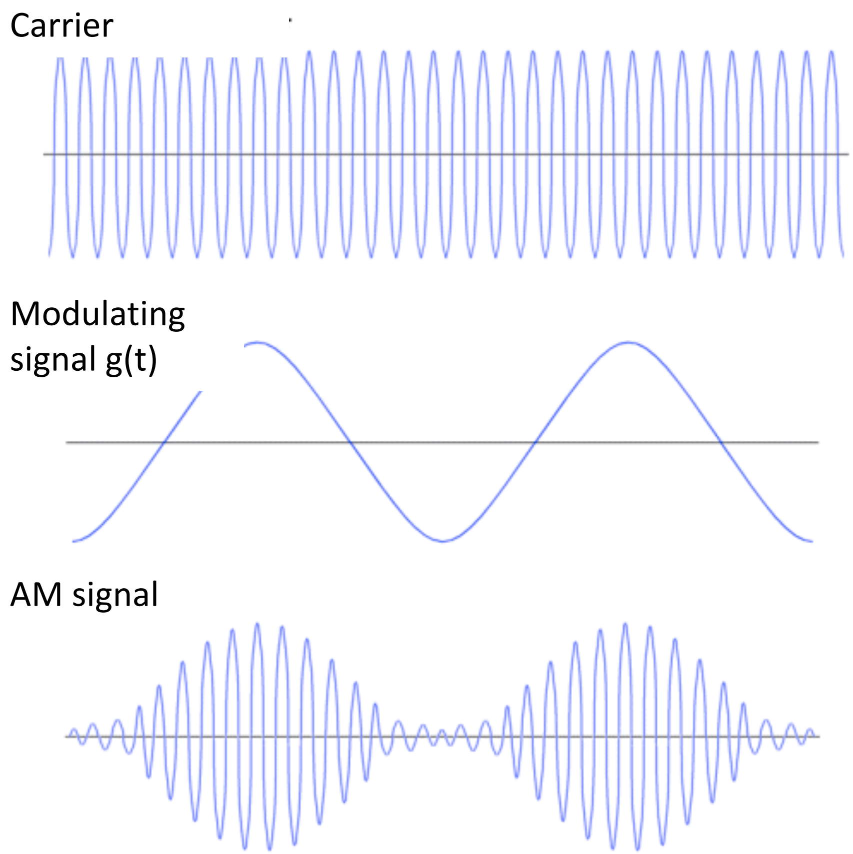
$$V_{AM}(t) = A(t)\sin(\omega_C t)$$

FM, on the other hand, operates at higher carrier frequencies: $88 \text{ MHz} < f_C < 108 \text{ MHz}$. FM information is contained in a time-varying phase:

$$V_{FM}(t) = A\sin(\omega_C t + \phi(t))$$

Why do we need a carrier at all? Why not just transmit electromagnetic waves at audio frequencies? First, we need to separate different stations so that everyone has an identifiable frequency. Second, low frequencies are difficult to transmit. Audio frequencies range from 10 Hz to 20 kHz. If we tried to transmit electromagnetic waves as these frequencies the wavelengths would range from about $\lambda = c/f = 10^4 - 10^8$ meters. For an antenna to efficiently transmit or receive a signals $\lambda/2$ in size, which is obviously completely impractical for audio frequencies. Instead, the information rides on a carrier with much higher frequency and therefore shorter wavelength so reasonable sized antennas can be used.

# *Amplitude Modulation*




*https://www.technologyuk.net/telecommunications/telecom-principles/amplitude-modulation.shtml*

The basic idea of AM is shown here. The information to be transmitted is some modulating function g(t). For simplicity, think of it as a single tone at frequency $f_{\text{mod}}$ and fixed amplitude a. Then,

$$g(t) = a\cos\omega_{\text{mod}}t$$

The actual modulated signal takes the form,

$$V_{AM}(t) = \big(A + g(t)\big)\cos\omega_C t$$

In order to eventually demodulate this signal we must choose A > a so the factor in parenthesis is always positive. An alternative way to write it is,

$$V_{AM}(t) = A\,(1 + m\cos\omega_{mod}t)\cos\omega_C t$$

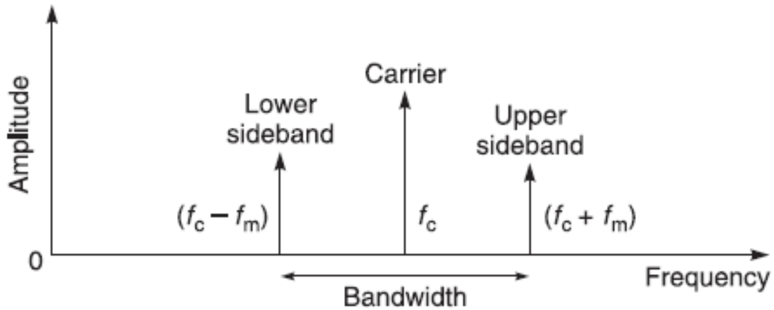
where 0 < m < 1 to ensure that the factor multiplying the carrier is always positive. m is called the *modulation index*. Using trig identities we have,

$$V_{AM}(t) = \cos\omega_C t + \frac{m}{2}\cos\left(\omega_C - \omega_{\text{mod}}\right)t + \frac{m}{2}\cos\left(\omega_C + \omega_{\text{mod}}\right)t$$

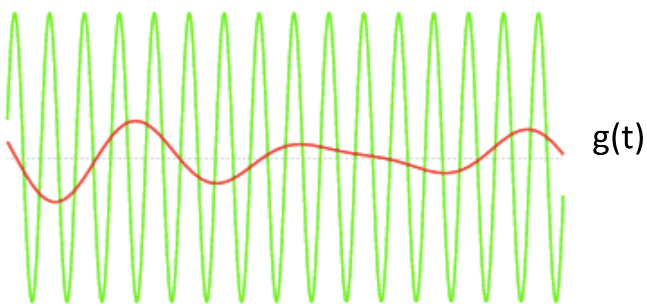
This AM signal contains the carrier frequency and two *sidebands*,

$$\omega_C\ ,\ \omega_C + \omega_{\text{mod}}\ ,\ \omega_C - \omega_{\text{mod}}$$

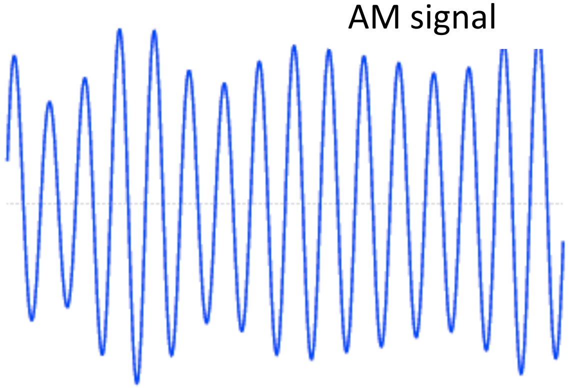
The AM spectrum for single tone modulation is shown to the right. A more realistic modulation signal as shown below contains a continuous range of modulation frequencies and amplitudes so the the spectrum would look continuous.

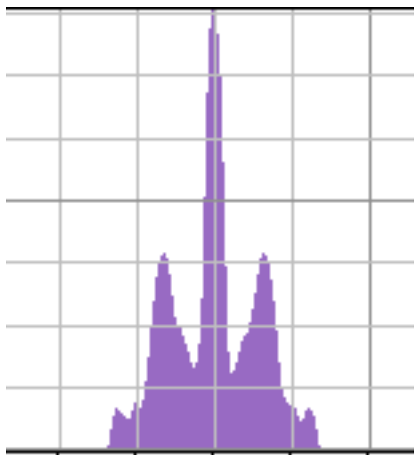



Carrier




AM signal





Spectrum a typical voice signal

*https://epxx.co/artigos/ammodulation.html*

The AM spectra shown above are symmetrical about the carrier frequency. Such signals are called double sideband (DSB) AM. However, this scheme wastes power since (1) half of the spectrum conveys no new information and (2) the carrier contains no information. This is important if you are spending money to generate kilowatts's of power for a commercial radio station. It is possible to generate AM signals with only one sideband and no carrier, known as *single sideband, suppressed carrier* (SSBSC) AM.

## *Phase and frequency modulation*

This time, the signal g(t) carrying the information will be used to modulate the *phase* of the carrier rather than its amplitude. There are two ways this is generally done:

$$V_{PM}(t) = \mathrm{A}\cos\big(\omega_C t + k_{PM}\ g(t)\big) \qquad (Phase\ modulation)$$

$$V_{FM}(t) = \mathrm{A}\cos\left(\omega_C t + k_{FM}\int g(t)\ dt\right) \qquad (Frequency\ modulation)$$

$k_{PM}$ and $k_{FM}$ are set by the circuitry that generates the modulated signal. In PM, the total phase is,

$$\theta(t) = \omega_C t + k_{PM}\ g(t)$$

The modulating function g(t) is proportional to deviations of $\theta(t)$ away from $\omega_C t$. That is, g(t) is proportional to phase deviations.

In FM, the modulating function is proportional to deviations in *frequency* away from $\omega_C$ . To see this, remember that the frequency is the tie derivative of the instantaneous phase,

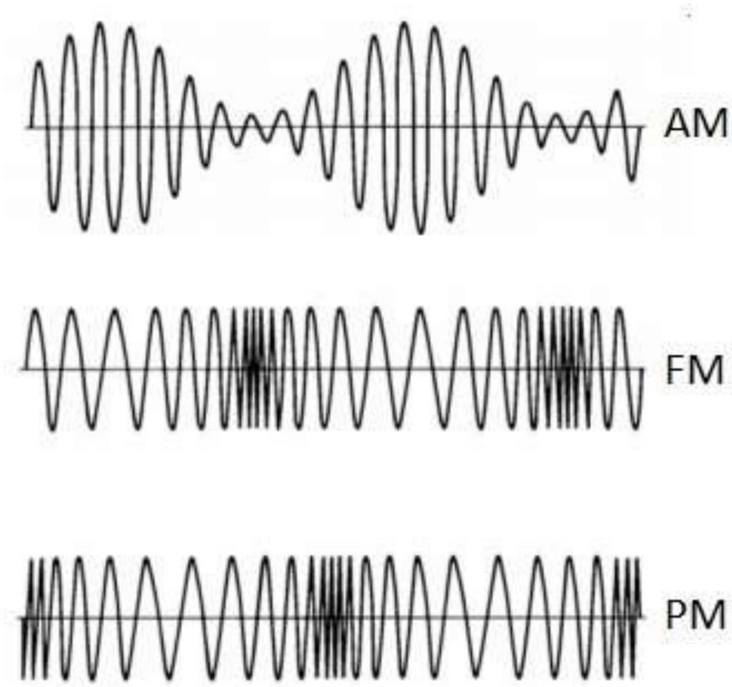
$$\omega(t) = \frac{d\theta(t)}{dt} = \omega_C + k_{FM}\ g(t)$$

for simplicity, let *g*(t) be a single tone modulation ;

$$g(t) = a\cos\omega_{mod}t.$$

For this case, the two modulated waveforms are,

$$V_{PM}(t) = \mathrm{A}\cos(\omega_C t + k_{PM}\ a\cos\omega_{mod}t) \qquad V_{FM}(t) = \mathrm{A}\cos\left(\omega_C t + \frac{k_{FM}a}{\omega_{mod}}\sin\omega_{mod}t\right)$$




The instantaneous frequency for an FM signal with single tone modulation is therefore,

$$\omega_{FM}(t) = \omega_C + k_{FM}a \cos\omega_{mod}t$$

It's more revealing to write it as,

$$\omega_{FM}(t) = \omega_C + \Delta\omega \sin\omega_{mod}t$$

$\Delta\omega = k_{FM}a$ is called the *frequency deviation*. So in FM there are 3 frequencies involved, even if you're transmitting a single tone:
(1) carrier frequency $\omega_C$ (2) modulation frequency (the tone) $\omega_{mod}$ (3) the frequency deviation $\Delta\omega$ which would be proportional to the loudness.

The figures are waveforms for AM, FM and PM with single tone modulation. FM and PM are similar looking but shifted by 90 degrees with respect to each other. Both PM and FM have a similar squeezing and stretching of the carrier waveform. The bunching up takes place every cycle of the modulation. Having defined the deviation frequency, the actual FM signal sent to the antenna can now be written as (again for a single-tone modulation) ,

$$V_{FM}(t) = \mathrm{A}\cos(\omega_C t + m \sin\omega_{mod}t) \qquad m \equiv \frac{\Delta\omega}{\omega_{mod}} = \frac{\Delta f}{f_{mod}}$$

Here *m* is called the *modulation index*. The big advantage of FM over AM is signal to noise ratio. It turns out that this increases with *m* as,

$$\frac{Signal}{Noise} \propto m^{3/2}$$

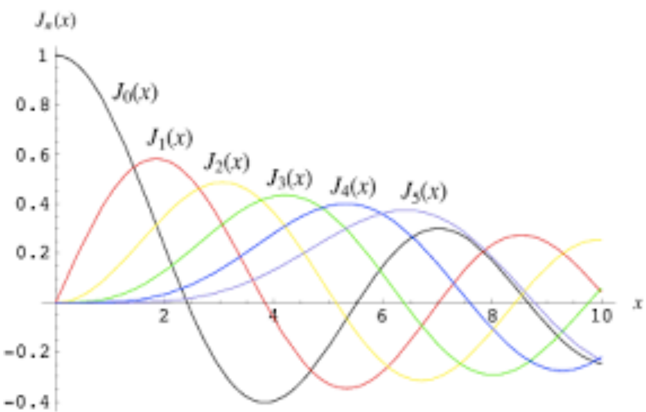
FM is much quieter than AM but the price you pay is increased bandwidth, which then requires a much higher carrier frequency.

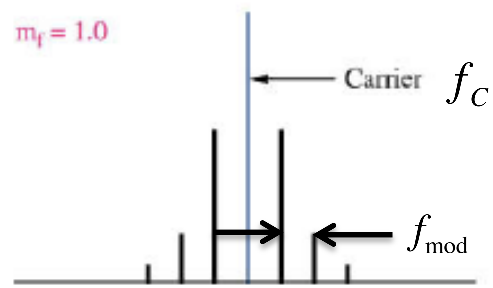
# *FM spectrum*

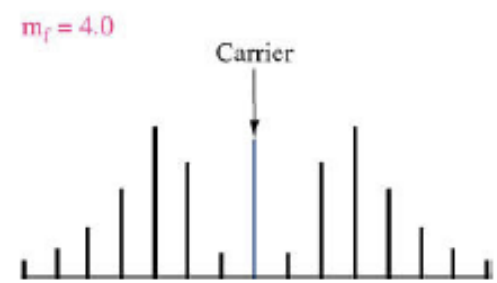
To find the FM spectrum for a single tone modulation we'll use an old mathematical identity:

$$V(t) = \cos(\omega_C t + m\sin\omega_{\text{mod}}t) = J_0(m)\cos\omega_C t + \sum_{l=1}^{\infty} J_l(m)\left\{\sin\left[(\omega_C + l\omega_{\text{mod}})t\right] + (-1)^l \sin\left[(\omega_C - l\omega_{\text{mod}})t\right]\right\}$$

The $J_l(x)$ are Bessel functions of the first kind which oscillate with an envelope that decreases with x. There are an infinite number of sidebands centered on the carrier frequency and separated by the modulation frequency. FM spectra for different modulation indices are shown below. Only a few sidebands are shown.



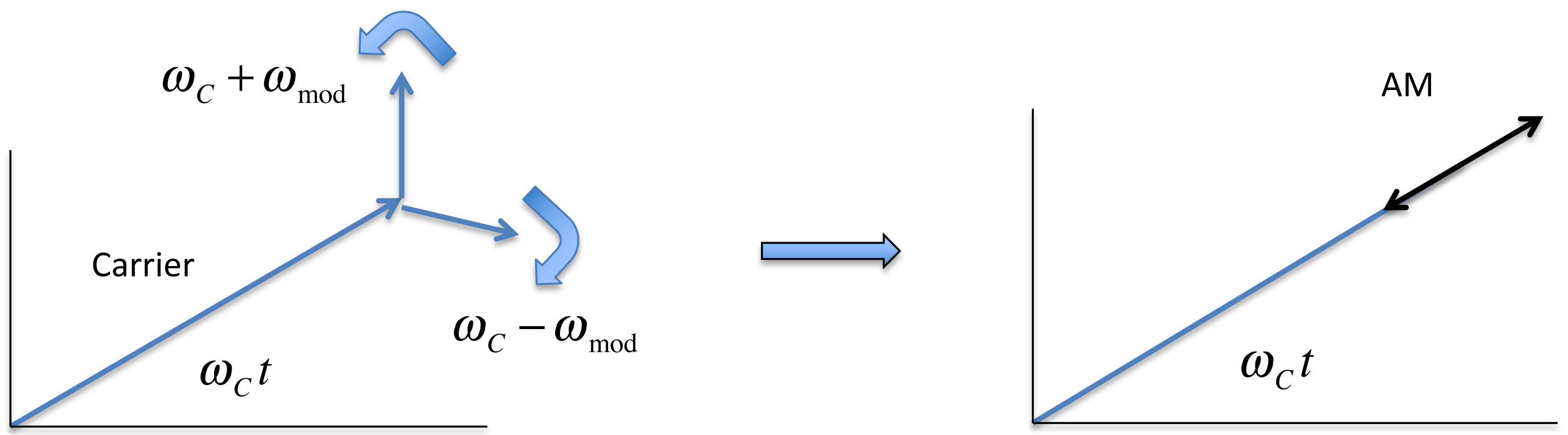

*http://cpassignments.blogspot.com/2015/04/fm-wave-generation_20.html*

The FM spectrum is infinitely wide but a satisfactory amount of information can be conveyed within a few sidebands around the carrier. In fact, 98% of the FM signal power is contained in a bandwidth of size,

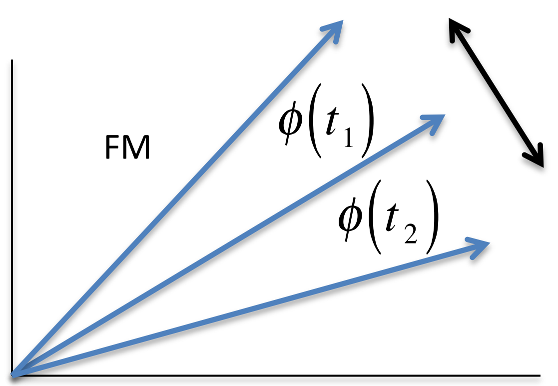
$$Bandwidth = 2(\Delta f + f_{mod}) \qquad m = \frac{\Delta f}{f_{mod}}$$

FM was developed by Edwin Armstrong in the 1930's but the rules for FM broadcasting were developed after WW II. Each FM channel is now allowed a maximum deviation of $\Delta f$ = 75 kHz and maximum modulation frequency of $f_{\text{mod}}$ = 15 kHz. This means a bandwidth of 180 kHz. The carrier frequencies range from 88 – 108 MHz.

It's useful to view modulation with phasors. For single tone AM, there are three phasors. The two sideband phasors rotate in opposite directions since they are above and below the carrier frequency. They therefore modulate the *length* of the carrier phasor but not its phase.

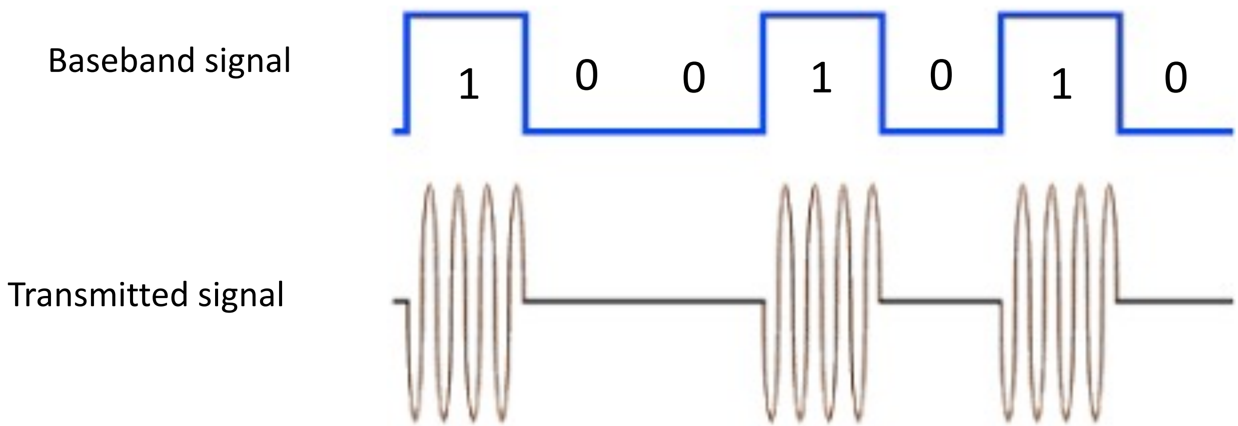



FM and PM are both phase modulation methods. The length of the phasor stays the same but there is an *angular* oscillation at the modulation frequency that is superimposed on the overall counterclockwise rotation at $\omega_C$ .



## *Digital modulation techniques*

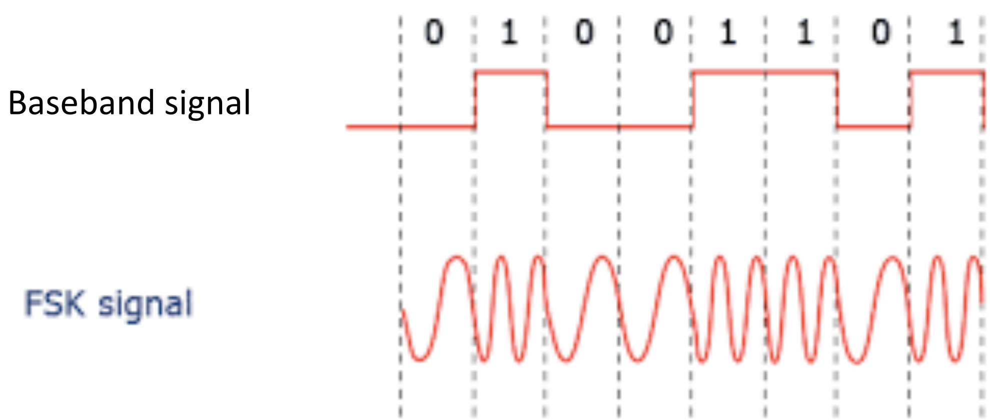
AM, FM and PM have evolved from analog to digital. Probably the simplest is called On-Off keying, shown below. This is digital AM in which the carrier signal is either on or off depending on whether the data bit is a 1 or 0. The string of 1's and 0's is called the *baseband signal*. It's the digital equivalent of the g(t) function we used for analog AM. On-off keying is used for mundane tasks such as garage door openers.




**Frequency shift keying (FSK)**

FSK is a type of digital FM. The signal is switched between two different frequencies that represent 0 or 1:

$$1: \quad f_C + f_{\text{mod}}$$
$$0: \quad f_C - f_{\text{mod}}$$



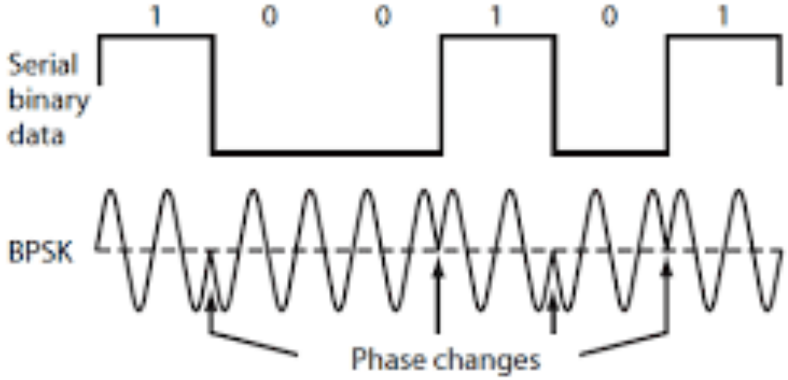

*http://www.as-interface.net/knowledge-base/as-i-wiki/frequenzumtastung-(fsk)?lang=en*

FSK has better signal to noise than on-off keying. In digital terms, it has a lower bit error rate (BER). Bluetooth uses FSK although in a more sophisticated way. In Bluetooth, the carrier frequency is randomly switched among 78 different values. This allows for much lower transmitter power levels but very good signal to noise because even if there is interference at a particular carrier frequency, the system rapidly switches to a new one.

## Phase shift keying (PSK)

In this case, 1 and 0 are represented by signals at the carrier frequency but having different phases. In what is called binary phase shift keying (BPSK) we might have,

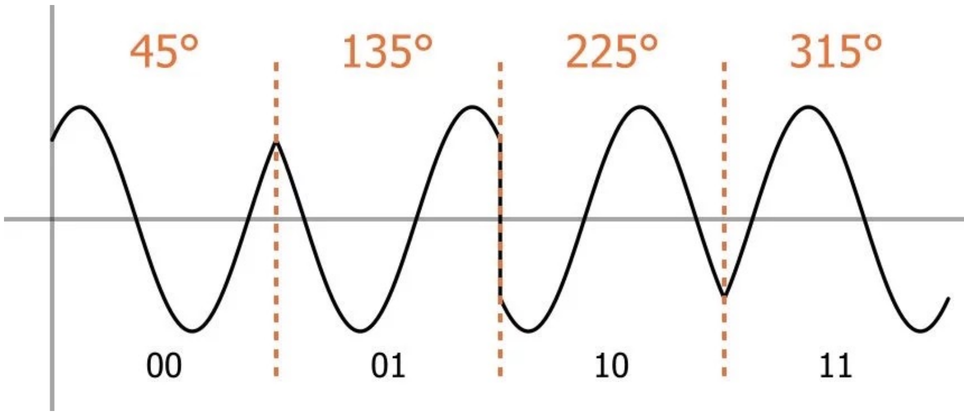
$$1 \;:\; +V_0 \sin \omega_C t$$
$$0 \;:\; -V_0 \sin \omega_C t$$




*agulmeera.org/bandpass-data-transmission/*

GPS signals are modulated using BPSK. The baseband signals have periods of 1 or 10 MHz depending on whether they modulate the L1 carrier (1575.42) or the L2 carrier (1227.60 MHz).

With more sophisticated electronics it is possible to compress even more information into variations of the carrier. The scheme shown to the right is called quadrature phase shift keying. (QPSK). Now, using the same carrier frequency but with 4 different phases, it's possible to transmit 2 bits of information in each time segment (separated by the dotted lines.) This means that the bandwidth required to transmit the same amount of information is half as much as it would be with BPSK.

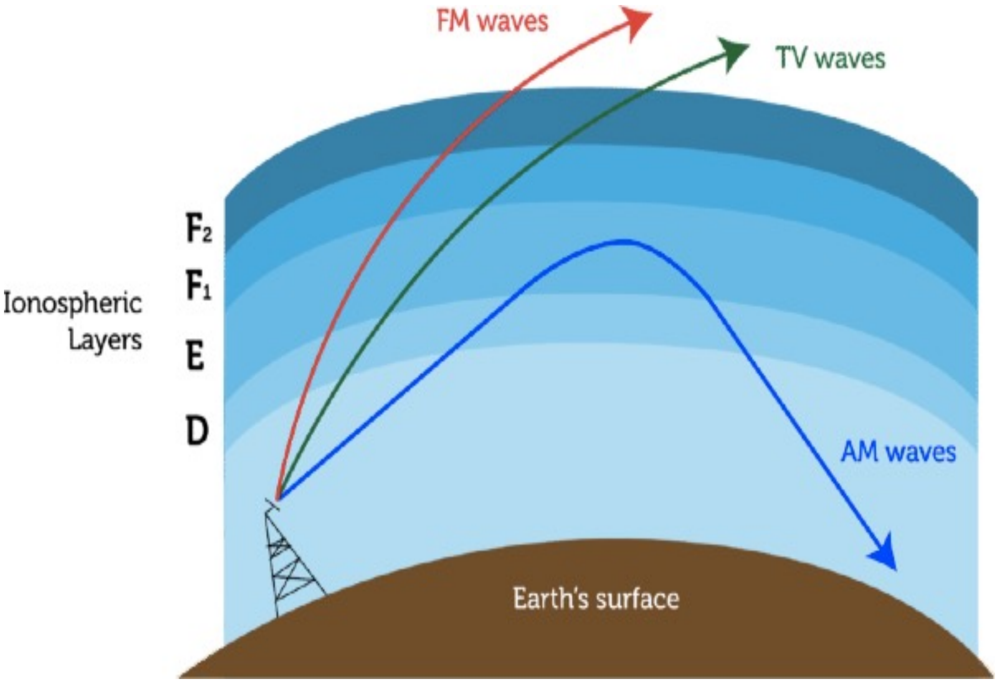



*https://www.allaboutcircuits.com/technical-articles/quadrature-phase-shift-keying-qpsk-modulation/*

## *The Ionosphere*

The atmosphere above about 60 km is partially ionized. High energy photons from the sun ionizie the various constituents including $N_2$, $O_2$, O, Ar, He, N, H. The predominant ion and its density varies with altitude and the time of day, sunspot activity and other things. There is far less ionization at night than during the day. The figure several identifiable layers that arise in the ionosphere. These are known as D, E, $F_1$ and $F_2$ . Each of these contains a dilute, neutral plasma consisting of ions and electrons along with unionized constituents.

The figures illustrates how the atmosphere affects the propagation of AM radio waves and FM radio waves. AM waves bounce off while FM waves pass through. That means that AM can be transmitted over greater distances than FM since it can bounce back and forth between the earth and the ionosphere. The phenomenon was discovered in 1901 by Marconi, who transmitted a radio signal 2200 miles across the Atlantic ocean from England to Newfoundland although he and many others at the time did not understand how it could happen. In the 1920's many radio amateurs were able to transmit signals over even greater distances. Signals that are transmitted by bouncing are sometimes called *skywaves.* To see how it works we'll need some basic E&M.



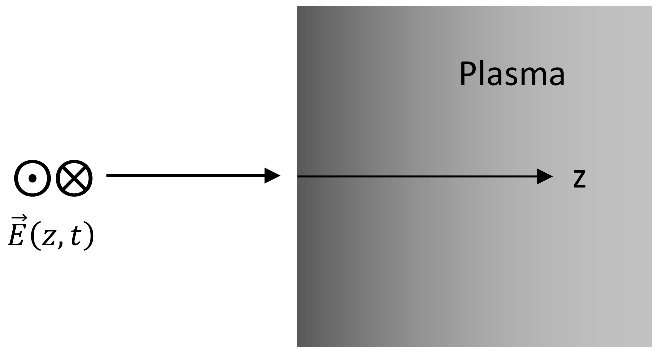

*http://data.allenai.org/tqa/the_electromagnetic_spectrum_L_0753/*

## *Propagation in a Plasma*

Start with Maxwell's equations.

$$div\,\vec{E} = \frac{\rho}{\epsilon_0} \qquad div\,\vec{B} = 0 \qquad curl\,\vec{E} = -\frac{\partial \vec{B}}{\partial t} \qquad curl\,\vec{B} = \mu_0 \frac{\partial\,\vec{P}}{\partial t} + \mu_0\epsilon_0 \frac{\partial \vec{E}}{\partial t} = \mu_0 \frac{\partial \vec{D}}{\partial t}$$

A plane electromagnetic wave polarized along the x-direction, travelling along the z-direction, is incident on the plasma which fills the half-space $z \geq 0$. Consider the plasma as an ideal gas of ions and electrons, each with number density $n_o$ per cubic meter.




Ions are so much heavier than electrons we'll consider them to be stationary, just providing a uniform background charge to keep the plasma neutral (on average.) Treat the electrons as non-interacting classical particles with mass *m* and charge – *e*. The electrons jiggle up and down in response to the electric field in the x-direction. Let force, electron displacement, polarization and electric field, take the complex exponential form form $\vec{F}(t) = Re\big(F(\omega)e^{i\omega t}\big)\hat{x}$ , $\vec{P}(t) = Re\big(P(\omega)e^{i\omega t}\big)\hat{x}$ , etc. Then,

$$\vec{F} = m\frac{d\vec{v}}{dt} = -e\vec{E}\,, \qquad \vec{P} = -n_0 e\vec{x} \quad \rightarrow \quad P = -\frac{n_0 e^2}{\omega^2 m}E$$

To find the dielectric constant use,

$$\vec{D} = \epsilon_0\vec{E} + \vec{P} \quad \rightarrow \quad D = -\epsilon\, E = \left(1 - \frac{\omega_p^2}{\omega^2}\right)E \qquad \omega_p = \sqrt{\frac{n_0 e^2}{\epsilon_0 m}}$$

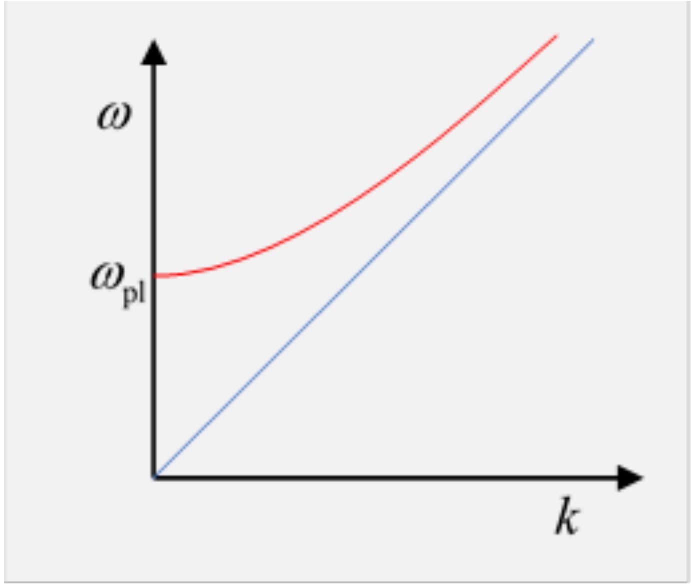
The quantity $f_p = \omega_p/2\pi$ is called the *plasma frequency*. The maximum density *n* of free electrons in the ionosphere varies with altitude, time of day and solar activity but a typical number is $n_0 \sim 10^6/cm^3$ corresponding to $f_p$ = 9 MHz. If a wave with $f < f_p$ encounters the plasma it's reflected back. You can see this by considering the phase velocity of the wave and wavenumber *k*,

$$V_{phase} = \frac{\omega}{k} = \frac{c}{\sqrt{\epsilon(\omega)}} = c\left(1 - \frac{\omega_p^2}{\omega^2}\right)^{-1/2} \rightarrow \quad k = \frac{\sqrt{\omega^2 - \omega_p^2}}{c}$$

The dispersion relation for the wave takes the form,

$$\omega^2 = c^2k^2 + \omega_p^2$$

which is shown on the right. If $\omega < \omega_p$ then the wavenumber k is imaginary and the wave won't propagate. The wave is evanescent and its energy is entirely reflected back.



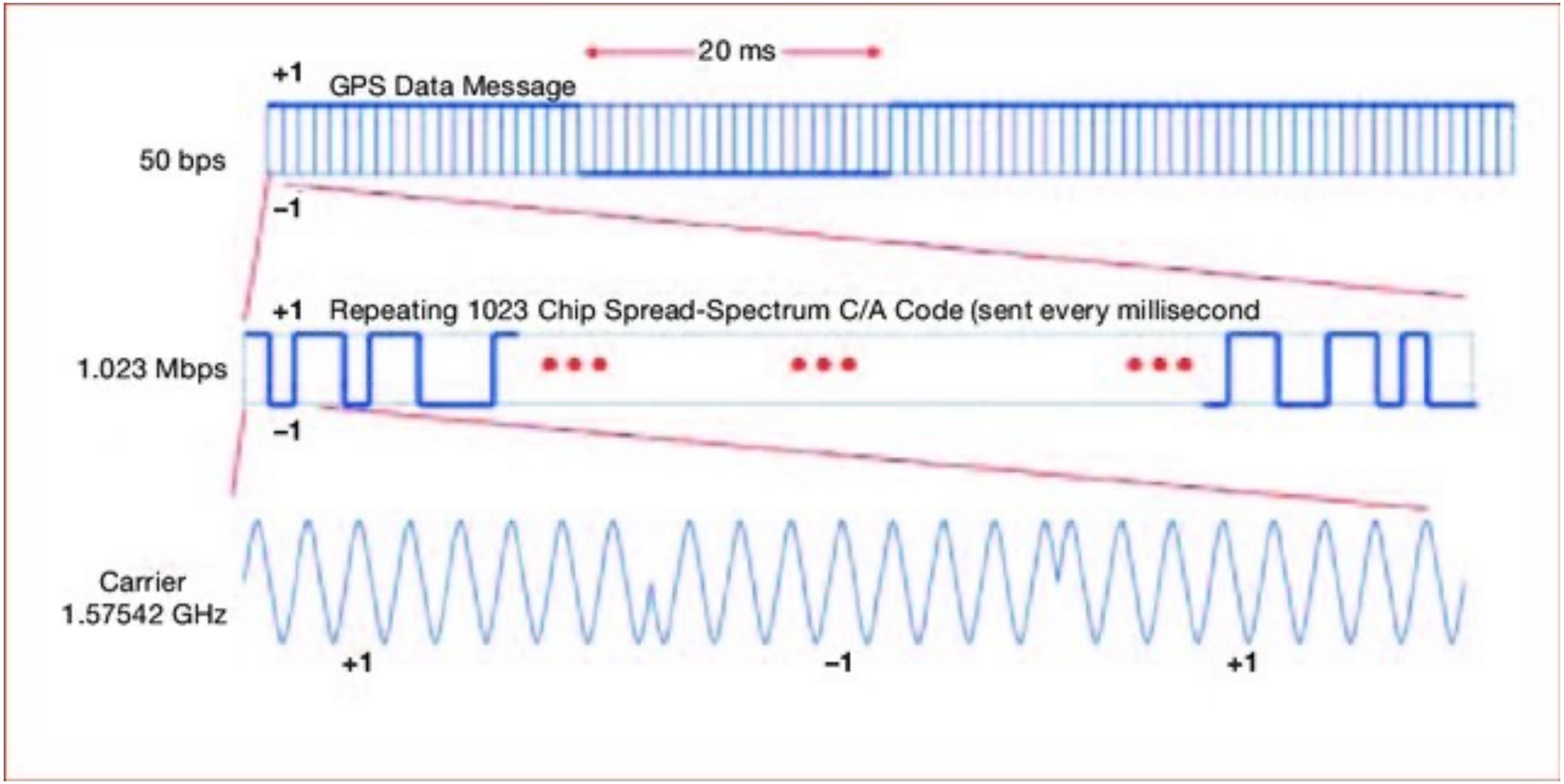

For the peak ionized electron densities in the ionosphere the plasma frequency $f_p \sim 10$ MHz, although this changes from day to night. If $f < f_p$ then a radio wave transmitted from the earth's surface can then be transmitted by successive reflections off the ionosphere and the earth. This works for AM since $f < 1.6$ MHz. FM signals, $88 < f < 108$ MHz, so they pass right through the ionosphere.

## *Ionosphere Delay*

Although electromagnetic waves with $f > f_p$ do indeed pass through the ionosphere, the passage still affects them. In particular, GPS signals, which operate at frequencies well above $f_p$ suffer a delay as they pass through the ionosphere and this causes significant problems. GPS signals are modulated with BPSK. The L2 carrier signal operates at $f = 1227.6$ MHz and is modulated every $10^{-3}$ seconds by a string of 1's and 0's about 1000 units ("chips") long – the baseband signal. This baseband code identifies the satellite and is used to figure out the transit time between the satellite and the receiver, and therefore the distance.

If you recall from E&M, we can define both a phase and a group velocity for a wave. For a wave in free space they are the same but in the plasma the wave has dispersion so they are not the same. In the plasma they are given by,

$$V_{Phase} = \frac{\omega}{k} = \frac{c}{\sqrt{\left(1 - \frac{\omega_p^2}{\omega^2}\right)}} \qquad V_{Group} = \frac{\partial\omega}{\partial k} = \frac{c^2}{V_{Phase}} = c\sqrt{\left(1 - \frac{\omega_p^2}{\omega^2}\right)}$$




*https://www.researchgate.net/publication/269285648_The_Use_of_GPS_Disciplined_Oscillators_as_Primary_Frequency_Standards_for_Calibration_and _Metrology_Laboratories/figures?lo=1*

The baseband signal, also called the C/A code, modulates the 1.57542 GHz carrier. It is used to determine the travel time between the satellite and the receiver. That information travels at the *group velocity*. Since it's travelling through the ionosphere the plasma frequency and therefore the group velocity depends on the altitude (x). Therefore the time for the modulated signal to go a distance D is the integral,

$$T = \int_0^D dx \frac{1}{V_{Group}(x)} = \int_0^D dx \frac{1}{c\sqrt{\left(1 - \frac{\omega_p^2(x)}{\omega^2}\right)}} \approx \int_0^D \frac{dx}{c}\left(1 + \frac{1}{2}\frac{\omega_p^2}{\omega^2}\right) = \frac{D}{c} + \int_0^D \frac{dx}{2c}\frac{\omega_p^2}{\omega^2} \qquad \frac{\omega_p^2}{\omega^2} < 10^{-2}$$

This expression shows that there is a *delay* due to passage through the plasma. The problem is, we don't know the plasma density profile through the ionosphere all that well and it changes from one instant to the next. In older versions of GPS it had to be estimated from other measurements. Modern GPS satellites transmit at 2 different carrier frequencies, L1 (1.2276 GHz) and L2 (1.5742 GHz). With two frequencies we'll get two different delay times.

$$T(\omega_1) = \frac{D}{c} + \frac{1}{2c\omega_1^2}\int_0^D \omega_p^2\,dx \qquad T(\omega_2) = \frac{D}{c} + \frac{1}{2c\omega_2^2}\int_0^D \omega_p^2\,dx$$

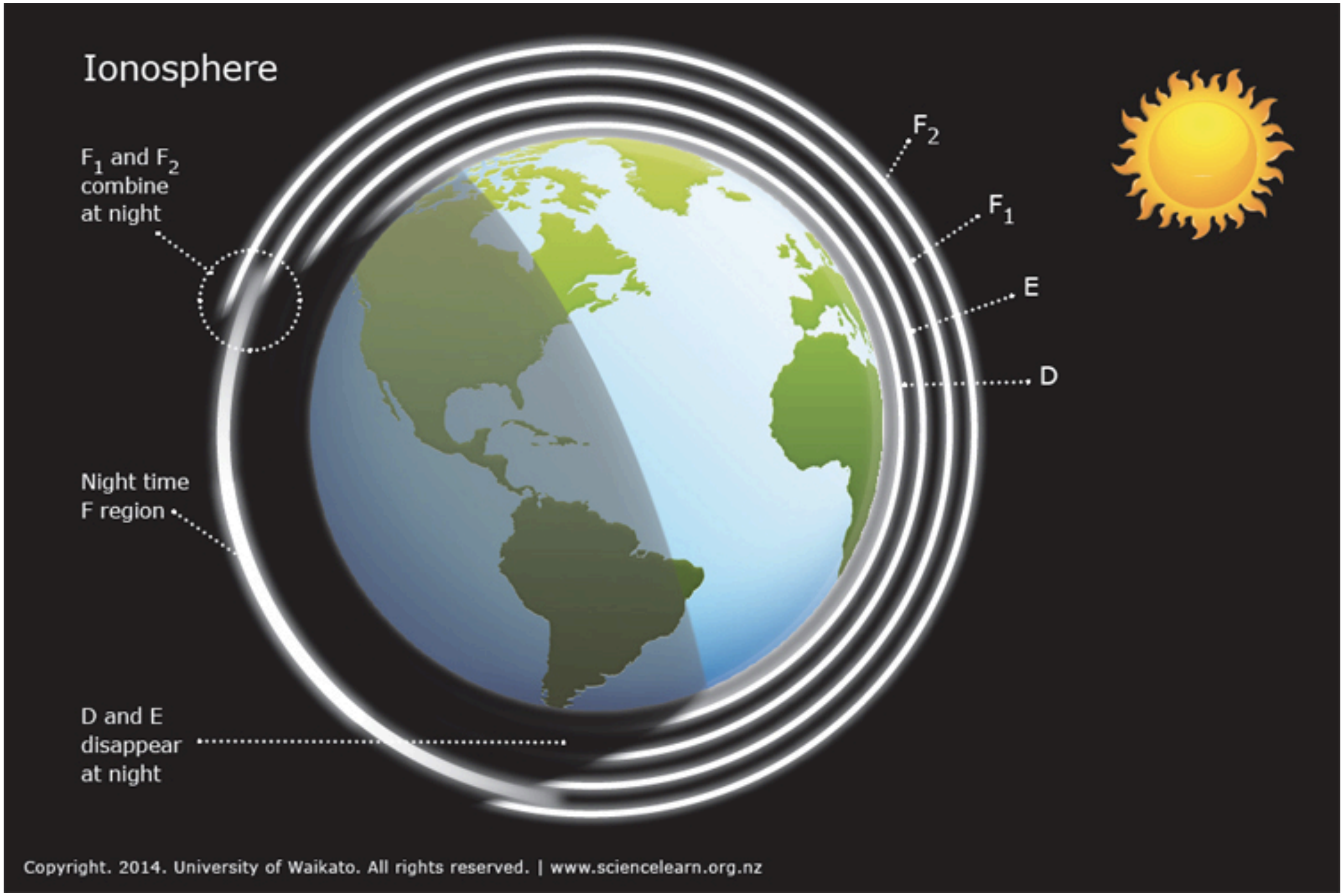
Subtracting them gives,

$$T(\omega_1) - T(\omega_2) = \frac{1}{2c}\left(\frac{1}{\omega_1^2} - \frac{1}{\omega_2^2}\right)\int_0^D \omega_p^2\,dx$$

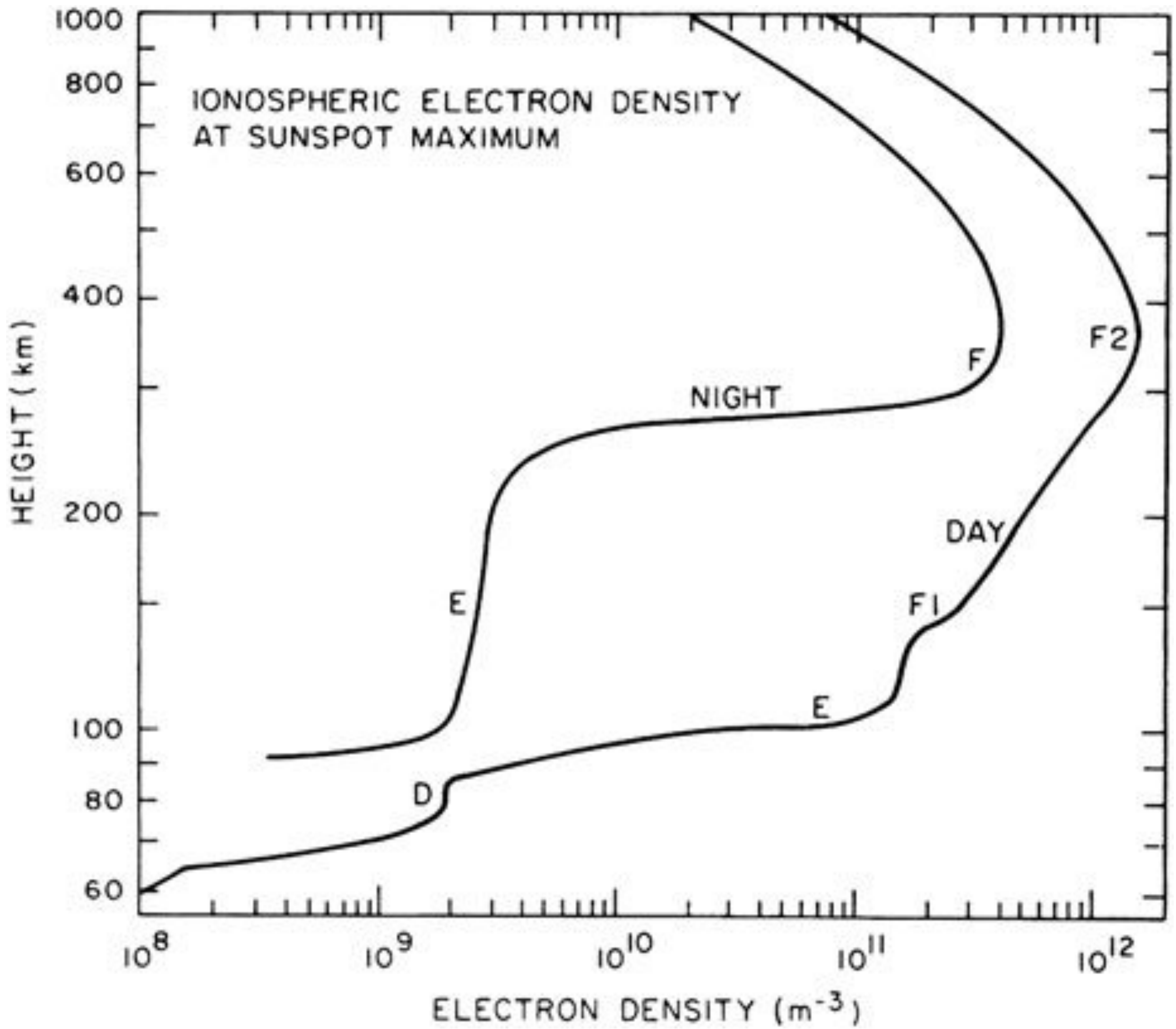
Since we've measured the delay times and we know the transmission frequencies we can obtain the unknown integral,

$$\int_0^D \omega_p^2\,dx = \frac{e^2}{\epsilon_0 m}\int_0^D n_0(x)\,dx$$

The integral measures the total amount of electron density the wave must travel through from the GPS satellite to the receiver. Ionospheric corrections have traditionally been a major source of error for GPS. The two-frequency approach is more costly and complex but vastly improves GPS precision.




*https://www.sciencelearn.org.nz/images/248-layers-of-the-ionosphere*




*https://www.researchgate.net/publication/314253214_Propagation_Effects_Ionized_Media/figures?lo=1*

## *Plasma oscillations*

In our calculation of electromagnetic waves, $f_p$ determines a cutoff frequency, below which waves won't propagate. What is this frequency? Consider an infinite plasma, With no disturbance, the ionization produces equal numbers of ions and electrons, each with number density $n_0$. The plasma is neutral. But now imagine a small deviation $\delta n(x,t)$ of the electron density that depends on x. The ions are heavy so assume they are stationary and maintain their equilibrium density $n_0$. The total *electron* density is,

$$n(x,t) = n_0 + \delta n(x,t)$$

Without the disturbance, the net charge density is zero and there's no electric field, *With* the disturbance the imbalance in charge density will produce an electric field via Gauss's law,

$$\nabla \cdot \vec{E} = \frac{\rho}{\epsilon_0} \quad \rightarrow \quad \frac{\partial E_x}{\partial x} = \frac{e}{\epsilon_0}(n_0 - n_0 - \delta n) = -\frac{e\,\delta n}{\epsilon_0}$$

The plasma must also obey the equation of continuity,

$$0 = \nabla \cdot \vec{J} + \frac{\partial \rho}{\partial t} = \frac{\partial J_x}{\partial x} - e\frac{\partial\,\delta n}{\partial t} = 0$$

Since the ions are assumed to be stationary, the current density $J_x$ comes entirely from the motion of electrons. Assuming that the disturbance is small, $\delta n \ll n_0$, we have,

$$J_x = -env = -e(n_0 + \delta n)v \approx -en_0 v$$

where $v$ is the electron velocity in the x-direction. As earlier, it obeys Newton's law,

$$m\frac{dv}{dt} = -eE_x$$

The equations are now

$$\frac{\partial E_x}{\partial x} = -\frac{e\delta n}{\epsilon_0} \qquad \frac{\partial J_x}{\partial x} - e\frac{\partial\, \delta n}{\partial t} = -en_0\frac{\partial v}{\partial x} - e\frac{\partial\, \delta n}{\partial t} = 0 \qquad m\frac{dv}{dt} = -eE_x$$

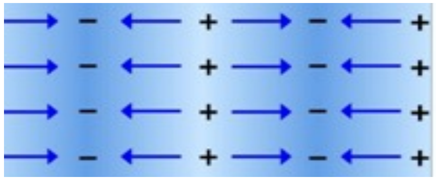
These are linear equations so we'll assume wave-like solutions,

$$\delta n(x,t) = \delta n(\omega,k)e^{i(kx-\omega t)} \qquad E_x(x,t) = E_x(\omega,k)e^{i(kx-\omega t)} \qquad J_x(x,t) = J_x(\omega,k)e^{i(kx-\omega t)}$$

Plugging in these solutions, cancelling out all the $e^{i(kx-\omega t)}$ factors and eliminating $E_x(\omega,k)$ and $J_x(\omega,k)$ we end up with,

$$0 = \delta n(\omega,k)\left(\omega^2 - \frac{n_0 e^2}{\epsilon_0 m}\right) \quad \rightarrow \quad \omega^2 = \frac{n_0 e^2}{\epsilon_0 m} = \omega_p^2$$

This is a disturbance that oscillates at the plasma frequency. However, there is no k-dependence! It's not a propagating wave but more like a collection of independent harmonic oscillators. Remember that these plasma oscillations are not like radio waves. Radio waves are *transverse*, i.e., the E and B fields are perpendicular to the direction of propagation. In a plasma oscillation there is only an electric field and the disturbance is *longitudinal* as shown below.



https://www.jeol.com/words/emterms/20121023.060058.php#gsc.tab=0

We've ignored any thermal energy of the electrons and assumed that the plasma extends infinitely in space. Non-zero temperatures and finite spatial extent will result in a propagating disturbance in which $\omega$ does depend on $k$. Plasma oscillations can be excited in all sorts of materials. In metals the densities are much higher than in the ionosphere so the plasma frequencies are extremely high. In aluminum $f_p$ = 3.8 x $10^{15}$ Hz so visible light, whose frequency is less than $f_p$, is reflected back.