The Physics of Global Navigation

Alex Rose

75725970

School of Physics and Astronomy

The University of Manchester

BSc Dissertation

February 2012

Abstract

An overview and discussion of various devices used for global navigation, and the physics that governs them, which covers a range of topics from the celestial navigation and the sextant, to the pseudorandom codes that allow GPS to function. Particularly focus is spent on Inertial Navigation Systems, Radio navigation, and the Global Positioning System.

Word count: 4409.

1. Introduction

Migration has always had significant implications for all life on Earth, allowing species to move to other locations and evolve to fit new conditions. Man, despite being a land mammal, has managed to traverse every continent on Earth, which in turn has allowed our population to grow to an incredibly large number for a species of such relatively large biomass.

The International Air Transport Agency projected in 2011 that by 2014, there will be 3.3 billion air passengers annually[1]. This suggests that the average human travels by aircraft once every 26 months. Combined, we consume around 250 million litres of jet fuel a year[2], and this is but one of our many modes of transport. It is fair to say that, as a species, we are rather interested in traversing the globe.

Furthermore, we have a penchant for the unknown. Many revered works of fiction revolve around exploration of new found lands. Especially in the past century, since space travel has become viable, we dream of permeating through space to explore the vast universe that eclipses us, and renders us trivial in its immensity.

But to accomplish such feats of motion, we have always required tools with which to navigate, to successfully reach our destinations as efficiently as possible. Some such devices seem elementary with our modern knowledge of physics. Others are technical marvels, using elaborate mechanisms to break the constraints of our own senses, and overcome complex problems presented by general relativity.

One of the earliest forms of navigation was by observing celestial bodies. By observing Polaris - the most visible star in Ursa Minor, which lies almost along the axis of rotation of Earth - one can tell the direction of true North to an accuracy of 1 degree. Eventually, as our knowledge of physics and astronomy were increased, celestial navigation became sophisticated, with the invention of the sextant.

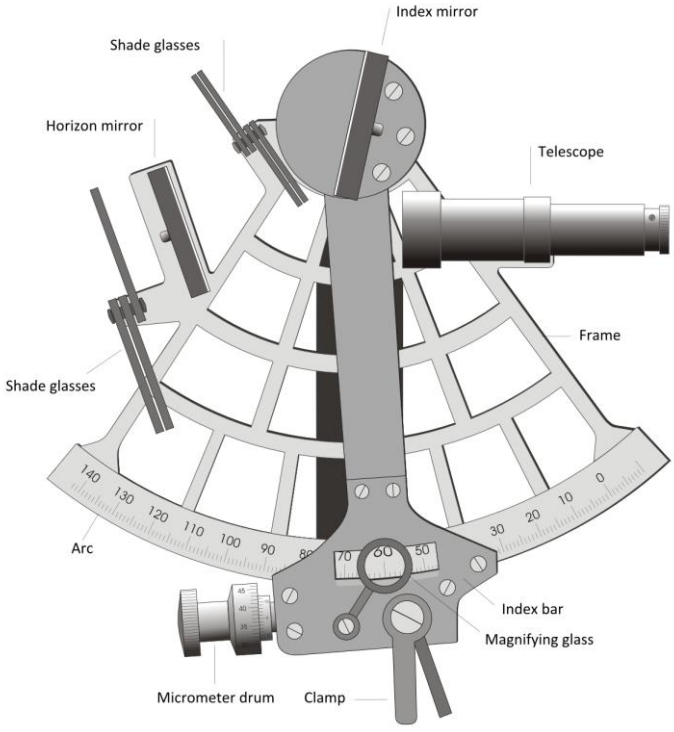

Fig. 1 – Diagram of a sextant[3]

The sextant is a device, which can be used to measure the angular separation between two objects. The user points the sextant at the first object of interest (usually the horizon in navigation), and then slides the index bar until the light from the second object is reflected off the index mirror and horizon mirror, such that the two images are horizontally aligned when observed through the telescope. The shades simply prevent eye damage. Using the vernier scale micrometer drum, it is then possible to read the angle to a high level of precision.

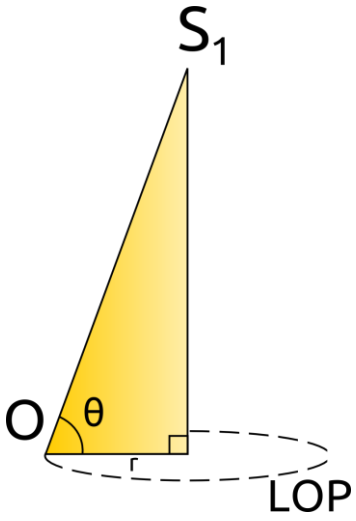

Fig. 2 – The LOP on which an observer O knows he lies, determined from the angle θ to a certain star, S_1.

To ensure that the object is at its lowest possible point, the sextant is rotated along the axis of the telescope, and the time of measurement and altitude are noted. By measuring the angle θ of a known star from the horizon, and using a book containing known reference values for the star in question at the time of measurement, it is possible to determine the distance r along the surface of Earth to the point at which the star is exactly overhead.

From this, a navigator knows he is on a specific circle along the plane of the surface of Earth, of radius r, known as the line of position (LOP). He does not know his direction from this point, however. By examining a second star, a second radius can be deduced. The navigator now knows he is at one of the points where both the circular functions meet. In the case that there is only one real root, he knows his position, although this is only possible if the radii are parallel. More realistically, he can look up a third star, in order to find the intersection between three circles. This will give him his location, to a certain precision (based on the precision of his watch, altimeter, and the divisions on the reference table). There is reasonable scope for both human error, and accumulative lack of precision.

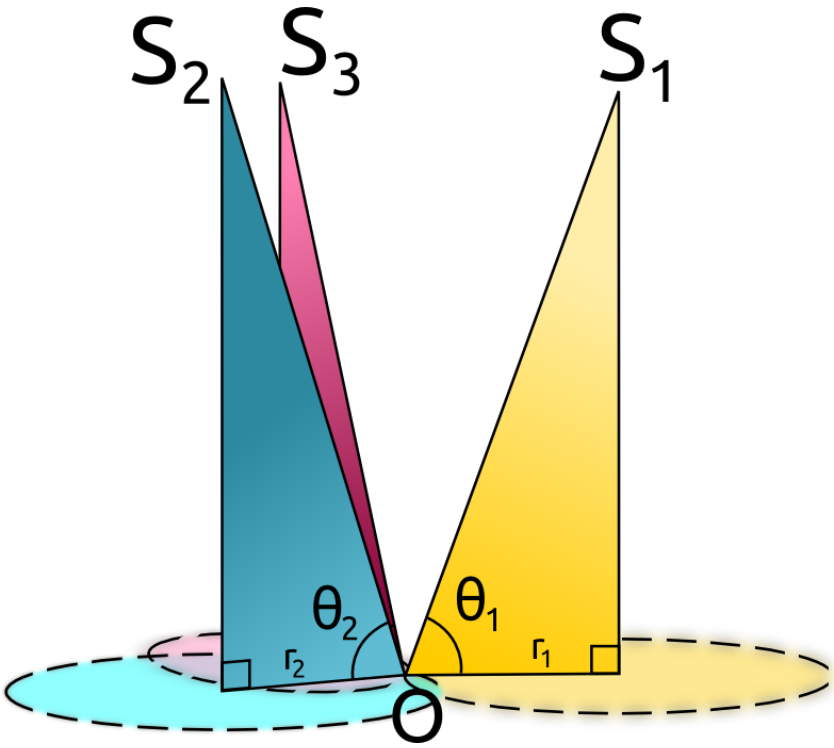

Fig. 3 – By measuring the angles to three known stars, an observer knows his location.

2. Inertial Navigation Systems

Inertial systems rely on dead reckoning. This means that they are initially finely calibrated, and subsequently continuously measure their motion in order to determine how far they have travelled in each direction. This is, obviously, subject to large accumulated errors.

In three dimensional space, there are six degrees of freedom, corresponding to translation and rotation through each of the three spatial dimensions. If a rigid body can continuously accurately measure its acceleration and angular velocity at all times, it can therefore calculate its precise location.

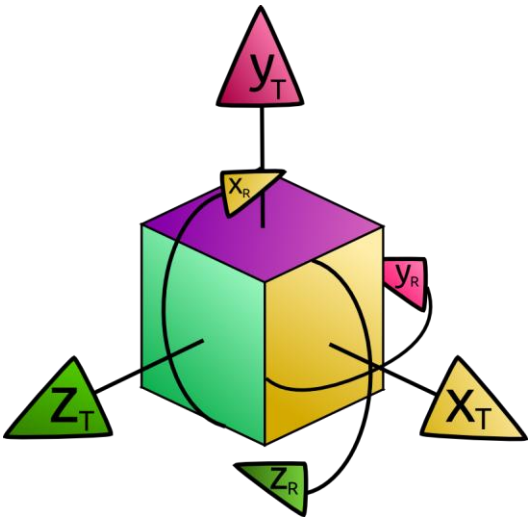

Fig. 4 – The six degrees of freedom a cube has in three dimensional space.

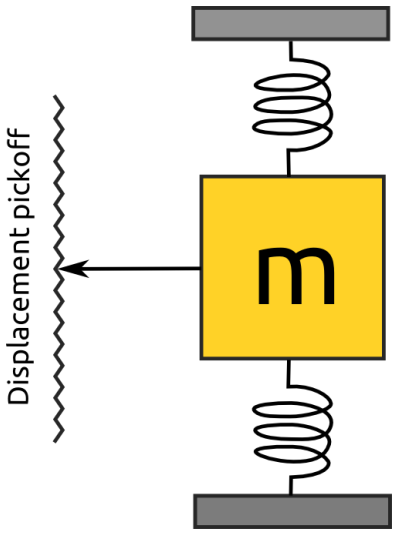

Fig. 5 – A mass suspended by two springs in a mechanical accelerometer.

Proper acceleration can be measured with an accelerometer, which is generally a mechanical or solid state device. A mechanical accelerometer uses a mass-spring system. A mass is suspended from springs, free to move in one axis. As the mass experiences a force, the spring in the corresponding direction will contract, which will produce a tension proportional to the displacement of the mass from its rest position. Using Newton's second law of motion, the acceleration of the mass can then be determined, using:

$$\ddot{x} = -\frac{kx}{m},$$

where k is the combined spring constant of the compressed springs, m is the mass of the mass, and x is its displacement.

A more convenient accelerometer, for smaller electronic devices, is a microelectricalmechanical system (MEMS). This consists of a seismic mass equivalent to the mass in a mechanical system - a comblike section of silicon, whose thin elastic edges allow it to move in one axis, equivalent to the springs. The "fingers" of its structure pass by static silicon pieces, creating a differential capacitor, and allowing current to flow. Thus, the motion can be determined by the amount of current flowing.

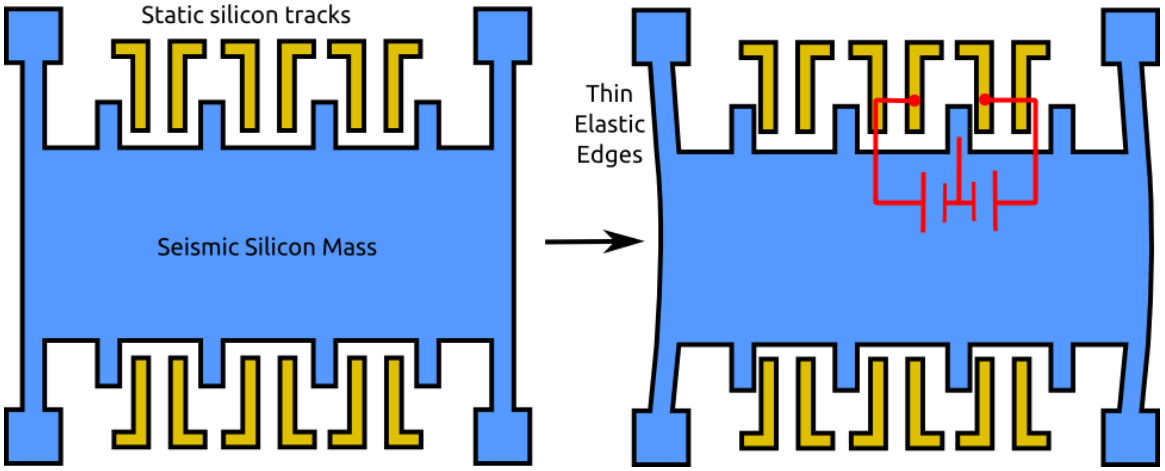

Fig. 6 – A silicon comb structure, which can move freely in one axis. As it moves, its "fingers" create a differential capacitor with the static silicon pieces. As the large segment moves, current flows.

If an accelerometer is fixed to a rigid body, since the rigid body and the accelerometer will move together, the proper acceleration of the rigid body is known. If a non-rotating object's initial velocity, position and orientation are all known, its position relative to inertial space can be found by integrating the measured acceleration twice. Since it is measured by a computer in discrete steps, an integration algorithm can be used.

It is important that a numerical integration method with a low order of error is selected and a low step size between measurements is used or the model will quickly become imbalanced. The Störmer-Verlet method, which has order $O(\Delta t^4)$, is a good selection.

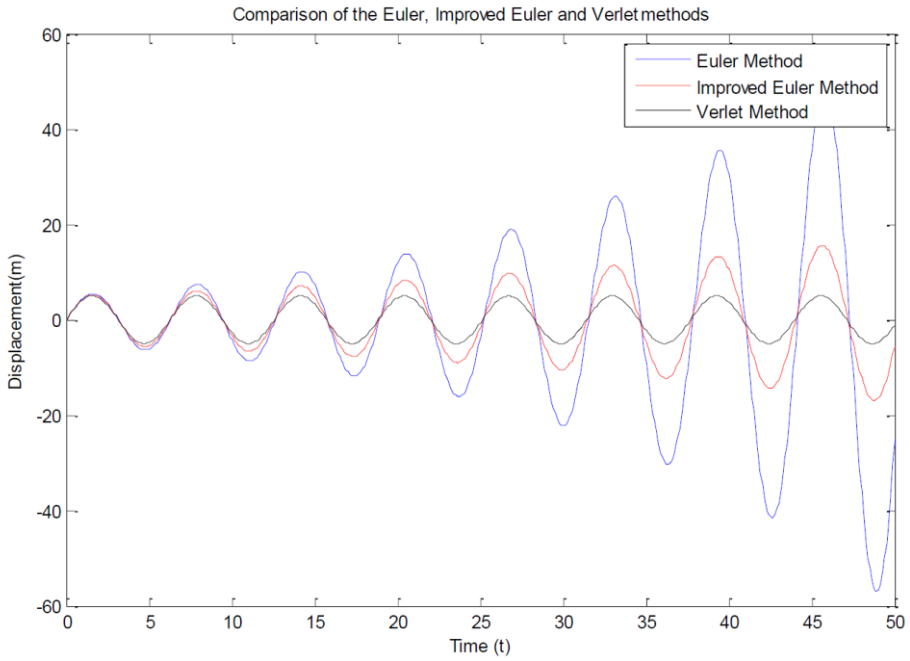

Fig. 7 – SHM in a mass spring system, as modelled using three different integration methods for the same step size. Ideally, the amplitude should remain constant, although over time errors accumulate, and this does not hold true.

As previously noted, when using accelerometers, only the *proper acceleration* is measured. This is the acceleration relative to an observer who is in a state of free fall. A correction must be applied, to account for the gravitational pull of Earth. This is also based on the altitude, so height should be measured with an altimeter in order to make an accurate correction.

By using three accelerometers, one for each dimension, it is possible to determine the force experienced on the object in any direction. However, in a strapdown system, since the accelerometers are aligned with the rigid body, they only tell the magnitude of the acceleration in directions *relative to the body*. If the orientation of the body is unknown, this data is meaningless. Thus, simultaneously, the orientation of the body must be measured.

This can be achieved using gyroscopes. A mechanical gyroscope features a rotating disc mounted on separate gimbals, which allow it freedom to rotate in any direction. As angular momentum is conserved, the disc will stay at a fixed orientation, and the angles between the gimbals will change, as the disc resists the change. These angle separations can be read continuously in order to determine the orientation of the device. This is subject to errors, as friction causes damping.

A much more elegant solution can therefore be used, in the form of a ring laser gyroscope (RLG). This exploits the Sagnac effect, in order to determine its absolute rotation. A light source fires a beam of light into a prism, which splits into two beams, which in turn are reflected by mirrors, returning to the prism, and entering a detector.

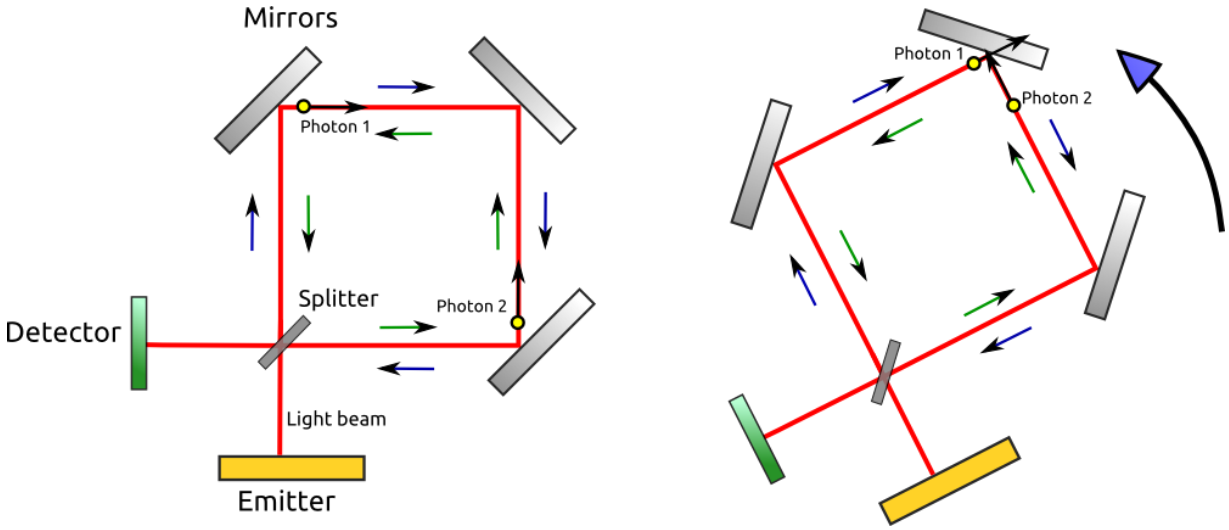

Fig. 8. Two photons obeying the Sagnac effect. In the non-rotating initial frame, they are at an equidistant point on their path. After the frame is rotated, however, one photon arrives at the middle mirror faster, as the speed of light is constant in all inertial frames.

As the frame rotates, the interference pattern changes, as photons split in one direction arrive sooner than in the other direction, as their speed is constant, and the mirrors are rotating towards or away from them respectively. By analysing these spectra, it is possible to take readings of the angular velocity, which can then be numerically integrated to determine the orientation.

A similar type of gyroscope, the fibre optic gyro, uses fibre optic cables coiled along each axis of rotation. The Sagnac effect, again, means that light has to travel less far when the fibre optic cable rotates towards the light. By firing light into both ends of

the cable, in opposite directions, the phase shift can be measured, and used to determine the angular velocity.

Alternatively, another MEMS can be used to measure the Coriolis effect inexpensively. In a rotating frame, two "fictitious" forces are experienced by rigid bodies – the centrifugal force and the Coriolis force, as the basis vectors of motion are time dependent. These forces can be expressed as:

$$F_{fict} = -2m\boldsymbol{\omega} \times \boldsymbol{v}' - m\boldsymbol{\omega} \times (\boldsymbol{\omega} \times \boldsymbol{r}),$$

where m is the mass of a rigid body in the rotating frame, $\boldsymbol{v}'$ is its velocity as determined in the rotating frame, $\boldsymbol{\omega}$ is its angular velocity, and $\boldsymbol{r}$ is its displacement.

The right hand term, the centrifugal force, causes motion away from the pivot of rotating frame. The left hand term, the Coriolis force, is responsible for motion in the opposite direction of the rotation.

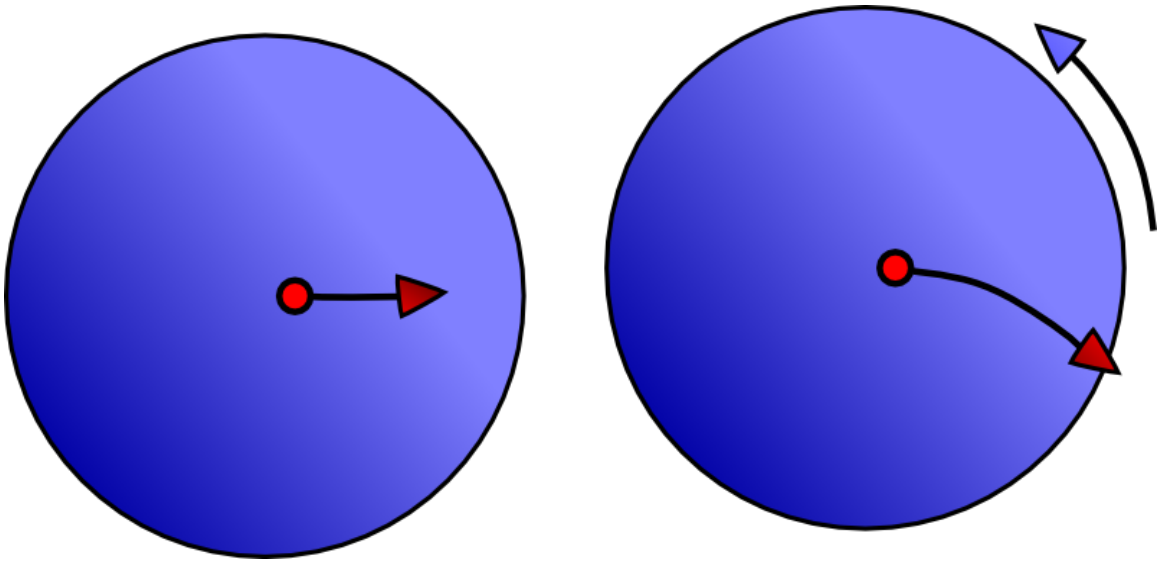

Fig. 9. A ball moves along a circular surface with a certain velocity. As the surface is then rotated, the ball takes an arced path along the surface, in the opposite direction to the rotation. For an observer in the rotating frame, the ball takes a curved path.

We can therefore see that the Coriolis effect can be observed from within the rotating frame itself. We can, thus, measure its effect and determine the angular velocity of the frame. A MEMS gyroscope does exactly this. A mass is oscillated along axis A, perpendicular to axis B, around which it will measure the rotation. Motion along a third axis, C, perpendicular to both of these, is sensed, and therefore any motion due to the Coriolis force arising from rotation around axis B is detected. From this, the angular velocity of the object can be deduced.

This becomes rather complicated and prone to error, as the Coriolis force due to Earth's rotation has to be taken into account at all times. This method is not as accurate as using RLGs, but it is relatively small and inexpensive, so it is used in small electronic devices.

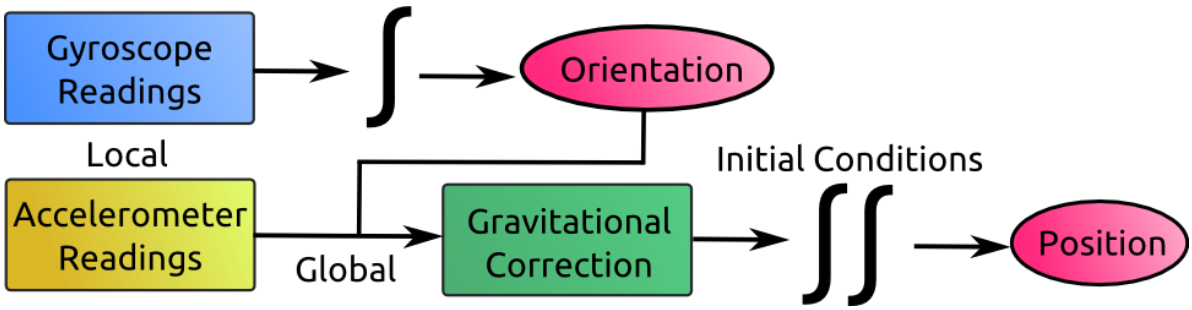

Fig. 10. The method of using locally measured accelerometer and gyroscope readings in order to determine position (and orientation) in a strapdown system.

An alternative to strapdown system is to use a gimballed system. In this system, a gyroscope is used to keep an accelerometer static in the global frame. As a rotation is made in one axis, the gimbals rotate until the horizontal accelerometers measure no force, and the vertical accelerometer measures a value of g, from Earth's gravitational field. Orientation can then be determined by the gimbal, but the accelerometer readings do not require the orientation data in order to determine position; the data is simply integrated twice, and as they always remain in the same orientation, they will measure the position accurately in each direction.

To improve the data received from an INS, additional sensors can be used to feed back information. For instance, a magnetic compass can be used to verify attitude. This is an important process, as inertial systems are prone to exponential buildup of errors, and require recalibration.

3. Radio Navigation Systems

Radio towers were initially used as waypoints, transmitting Morse code or specific frequencies in order to provide information with which navigators could, using a detector and an oscilloscope, observe traces and use reference tables to determine their position. As these could not be used to pinpoint their location anywhere on Earth, only to aid a human in calculating their rough position, they were classed as navigation aids.

One of the first methods of radio navigation was Very Low Frequency (VLF) navigation. The ionosphere derives its name from the charged particles which inhabit it, which allow it to act as a conductor. The volume between the Earth and the ionosphere behaves a huge waveguide, with the surface acting as the ground plane. VLF and ELF waves can therefore propagate throughout the entire globe relatively easily. Very few VLF transmitters are actually, therefore, required.

Another advantage to VLF waves is that they can penetrate water to a depth of several metres, so underwater vessels can transmit and receive VLF signals without needing to surface. A VLF detector can even be mounted to a float and allowed to drift upwards, thereby allowing even greater depths to be maintained, while still retaining the capability of communication.

Because of the range of VLF, locating a precise location on Earth is possible using a few radio towers. One radio tower is a master transmitter. This emits randomly generated pulses based on a specific seed. Secondary towers take the transmitted signal and simply repeat it.

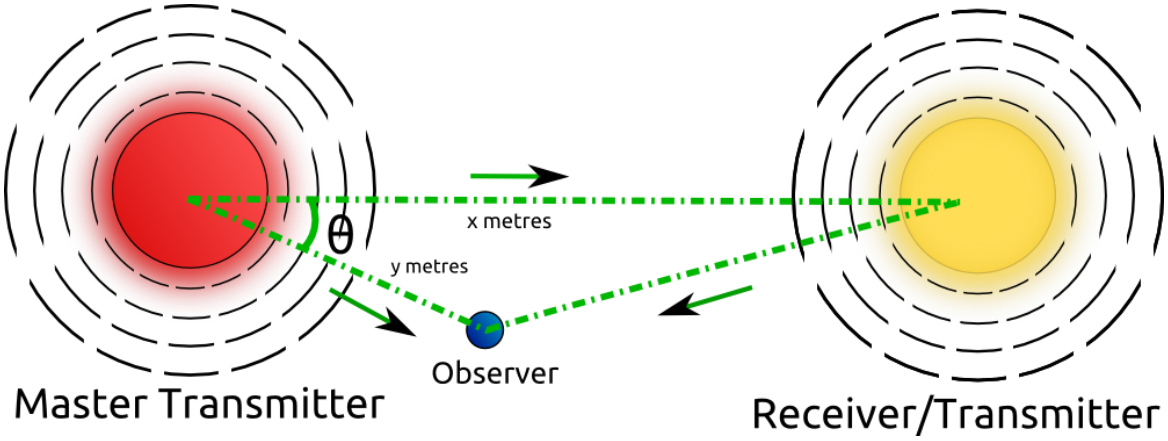

Fig. 11. Transmission of data between a master transmitter and a receiver, which retransmits the exact same data again.

Say, for instance, the transmitter and receiver are x metres apart, and the lag between receiving signal and retransmission is negligible. The master transmitter begins broadcast at $t = 0$. At $t_1 = x/c$ seconds, the secondary transmitter will receive the initial signal and transmit it. Now assume that an observer is y metres from the master transmitter. He will receive the initial signal at $t_2 = y/c$ seconds, and the repeat of the original signal at $t_3 = \frac{x+\sqrt{x^2+y^2-2xycos(\theta)}}{c}$ seconds.

Because the pulses are generated from a seed, as long as the observer's onboard computer has been synchronised with the transmitter, it can calculate how much time has passed since the data was transmitted by the time delay on the pulse sequences. Thus, he knows when $t = 0$, and that he is ct_2 metres from transmitter 1, and $\text{acos}\left(\frac{(x^2+(ct_2)^2-(ct_3-x)^2)}{2xct_2}\right)$ radians from the line which joins them. This is not sufficient to know his location, however, as there are two solutions to this. A third tower can be used to fix this problem.

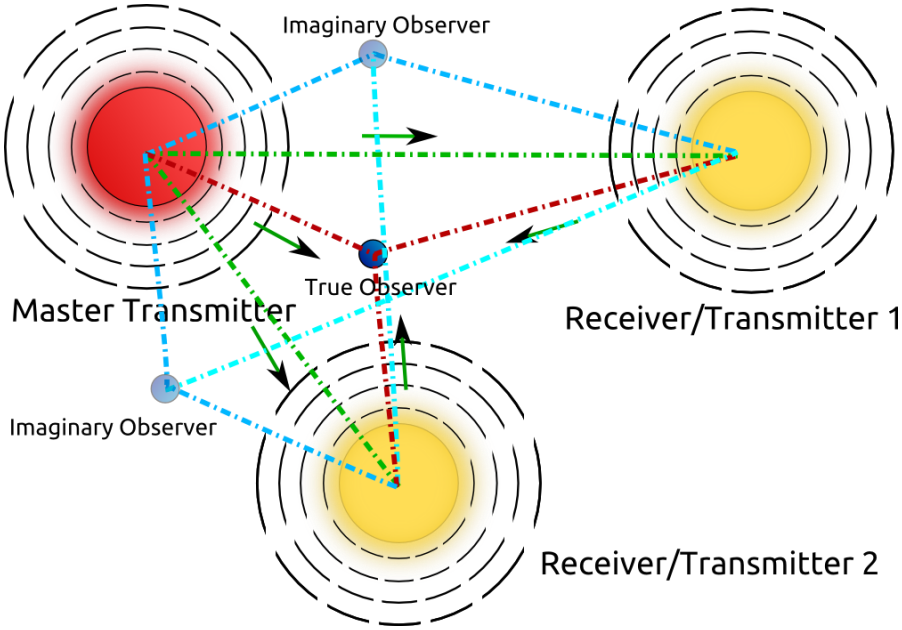

Fig. 12. Imaginary observers receives the same data as the true observer from only the repeaters, symmetrical about the lines joining them with the master transmitter. Using all three towers, though, only one solution, the true observer, satisfies all conditions.

The third tower shows, again, two possible solutions along an arc, which satisfy both the correct angle and distance. Only the real observer will satisfies the conditions of both equations, so the system then knows its position.

However, the angle does not even technically have to be calculated. One can simply plot their distance from each receiver in an arc on a chart with a pair of compasses, and the point where they all cross is his distance.

A way of improving this, perhaps, would be to use two towers on opposite sides of the equator. As it is fairly easy to determine what hemisphere you are (e.g. at night you could simply look for Polaris/Crux, see the orientation of the moon, measure the direction of the Coriolis force etc.), you would not even need a third tower, because you already know which solution is the correct one.

A more sophisticated method of transmitting location data is to use Very High Frequency omnidirectional range (VOR). As these are in the VHF band, the range is much shorter, so many more towers are required for such navigation. A single tower sends two signals out – the original master signal, and the same signal again out of phase, with each different phase corresponding to each direction, which determines the angle it was transmitted at.

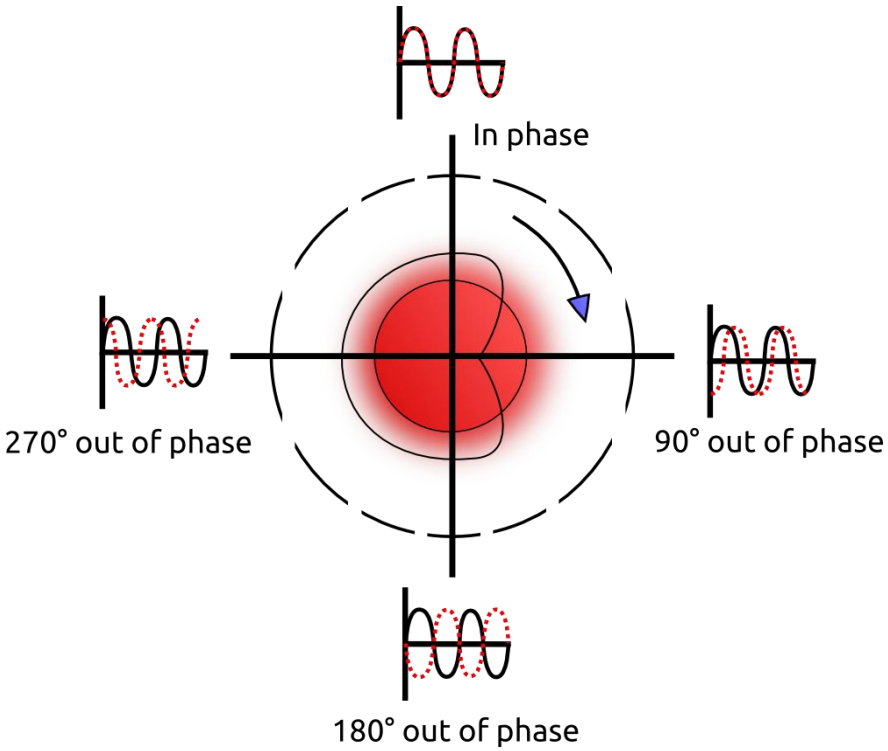

Fig. 13. The phase differences at the various radial directions.

This was initially achieved by mechanically rotating the second transmitter. The continuous wave would be sent in all directions, and the second broadcast would be constantly transmitted in a single direction, starting due North exactly in phase with the continuous transmission, and eventually rotating 360° to due North again. By measuring the time lag between the initial pulse and the second pulse, a receiver can calculate its bearing.

By receiving such signals from two towers it is possible to obtain two bearings, which, when plotted on a chart, will give the current location at the point of intersection.

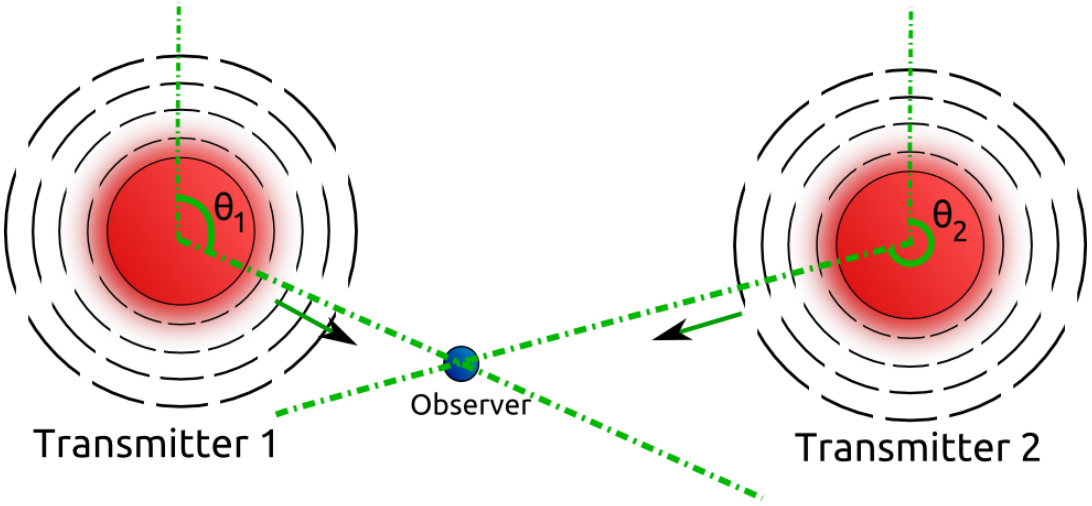

Fig. 14. The intersection between the lines provided by the bearing information transmitted by two radio masts shows the observer where he is.

This must be mapped onto a chart which reflects the curvature of the Earth accurately, or the position will not be correctly mapped. Generally, VOR transmitters will provide angled phase transmissions in steps of 1°, so the further from the towers a navigator moves, the less accurate the position will be.

An improved version of VOR is Doppler VOR (DVOR). Rather than mechanically rotating towers, it uses an electronically rotating signal, using an assortment of individual aerials arranged in a circle. The radial velocity of the circle causes a Doppler shift. As the transmission's direction is rotated around, and approaches the observer, the frequency will shift up. Likewise, as it rotates further, the frequency will shift down. In this way, the signal can give a much more accurate bearing, and is less prone to interference.

4. The Global Positioning System

The Global Positioning System is a military operation provided by the United States Department of Defence. Restricted access to the system is available to the general public. GPS blows all other navigation systems completely out of the water with its accuracy. The public can locate their position to an accuracy of *nine metres.* Military personnel can find their position to the order of *centimetres*.

GPS calculates position using a constellation of satellites. At the time of writing, the constellation contains 32 operational satellites[4]. Their orbits ensure that, at any time on any point on Earth, there will always be at least four available satellites. The satellites constantly such that their antennae face Earth and the solar panels which power them face the sun.

The satellites all transmit microwaves at two frequencies. Transmissions of the first frequency contain an embedded code, which can be access by the general public, known as the course acquisition (C/A) code. The other transmissions use a separate code, the precision (P) code, which only the US military have access to.

Fig. 15 – The GPS satellite constellation.[5]

These transmissions are encoded with their position, and timestamps, as measured by extremely accurate atomic clocks (measuring caesium vibrations) on the satellites, with several backup clocks. The time at which the signal is received is recorded by the user, but unfortunately most devices which receive GPS do not have the luxury of atomic clocks, so their recorded times are not nearly as precise. The calculated time difference multiplied by the speed of light is, therefore, known as the pseudorange to that satellite, as the receiver's clock is so inaccurate that it cannot calculate the true range.

This time difference, rather than simply being accepted as a fault of the system, is treated as an unknown value. This creates four unknowns – the three spatial coordinates, and the difference in time. Therefore, to solve these simultaneous equations, four different signals are required. This is the reason that the GPS constellation ensures that there are always four satellites available at any point on Earth. GPS navigation devices will use as many satellites as they can access, however, since this will increase the accuracy of their location.

GPS is handled in four sections:

- The satellite constellation, which broadcasts the encrypted information.
- The control station, which ensures that the times of the satellites are correctly synchronised, and that readings are as expected.
- The user hardware, which measures time, decrypts the signals, compares the timestamps of the various satellite messages, and on more advanced units, solves the equations itself to determine its position.
- The tracking stations on the ground, which, in most cases with civilian devices, handle the calculations for determining position, as well as restricting access.

The GPS satellites orbit in near circular orbits, with an elliptical eccentricity of 0.02. The primary 24 satellites each form six planes around the Earth each plane containing 4 satellites. These planes ensure that there will always be enough accessible satellites. The satellites orbit with a period of 11 hours and 58 minutes, completing two full orbits in just under a day. (The four minute difference is due to the sun's change in position in the sky as the Earth rotates around it).

From a fixed location, the same set of satellites will be used every day, as they pass overhead again, meaning that any errors or difficulties in picking up signal will repeat on a daily basis. Errors are dealt with by many stations, which monitor signal and report the information to the control station. These improve the transmitted information on position and time, and determine the future locations of the satellites. Corrections are transmitted to the satellites hourly, which apply them to their encoded transmissions. Satellites can also be remotely controlled, and shifted into different orbits.

The C/A code is a sequence, which repeats every millisecond. It is pseudo-random, that is, based on an algorithm that generates seemingly random numbers from an initial state (or seed). These transmissions are encoded with the time of

transmission. If this is wrong to a single millisecond, the calculated position will be incorrect by 293km. Every satellite has a unique code, so a receiver knows which satellite transmitted the information.

The P code, however, is only repeated once every 267 days. It works in the same way as the C/A code, but with a resolution ten times greater. A third transmission, the Navigation Message, broadcasts information about its orbit, clock corrections, atmospheric conditions and any technical faults.

The generated codes are a series of binary signals. The algorithm which produces them is a linear feedback shift register (LFSR). Each digit produced is based on a XOR relationship. For instance, in a three stage register, initially the first two values of a binary sequence A_N is initialised (e.g. $A_1 = A_2 = 1$). Each subsequent value in the sequence is determined by:

$$A_N = XOR(A_{N-1}, A_{N-3}),$$

such that 50% of the generated values are 1s and the other 50% are 0s, repeating the series after every 7^{th} digit. By knowing the equation and the two initial conditions (the values of $A_1$and A_2), anyone could then calculate all subsequent values of this sequence. By changing the formula so that:

$$A_N = XOR(A_{N-1}, A_{N-R}),$$

this becomes an R stage register, where the sequence length follows the formula:

$$S = 2^R - 1.$$

The C/A code uses an R value of 10, so the sequence is 1023 digits long. One would need the first 9 digits in the sequence in order to replicate the code. The method of creating the code is more complicated than a single XOR gate, however, as it uses two separate LFSRs to generate values, and feeds those values into another XOR gate in order to create the code. Each satellite can, therefore, have a unique code simply by changing the "phase" (as in, the difference in index number from each individual LFSR to be fed into the final XOR gate).

This code is then transmitted at 1.023MHz, so the whole sequence is transmitted exactly once every millisecond, corresponding to around 300km in difference.

The P code uses two registers also, but takes around 270 days to complete the sequence, where each week has its own code with which to identify the different satellites.

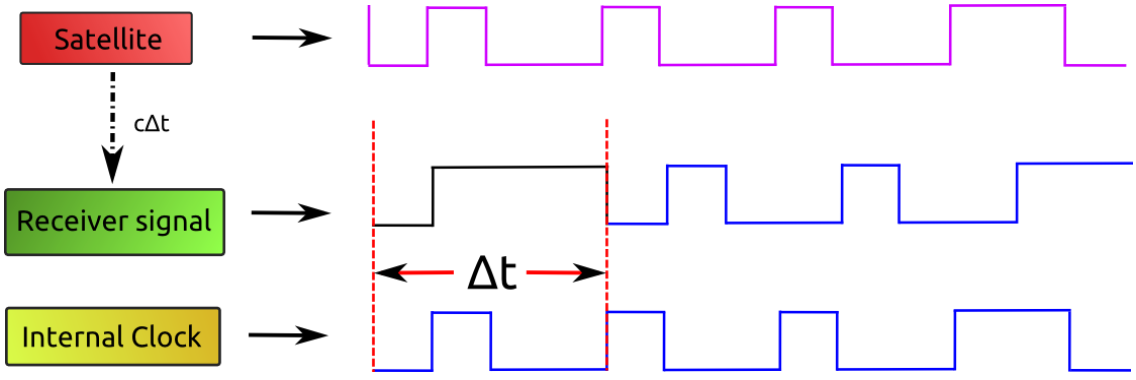

Fig. 16. The satellite and the internal clock of the receiver generate the codes simultaneously, and the time difference can be determined by comparing the two signals, and multiplied by c in order to find the pseudorange.

GPS receivers have their own replicas of the codes, each digit of which is associated with a certain time. It can compare the received digits to the digits it generates itself and find the time difference required to synchronise them. This is achieved through autocorrelation. The first bits of each sequence are multiplied together to give a new bit. Next the second bits of each sequence are multiplied together, and so forth. If the values are perfectly synchronised, the average value is equal to 1.

This time difference is subject to the error due to the lack of precision of the receiver, so this is fixed, as discussed before, by solving four simultaneously equations using data from at least four satellites.

There is also another issue involving time – relativity. The atomic clocks in satellites experience the effects of both special relativity and general relativity. The satellites are moving at around 4000m/s. Normally, such a non-relativistic velocity would be irrelevant in calculations, but because pinpoint accuracy is required, this is a really serious issue. Every day, due to special relativity, the time dilation experienced by the satellite would be around 7.3μs - around 2.7ms in a year. As 1 millisecond of difference equates to almost 300km, this is highly significant change. Timekeeping must be extra vigilant, therefore, using the Lorentz time dilation formula.

This, however, is minute compared to the effects of general relativity, which accounts for 38μs a day (or 13.9ms a year), as we on the surface of Earth are so close to such a massive body, which bends the spacetime around it. Gravitational time dilation is exponentially dependent on gravitational field strength, which is based on the inverse square law of gravitation. Therefore, our receivers experience far more time dilation than the satellites, which orbit at over four Earth radii.

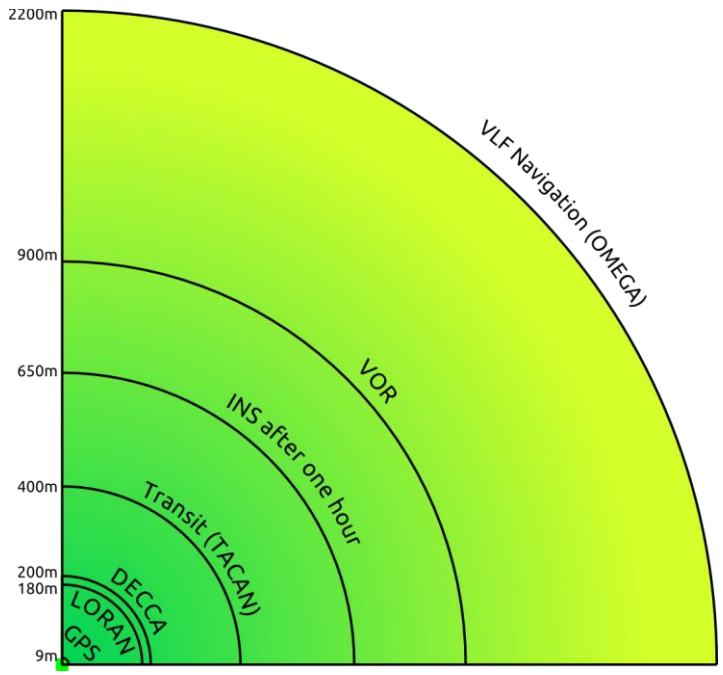

Fig. 17. The relative accuracy of the major navigation systems discussed. It is also worth noting that an expert with a professional sextant can successfully navigate to an accuracy of 460m.

5. Conclusion

As time has progressed and our knowledge of physics has increased, our position systems have become extremely sophisticated, increasing in precision by several orders of magnitude. Knowledge of such systems will be vitally important if the human race eventually achieves frequent space travel, and needs to find methods of charting positions not simply on a globe, but a galactic scale.

Despite how advanced some of our technology has become, it is comforting that we still have methods of navigation, like the sextant, which are reasonably accurate and entirely non electrical, and that we can also rely on mechanical systems. INS systems are still extremely useful in fast moving, self-contained systems, and not only this, but their principles can be extended to computer modelled physics engines.

It is also incredible that humans now all have access to a completely free system of pinpoint positioning, although, being run by the US Military, there is a fair amount of potentially interesting information (e.g. on how the P codes are generated) that a general member of the public doesn't have the clearance to access.

But regardless, if international relations go awry, we still have our advanced radio systems that work to a very accurate level, which when coupled with INS systems, are enough to guide any aircraft to its destination.

References

[1] *Industry Expects 800 Million More Travelers by 2014 - China Biggest Contributor* - IATA - http://www.iata.org/pressroom/pr/pages/2011-02-14-02.aspx [Date Accessed: 10 March 2013]

[2] *World Jet Fuel Consumption by Year* – Index Mundi - http://www.indexmundi.com/energy.aspx?product=jet-fuel&graph=consumption

[Date Accessed: 10 March 2013]

[3] *Sextant Diagram* - Wikimedia Commons - http://upload.wikimedia.org/wikipedia/commons/4/4a/Marine_sextant.svg

[Date Accessed: 10 March 2013]

[4] *Current GPS Constellation* – United States Naval Observatory http://tycho.usno.navy.mil/gpscurr.html [Date Accessed: 10 March 2013]

[5] *GPS Diagram* - http://www.pocketgpsworld.com/reviews/howgpsworks/gps1.jpg [Date Accessed: 10 March 2013]

Bibliography

Groves, Paul – "Principles of GNSS, Inertial, and Multisensor Integrated Navigation Systems" [2009]

"Inertial Navigation System", Department of Control Engineering, Aalborg University [2008]

Radio Navigation - Nordian [2009]

Woodman, Oliver – "An introduction to inertial navigation", *Technical Report* Number 696 [2007]