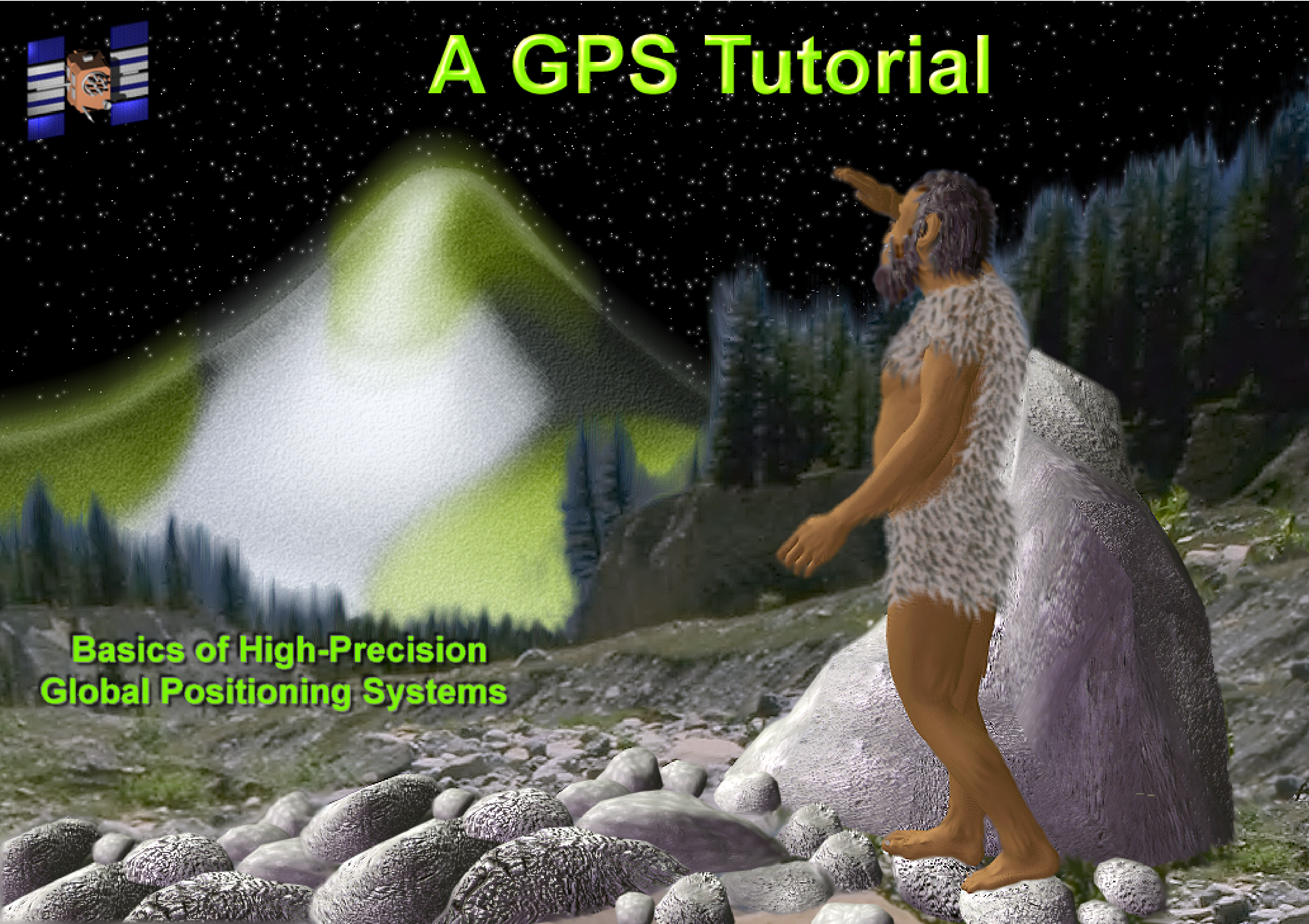
A GPS Tutorial



Basics of High-Precision Global Positioning Systems

Table of Contents

Chapter 1. From Stones to Satellites

Where am I? How do I get to my destination? These questions are as old as the history of mankind.

The Stone Age

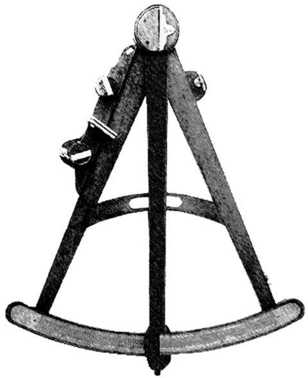
Identifying and remembering objects and landmarks as ***points of reference*** were the techniques that the early man used to find his way through jungles and deserts. Leaving stones, marking trees, referencing mountains were the early navigational aids. Stones, trees and mountains were the early examples of "points of reference", a concept that has evolved through times with the advent of (and the need for) more sophisticated techniques, objects and instruments.

The Star Age

Identifying points of reference was easy on land; but it became a matter of life and survival when man started to explore the oceans, where the only visible objects were the Sun, the Moon and the stars. Naturally, they became the "points of reference" and the era of celestial navigation began.

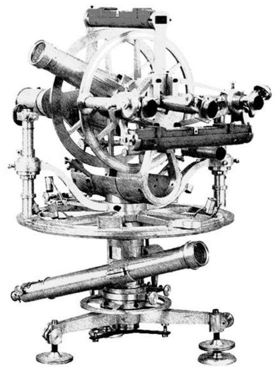
Celestial navigation was the first serious solution to the problem of finding one's position in unknown territories, where the Sun, the Moon and stars were used as points of reference. The relative position of stars and their geometrical arrangement look different from different locations on Earth. Therefore, by observing the configuration of stars one could estimate his position on Earth and the direction that he should take for his destination. The Great Bear and Small Bear constellations are two examples. The geometrical configurations of stars from the observer's point of view were more accurately determined later by measuring the relative angles between them. For better accuracy, special optical instruments were invented to measure the angles of view between stars. These measured angles were then used to determine the position of the observer with the aid of published pre-calculated charts that eased the tedious computation task.

The process of measuring the angles of the stars with optical instruments was time-consuming and inaccurate. It could not be used during the day or on cloudy nights. The measured angles had to be transferred to special charts and after tedious calculations, the derived position was good only to about several miles.



The calculation process was the basic triangulation geometry, where the stars became the known points of reference and the measured angles between them and the navigator would solve for the triangles' components and determine the navigator's position.

The triangulation would have been simpler if distances to the stars could have been measured also. In fact, they could have been used to solve for the triangle's components instead of angles, but such measurements were not possible.



In frustrated moments of trying to determine a position, many navigators must have dreamed, conceivably, of gadgets that would do such a task automatically and more accurately. There were probably people that pictured a device, or even worked on building one, that aligned itself with stars quickly, measured angles to these points of reference and computed its position automatically.

The idea of automatic computation of position through measurement of distances to points of reference became a reality only recently when radio signals were employed and the age of radio navigation began.

The Radio Age

About the middle of this century, scientists discovered a way to measure distances using radio signals. The concept was to measure the time it took for special radio signals to travel from a transmitting station to a special device designed to receive them. Multiplying the signal travel time by the speed of the signal gives the distance between the transmitter and the receiver. The speed of radio signals is the same as the speed of light — about 300,000,000 meters per second (about 186,500 miles per second). Accurate measurement of signal travel time is important since one microsecond (one millionth of a second) of error in measuring the travel time is equal to 300 meters of error in distance. For precision positioning, therefore, the receiving radio should be able to measure the travel time much better than one millionth of a second, perhaps to one billionth of a second (one nano-second).

How could such a radio signal transmitter-receiver system be used to determine a person's location?

Assume that a transmitting tower is installed at a known point, A, on the earth and we have a special radio that can receive signals from transmitter A and measure the distance to the transmitter. The exact location of point A is programmed in our special radio receiver. We are in some unknown location. We turn on the receiver and measure our distance to the transmitter as 12,325 meters. This does not tell us where we are, but it narrows our position to a point on a circle with the radius of 12,325 meters around the transmitter, as shown in the Figure 1.

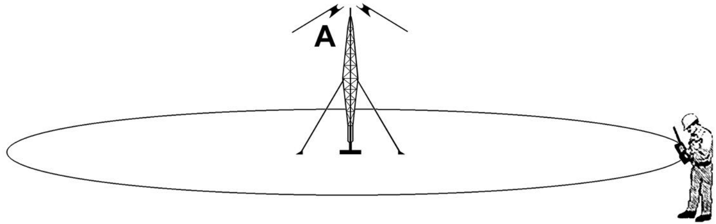


Figure 1

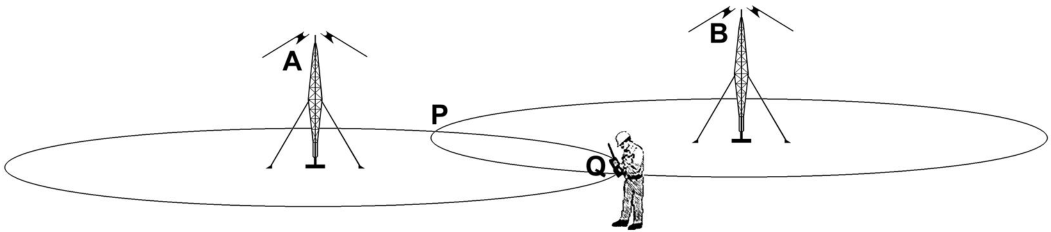


Figure 2

Next, assume that a second transmitter tower is installed at another known point, B, on the earth. The same special receiver measures our distance to transmitter B as 9,792 meters. This tells us that we are somewhere on a circle with the radius of 9,792 around the transmitter B. Now we have two pieces of information: our distance to point A is 12,325 meters and our distance to point B is 9,792 meters. So we are on circle A and circle B at the same time. We must be at the intersection of the two circles, one of the two points P or Q shown in Figure 2.

Measuring our distance to a third transmitter C would identify exactly where we are. Now you can imagine how the system works: we turn on our special radio receiver; it quickly measures the distances to transmitters A, B, and C and will compute our location. Remember that the exact location of transmitters A, B, and C were previously programmed in our special radio receiver. Transmitters A, B, and C together are called a transmitter "chain". A chain may have four or more transmitters in order to have better coverage. The range of a radio transmitter is about 500 kilometers.

Navigational systems that use such radio signals to measure distances to several transmitting towers located at known points are called radio navigation systems.

The LORAN Age

LORAN (LOng RAnge Navigation) is one such radio navigation system that became operational around 1950. Each LORAN chain consists of at least four transmitters and typically covers areas of about 500 miles. To provide LORAN coverage for larger areas, several LORAN chains are used. For example, two LORAN chains cover the West Coast of the United States.

Each LORAN transmitter chain broadcasts radio signals on its own designated frequency. A LORAN receiver tunes in to the radio signals of the transmitters of the chain, measures distances to them automatically, and computes the position of the receiver. A LORAN receiver has the exact locations of all LORAN transmitter chains in its database. In a journey, one may pass through several LORAN chains. So, the navigator needs to know and tune in to the frequency of each LORAN chain he is passing through; in the same manner that one needs to change the frequency of an FM radio when leaving the coverage area of one FM station and entering into a coverage area of another.

The entire operational LORAN chains worldwide cover only a small portion of the earth. They are operated by local governments and are generally situated near coastal areas that have high traffic volume.

Although LORAN was a major breakthrough for navigation, it has the following shortcomings:

- LORAN coverage is limited to about 5% of the surface of the earth where the chains are established. It is not a global system.
- LORAN transmitters send out signals along the surface of the earth and can therefore provide only two-dimensional position information (latitude and longitude). It cannot provide information about height and, for example, cannot be used in aviation to provide altitude.

In general, the accuracy of LORAN is good to only 250 meters.

The Satellite Age

To overcome these limitations, satellite-based radio navigation systems were conceived in which improved radio transmitters were put aboard satellites orbiting the earth at high altitudes to give wider coverage. This is similar to the concept of a local TV station versus satellite TV. You can receive your land-based local TV station only in your city and within a distance of no more than 100 miles, while a satellite TV transmitter can cover an area as wide as the US (about 3000 miles). Signals from navigation satellites can cover large areas of the earth, and several satellites can cover the whole planet.

The theory behind the operation of the satellite navigation systems is similar to that of the land-based systems. In land-based navigation systems, the transmitting towers are the reference points located on the earth and the distance to them is measured by the receivers to compute the two-dimensional position

(Latitude and Longitude or X and Y) by finding the intersection of several circles. In satellite-based systems, the satellites act as the reference points and the distance to them is measured to determine the three-dimensional position (Latitude, Longitude, and Altitude or X, Y, and Z) by finding the intersection of several spheres.

In land-based systems, the location of the transmitting towers are fixed, accurately known, and stored in the data base of the receivers. In satellite-based systems, the locations of the satellites are not fixed. They orbit the earth at high speeds. However, satellites have a mechanism of giving information about their location at any instant in time. The accuracy of the calculated location of the satellites, at the time at which distances to them are measured, affects the accuracy of calculated position of the receiver. In other words, the accuracy in computing our position depends on the accuracy in computing the location of our reference points.

In a positioning satellite system, satellite locations and their orbits are continuously monitored from several observation centers around the world by the organization responsible for keeping the orbit of the satellite within acceptable boundaries. This organization also predicts the orbit of the satellite for the next 24 hours based on the actual orbit information received by the observation posts for the previous 24 hours (similar to weather predictions). The predicted orbit information for the next 24 hours is relayed to each satellite by the control organization, so that it can be sent to receivers. Satellites broadcast their orbit information as part of their radio signal structure.

With satellite systems, we are once again "looking" up to the sky. This time, however, we are "looking" at man-made objects instead of stars. And unlike celestial navigation utilizing stars, man has now devised a scheme (using radio signals and receivers) to measure distances to reference points.

One of the first satellite navigation systems was Transit. The experience gained from Transit and several other experimental systems led to the development of the current Global Positioning System (***GPS***) by the United States of America and GLObal Navigation Satellite Systems (***GLONASS***) by the Russian Federation. GPS and GLONASS are very similar, as we will discuss in the rest of this book.

Summary

In this chapter, we learned:

- *Stones, trees, and mountains were the early examples of "points of reference".*
- *Celestial navigation was the first serious solution to the problem of finding one's position in unknown territories, where the Sun, the Moon and stars were used as points of reference.*
- *Automatic computation of position, through measurement of distances to points of reference, became a reality only recently, when radio signals were employed and the age of radio navigation began.*
- *LORAN coverage is limited to about 5% of the surface of the earth where the chains are established. It is not a global system.*
- *Signals from navigation satellites can cover large areas of the earth, and several satellites can cover the whole planet.*

Chapter 2. Navigation Satellites

In the previous chapter we discussed that if our distances to several satellites of known locations could be measured, then we can compute our position. We have put forward three major prerequisites:

1. Satellites — We need satellites as our points of reference to which we can measure our distances. At any given time, we also need to know the exact location of each satellite.

    A satellite by itself is nothing more than a vehicle. You may call it a space vehicle. The function of each satellite depends on what equipment it carries. If the equipment on board is something like a TV station, then it becomes a TV satellite. If the equipment on board is weather observation equipment, then it becomes a weather satellite whose function is to prepare large-scale images of clouds, storms, hurricanes, etc. Obviously, we need special satellites for our purpose.

2. Coordinate System — How do we express the location of each satellite and how do we express our position? As we recall from algebra, We can express them with sets of numbers related to some coordinate system. We need to define the coordinate system such that they are recognizable by everyone. Thus, we need a universal coordinate system.

3. Distance Measurement — What type of electronic signals must the satellites emit to enable us measure our distances? How do we measure these distances? How accurately can we measure these distances?

And how do these measurement errors relate to the accuracy of our calculated position? We will attempt to answer these questions in the remainder of this chapter.

Measuring Distances to Satellites

Time Is Distance

Have you noticed that during a thunderstorm, you hear the sound sometime after you see the light? The reason is that sound waves travel much slower than light waves. We can estimate our distance to the storm by measuring the delay between the time that we see the thunder and the time that we hear it. Multiplying this time delay by the speed of sound gives us our distance to the storm (assuming that the light reaches us almost instantaneously compared to sound). Sound travels about 344 meters (1,130 feet) per second in air. So if it takes 2 seconds between the time that we see the lightning and the time that we hear it, our distance to the storm is 2 x 344 = 688 meters. We are calculating the distance to an object by measuring the time that it takes for its signal to reach us.

In the above example, the time that we see the lightning is the time that the sound waves are generated in the storm. Then we start to measure the delay until the time that we hear the sound. In this example, the light is our start signal. What about the cases for which we don't have a start signal? Consider the next example.

Codes and Patterns

Assume that your friend at the end of a large field repeatedly shouts numbers from 1 to 10 at the rate of one count per second (10 seconds for a full cycle of 1 to 10 count). And assume that you are doing the exact same thing, synchronized with him, at the other end of the field. Synchronization between you and him could have been achieved by both starting at an exact second and observing your watches to count 1 number per second. We assume that you both have very accurate watches. Because of the sound travel time, you will hear the number patterns of your friend with a delay relative to your patterns. If you hear your friend's count with a delay of one count relative to yours then your friend must be 344 meters away from you (1 sec x 344 meters/sec = 344 m). This is because the counts are one second apart.

Now assume that you and your friend count twice as fast, two counts in one second. Then at the same distance between you and your friend you will hear a two-count delay. This is because now each count takes 0.5 seconds and each count delay measures 172 meters. If you could count 100 times faster then each count would take 0.01 seconds and each count delay between you and your friend would measure the distance of 3.44 meter. Counting faster is like having a ruler with finer graduation. Of course in real world, you need appropriate devices and instruments to generate and receive very fast counts.

Next assume that you and your friend are far apart and counting very fast, say each count in 0.01 second (each delay count is 3.44 meters), and, as before, both are repeatedly counting from 1 to 10. Assume when you say 7 you hear your friend's voice say 5. You hear a delay count of 2 but you know your distance is more than 6.88 meters. This is because the delay is not just only 2 counts, but rather 2 counts plus some multiples of 10 counts (i.e. some multiples of the pattern cycle). This is as if your measuring tape is not long enough and there are some multiples of the full length of measuring tape plus some fraction. We refer to this unknown number of full pattern delays as ***unknown integer***. If you and your friend were to count repeatedly from 1 to 1000 (instead of 1 to 10) then you could hear 212 count delays between the numbers that you hear and your numbers, which would produce the distance of 212 count delays x 3.44 meters = 729.28 meters. This is 21 full cycles of the 1-to-10 pattern, plus 2 counts. The number of full cycles, 21, that we were not able to observe with our short pattern is our unknown integer.

What we demonstrated above are the concepts of *pattern granularity* (fineness of tape marks) and *pattern length* (tape length).

The concept of measuring distances to satellites is much like what we discussed above, but satellites transmit electronic patterns rather than voice counts. Likewise, our receiver generates similar electronic patterns for comparison with the received patterns from satellites in order to measure the distances to them.

Satellites generate two types of patterns: One has a granularity of about 1-millimeter and a length of about 20 centimeters. The other has a granularity of about 1 meter and effectively an unlimited length. In satellite terminology, the first pattern is called ***"carrier"*** and the second is called ***"code"***. The distance measured by carrier is called ***"carrier phase"*** and the distance measured by code is called ***"code phase"***. Because code pattern is long, the code phase measurements are complete and do not have any unknown integer. We can measure our distance to a satellite as 19,234,763 meters, for example. In contrast, the carrier pattern is short and carrier phase has a large unknown integer. You may think that it is useless to say, for example, that our distance to satellite is 13.2 centimeters plus an unknown number of carrier cycles. The unknown integer is in the order of several tens of millions. You may ask what good will it do to measure the fractional part so accurate when millions of full cycles are missing? We will explain more.

Initial Unknown Integer (Integer Ambiguity)

In the previous counting example with a short pattern, assume that you and your friend are standing next to each other and synchronized together counting fast from 1 to 10. You hear no delay because you are standing next to each other. Then your friend starts to move away. The count delays start to grow from 0 (no delay) to 9. After it reaches 9 it will drop back to 0. This is actually 10 and not zero. You know that this is the case (that the zero count delay actually represents one full cycle count) because you have been following the count delays continuously. You will keep in mind, as your friend moves away, to count the whole number of cycles that are being added to your distance. In this case, there is no unknown integer as long as you keep track of him continuously.

If, instead of starting next to each other, you start at some unknown distance, then you are starting from an unknown integer of cycles. However, if after starting your friend moves away from or towards you, you can account for the number of full cycles that must be added to or subtracted from the ***initial unknown integer***. All the distances that you measure every second contains the same initial unknown integer. This is true as long as you keep track of him continuously. If you don't hear him for some period of time, then you don't know how many full cycles he moved and you will have to start with another unknown number of cycles. The point is that as long as you keep track of him you have only one initial unknown integer.

The concepts of code and carrier are very important. Let us use another analogy for better understanding. You may consider that code phase is like a watch that only has an "hours" hand (call it code watch). At any time you can look at this watch and know the time of the day approximately. You may consider carrier phase like a watch that only has a "seconds" hand (call it carrier watch). You can keep track of the elapsed time with this watch with the accuracy of one second as long as you monitor the watch continuously to keep track of the elapsed full minutes. If you somehow can determine the number of full minutes initially (the initial unknown integer when you started looking at this watch) then you can keep track of time very accurately. If you get distracted and lose track of the number of minutes, then you have a new "initial unknown integer" that you somehow must determine again. With code phase watch, you always get the time of the day instantly but with the accuracy of not better than 10 minutes by estimating the location of the hour hand. The code watch can narrow the estimate of unknown minutes (integers) of the carrier watch to plus or minus few minutes. You see that there is a gap between the seconds hand and the hours hand. We are missing the minutes hand. GPS manufacturers have developed techniques to narrow the gap such that code phase and carrier phase can make unambiguous and accurate distance measurements as fast as possible. We will explain the reason for the gap later.

The good news is that the integer ambiguity of carrier phase can be determined by tracking satellites for some period of time. This is the fundamental concept in precision applications like geodesy.

With carrier phase, tracking the correct number of full cycles that the distance to satellite is changing is very critical. You will miscalculate this number if you miss a cycle or add an extra cycle. In GPS terminology, this is called a ***"cycle slip"***. In our previous example, cycle slips can happen if you don't hear your friend's voice correctly due to noise or other effects, or if he suddenly jumps a very long distance. Cycle slips is like missing the meter marks while you are concentrating on reading the millimeter ticks. It can create large errors. Most GPS systems are able to detect and repair cycle slips.

Note that not all receivers can measure carrier phase. Carrier phases are typically used in high precision receivers.

We can measure the distances to the satellites with the accuracy of 1 meter with code phase and 1 millimeter with carrier phase. This does not mean that we can determine our position with a GPS receiver with the accuracy of one meter or one millimeter. Several sources introduce inaccuracies into the GPS measurement. We will discuss them in the next Chapter.

Summary

In this chapter, we learned:

- *We can calculate the distance to an object by measuring the time that it takes for its signal to reach us.*
- *Satellites generate two types of patterns, carrier and code. Carrier has a granularity of about 1-millimeter and a length of about 20 centimeters. Code has a granularity of about 1 meter and effectively an unlimited length.*
- *The integer ambiguity of carrier phase can be determined by tracking satellites for some period of time.*

Chapter 3. Sources of Inaccuracy: The Problems

In our discussion of measuring distances using patterns of numbers, we assumed that you and your friend started counting numbers simultaneously. If your watch is one second off, this will translate into 344 meters of error in measuring distance, because sound travels 344 meters in one second. With satellites, the electronic signals travel about 300,000,000 meters per second (the speed of light). So the errors in the satellite clock and the receiver clock contribute profoundly to errors in distance measurements.

Satellite Clock

One billionth of a second (one nanosecond) of inaccuracy in a satellite clock results in about 30 centimeters (one foot) of error in measuring the distance to that satellite. For this reason, the satellites are equipped with very accurate (Cesium) atomic clocks. Even these very accurate clocks accumulate an error of 1 billionth of a second every three hours. To resolve the satellite clock drifts, they are continuously monitored by ground stations and compared with the master control clock systems that are combinations of more than 10 very accurate atomic clocks. The errors and drifts of the satellites' clock are calculated and included in the messages that are transmitted by the satellites. In computing the distance to the satellites, GPS receivers subtract the satellite clock errors from the reported transmit time to come up with the true signal travel time.

Even with the best efforts of the control centers in monitoring the behavior of each satellite clock, their errors cannot be precisely determined. Any remaining satellite clock errors accumulate typically to about a few nanoseconds, which cause a distance error of about one meter.

Receiver Clock

Similar to satellite clock errors, any error in the receiver clock causes inaccuracy in distance measurements. However, it is not practical to equip receivers with very accurate atomic clocks. Atomic clocks weigh more than 20 kilograms, cost about US$50,000, and require extensive care in temperature control.

Assume that at a given time our receiver clock has an error of one millisecond, causing a distance error of about 300,000 meters. If the distances to all satellites are measured exactly at the same time, then they are all off by the same amount of 300,000 meters. We can, therefore, include the receiver clock error as one of the unknowns that we must solve for. Remember from Chapter 1 that we had three unknowns (X, Y, Z) for the position. Now we have four unknowns: three components of position and the new unknown of receiver clock error. We will need four equations in order to solve for the four unknown. Measuring distances to four satellites can provide us such four necessary equations. Instead of three satellites before, now we need four, but in return we can use inexpensive clocks in our GPS receivers.

Note that the concept of receiver clock being one of the unknowns is valid only if we take measurements to all satellites exactly at the same time. If distances to all satellites are not measured at the same time, then for each measurement we have a different clock.

Making simultaneous measurements to four satellites, we not only compute the three dimensions of our position, but we also find the error in our receiver clock with very good accuracy. A typical clock has a drift of about 1000 nanoseconds every second, but we can now adjust the receiver time to the accuracy of the GPS clock. This will make the inexpensive clock of the receiver as good as an atomic clock. Receivers correct their clock every second and provide a corrected tic signal for outside use for those who need accurate time. If we put a receiver in a precisely known location, then we need to track only one satellite to continuously calculate the receiver clock error and adjust it.

Four is the minimum number of satellites that we need to compute position and time. The more satellites we have the more accurate results we can get. This is discussed later in the GDOP section.

Satellite Orbit Error

As we discussed before, the accuracy of our computed position also depends on how accurately we know the location of the satellites (the points of references). The orbits of satellites are monitored continuously from several monitoring stations around the earth and their predicted orbital information is transmitted to the satellites, which they in turn transmit to the receivers. The history of GPS has shown, thus far, that the accuracy of the orbital prediction is in the order of a few meters. This will create about a few meters of error in computing our position. In the next chapter we shall see how to remove this error.

Atmospheric Errors: Ionosphere and Troposphere

Ionosphere

In computing distances to satellites, we first measure the time it takes for the satellite signal to reach the receiver and then we multiply this by the speed of light. The problem is that the speed of light varies due to atmospheric conditions. The upper layer of the atmosphere, called the ionosphere, contains charged particles that slow down the code and speed up the carrier.

The magnitude of the effect of the ionosphere is much more during the day than during the night. The magnitude also has a cyclical period of 11 years that reaches a maximum and a minimum. For the current cycle, the ionosphere will reach its peak magnitude in 1998 and its minimum in 2004. The cycle will then be repeated. The effects of the ionosphere, if not mitigated, can introduce measurement errors greater than 10 meters.

Some receivers use a mathematical model for the effects of the ionosphere. With the approximate knowledge of the density of the charged particles in the ionosphere (broadcast by satellites), the effect of the ionosphere can be reduced by about 50%. The remaining error is still significant.

The impact of the ionosphere on electronic signals depends on the frequency of the signal. The higher the frequency, the less is the impact. So if we transmit the patterns simultaneously via two different frequencies, the ionosphere may delay the code on one frequency, for example, by 5 meters and on the other frequency, say, by 6 meters. We cannot measure the magnitude of these delays, but we can measure their difference by observing the difference on their arrival time, which in this case translates into 1 meter of effective distance between them. By measuring this difference and using known formula for frequency dependency of the ionosphere delay, ionosphere effect can be removed.

It is exactly for this reason that all GPS satellites transmit information in two frequencies, called ***L1*** and ***L2***. Precision receivers track both signals to remove the effect of the ionosphere. All non-precision receivers track only the L1 signal. This is one of the main distinguishing features between different types of receivers. The L1 receivers are also called single frequency receivers, while the receivers that track L1 and L2 are called dual frequency receivers. Dual frequency receivers practically remove the ionosphere effects.

Since the L2 signal is not entirely available to the general public, sophisticated techniques have been implemented in receivers to extract the code and carrier information, even with the partial availability of the L2 signal. These techniques fully satisfy the requirements of the users for non-military applications, while not compromising the Anti Spoof policy and security objectives of the US Department Of Defense (DOD).

There has been some discussion on allowing a different frequency for civilian applications to separate the DOD and civilian requirements. We believe, however, that the existing system fully satisfies the civilian requirement, particularly since advancements in electronics integration have made the technology affordable for broad civilian applications.

Troposphere

The lower level of the atmosphere, which contains water vapors, is called the troposphere. It has the effect of slowing down both code and carrier. The effects of the troposphere cannot be removed using dual frequency systems. The only way to remove the effects of the troposphere is by measuring its water vapor content, temperature and pressure, and applying a mathematical model that can compute the delay of the troposphere.

Multipath

In measuring the distance to each satellite, we assume that the satellite signal travels directly from the satellite to the antenna of the receiver. But in addition to the direct signal, there are reflected signals, from the ground and the objects near the antenna, that also reach the antenna through indirect paths and interfere with the direct signal. The compound signal creates an uncertainty about the true signal arrival time, much the same way as the echo from nearby mountains may cause uncertainty in the exact time you hear your friend's voice. If the indirect path is considerably longer than the direct path (more than 10 meters) such that the two patterns of signals can be separated, then the multipath effect can be substantially reduced by signal processing techniques.

Receiver Errors

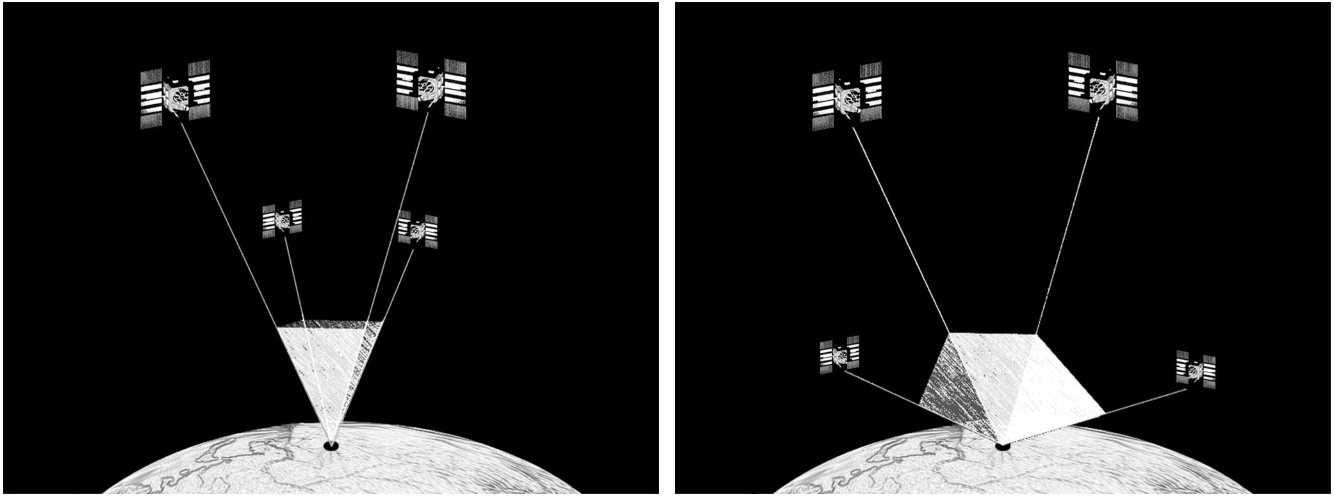
Receivers may introduce some errors by themselves in measuring code or carrier. In high quality receivers, however, these errors are negligible (less than one millimeter) for carrier phase and a few centimeters for code phase.

Geometric Dilution Of Precision (GDOP)

We have been talking about the errors in measuring distances to satellites, which are commonly referred to as ranging or range errors. The question is what is the relationship between the range error and the error in computed position. Or, in other words, how many meters of error are introduced in our computed position as a result of one meter of error in measuring distances to the satellites?

The answer is that it depends on the number and the geometry of the satellites used. If four satellites are clustered near each other, then one meter of error in measuring distance may result in tens or hundreds of meters of error in position. But if many satellites are scattered around the sky, then the position error may be less than 1.5 meters for every meter of error in measuring distances. The effect of the geometry of the satellites on the position error is called Geometric Dilution Of Precision (***GDOP***), which can roughly be interpreted as the ratio of the position error to the range error.

Imagine the tetrahedron that is formed by lines connecting the receiver to each satellite used. The larger the volume of this tetrahedron, the smaller (better) the GDOP. In most cases, the larger the number of satellites the smaller the GDOP.



Selective Availability: The Man-Made Errors

Errors in the satellite clock, the satellite orbit, the ionosphere, the troposphere, the multipath, and the receiver typically amount to less than 10 meters of range error which, under typical GDOPs of about 2, results in a position accuracy of about 20 meters.

The US Department of Defense has determined that providing this level of precision to the general public is against the US national interest. Therefore, DOD has introduced man-made intentional errors to degrade the position accuracy of GPS to about 100 meters. This intentional degradation is called ***Selective Availability (SA)*** and is implemented by tethering the satellite clocks and reporting the orbit of the satellites inaccurately. Military receivers are equipped with special hardware and codes that can mitigate the effect of SA. SA can be turned ON or OFF through ground commands by the GPS system administrators.

Summary

In this chapter, we learned:

- *We can adjust the receiver time to the accuracy of the GPS clock. This will make the inexpensive clock of the receiver as good as an atomic clock.*
- *Four is the minimum number of satellites that we need to compute position and time. (But the more satellites we have the more accurate results we can get.)*
- *Dual frequency receivers practically remove the ionosphere effects.*
- *If the signal indirect path is considerably longer than the direct path, more than 10 meters, then the multipath effect can be substantially reduced by signal processing techniques.*

Chapter 4. Sources of Inaccuracy: The Cures

Neither the accuracy of 100 meters (SA On) nor the accuracy of 20 meters (SA Off) is enough for many civilian applications. Since the inception of GPS, methods have been, and are still being, developed to reduce errors and enhance the accuracy, even with the implementation and presence of SA and partial availability of L2 (AS).

Differential Mode

Assume you have two receivers not too far from each other. The errors due to the satellite clock, the satellite orbit, the ionosphere, the troposphere and SA affect both receivers the same way and with the same magnitude. If we knew the exact location of one receiver, we could use that information to calculate errors in the measurement and then report these errors (or correction values) to the other receiver, so that it could compensate for them. This technique is called Differential mode.

The receiver in the known location is called ***"base receiver"*** and the other receiver with unknown location is called ***"rover receiver".*** The base receiver computes its instantaneous range to each satellite, based on its known position and the instantaneous location of each satellite. Then it compares each calculated value to its measured range for the corresponding satellite. The difference between the two is the range error (or correction value) for the corresponding satellite, which is reported to the rover receiver. The rover receiver subtracts the reported correction values from its measured ranges for all corresponding satellites and computes its own position with much better accuracy.

Due to the motion of the satellites and changes in their clocks, correction values change rapidly with time. Therefore the base receiver must quickly compute the range errors and transmit them to the rover.

Note that the accurate knowledge of the position of the base directly impacts the accuracy of the position computed by the rover. If we enter a position for the base receiver that is off in some direction, then all range errors computed and transmitted by the base receiver will be off in such a way that the computed rover position will be off by the same amount and in the same direction as the base.

The distance between the base and rover receivers is called "baseline". When the baseline is small, i.e. when the receivers are very close to each other, the range errors for the two receivers are nearly identical; therefore, we could use the range errors calculated by the base to correct for the rover position. As the baseline gets longer, the correlation between the range errors becomes weaker. In other words, there will be some residual errors in the computed position of the rover that depend on its proximity to base. As a rule of thumb, you can expect an additional one millimeter of error or uncertainty for every kilometer of baseline when dual frequency receivers are used. This is abbreviated as 1 ppm (one part per million). For single frequency receivers this error increases to 2 ppm.

The differential mode will remove most of all errors except multipath and receiver errors. These errors are local to each receiver and will not be canceled by the differential mode.

The receiver error (or noise) is typically about 10 cm for the code phase and about 1 mm for the carrier phase. In high quality receivers, these errors are even smaller by several times. The multipath error, on the other hand, could be as much as several meters for the code phase and several centimeters for the carrier phase. Therefore, if we somehow deal with the multipath errors, we can obtain millimeter level accuracy with carrier phase and decimeter accuracy with code phase. How to deal with multipath errors is a separate topic that we will cover in Chapter 5; for now let us assume that multipath errors can somehow be mitigated and continue our discussion.

Historically, differential mode with code phase has been called DGPS and with carrier phase called ***Carrier Phase* Differential (CPD)**. Real-time carrier phase differential has been called ***Real-Time Kinematic (RTK)***. In non-real time applications, terms like static or kinematic have been used. Next, we discuss DGPS and RTK. These discussions apply to non-real time applications the same way.

In carrier differential, computations are much more complex because we have to deal with the additional unknowns of initial full cycles. It may take several minutes to determine the full cycle ambiguities. After determining the initial full cycle ambiguities, then each subsequent position computation is instantaneous. As long as a minimum of four (or for reliability five) satellites are tracked, the loss of lock to other satellites will not disturb the continuity of position computation. But as soon as we end up with less than four satellites, then we need to resolve the initial full cycle ambiguities again when we re-lock to satellites, which can take up to several minutes. With code DGPS results are instantaneous, but not as accurate.

DGPS

In DGPS, if range errors are transmitted from the base receiver to the rover receiver in real-time, (i.e. with a radio link) then the system is called real-time DGPS in which accurate results can be obtained in real time. This is desirable for applications in which some actions need to be performed in the field, such as placing markers or moving objects to exact locations. If real time results are not needed (e.g. for making accurate maps), the measurements are time tagged and recorded in the base and rover receivers and later transferred to a computer to calculate the accurate position of the rover the rover at each instant. This is called post-processed DGPS.

DGPS is based on measuring distances to satellites with code phase. Code phase is like a measuring tape that has tic marks and numbers only every meter. The meter-marks and numbers of this measuring tape appear instantly after we lock to satellites, therefor we can measure distances instantly but not accurately.

RTK

RTK is based on measuring distances to the satellites with carrier phase. Another analogy of carrier phase is that it is like a measuring tape that has meter-marks and millimeter-marks. But with this measuring tape the meter-mark numbers do not appear instantly when we lock to satellites. We have to wait for the meter-mark numbers to appear and become clear (like a Polaroid picture) to be able to measure the distances. This is the time that we have to wait to determine the "initial unknown integers". The more we wait the meter-mark numbers become more clear (like a Polaroid picture). When meter-mark numbers become clear they remain clear and we can make instantaneous accurate measurements repeatedly until we lose lock to satellites in which case the meter-marks disappear again. When this happens, we have to wait for them to re-appear after we re-lock to satellites.

You can re-lock to satellites quickly and resolve their integer number immediately if you maintain tracking of at least five satellites.

When satellite interruptions are very brief, receiver may be able to continue based on the integer estimation that it had before.

Estimating the integer numbers incorrectly, or having cycle slips, is like reading the wrong number on the meter-mark. You can imagine examples when you measure something to be 3.874 meter while it actually was 4.874 meter. You read the millimeter-marks very accurately but misread the meter-mark number.

After the receiver resolves the ambiguities correctly, the accuracy of each position computation is between 0.5 to 2 cm horizontal and 1 to 3 cm vertical (depending on antenna multipath rejection capability) plus 1 ppm for double frequency and 2 ppm for single frequency. All RTK accuracy specifications from all manufacturers are within this range. They are all based on the assumption that the integers are estimated correctly.

Resolving the integers correctly is the key in RTK. The big question here is how long it will take to resolve the integers reliably after satellites are locked. If they are resolved incorrectly, it is like reading the meter-number wrong but continue to concentrate on reading the millimeter marks.

How long do I have to wait for the ambiguities to be estimated correctly and with a good assurance?

For short baselines (less than 10 kilometer) the time that you have to wait depends upon:

— The level of assurance (or confidence level) that you require for correct integer estimation.

— The number of satellites.

— Whether you have a single or a dual frequency receiver.

— The strength of the multipath signal (reflection coefficient of ground).

— Multipath mitigation characteristic of the antenna.

The sensitivity of time to resolve the ambiguities to the above factors are described below quantitatively, in order to help you understand the relative importance of each factor:

— Everything else being equal (i.e. for a given set of numbers for all other factors), it may require 10 seconds to resolve ambiguities with 99% confidence but 100 seconds to get to 99.9% confidence.

— Everything else being equal, it may require 1 second to resolve ambiguities with 15 satellites but 100 seconds with 8 satellites. It may take 10 minutes if there are only 6 satellites.

— Everything else being equal, it may require 10 seconds to resolve ambiguities using dual frequency receivers but 2 minutes with single frequency. Having dual frequency for short baselines is like having 50% more satellites. For example if you track dual frequency of 6 satellites, it is like tracking 9 single frequencies in short baselines.

— Everything else being equal, it may require 10 seconds to resolve ambiguities when reflected signal is from a dry ground but 20 seconds when it is reflected from the wet ground. The reflected signal from a wet ground is almost as strong as the direct signal.

— Everything else being equal, it may require 10 seconds to resolve ambiguities with antenna that has good multipath rejection (-35 db down/up ratio) but 200 seconds with a regular antenna (-15 db down/up ratio). Multipath is a main cause of wrong integer estimation.

The number of satellites is the most important factor in resolving ambiguities reliably and quickly. As a rule of thumb, you need at least 8 satellites for short baselines. For short baselines, the role of dual frequency is that it multiplies the number of satellites by 1.5.

For long baselines, you must have at least 8 dual frequency satellites.

With GPS alone only 32% of the time, we have 8 satellites and only 5% of the time 9 satellites and very rarely more than 9.

GPS has 6 or more satellites 96% of the time. So Dual frequency GPS is marginally sufficient for short baseline RTK. This is because 6 dual frequency is equivalent to 9 single frequency for short baselines.

GPS+GLONASS always have 9 or more satellites, 99% of the time more than 10, 96% of the time more than 11 satellites.

Single frequency GPS can rarely work for short baselines RTK and never for long baselines.

Single frequency GPS+GLONASS can always work for short baselines but not for long baselines.

Dual Frequency GPS can always work for short baseline but rarely for long baselines.

Dual frequency GPS+GLONASS is the only RTK system that can always work for short and long baselines.

Adverse Signal Conditions

While, as we discussed above, the number of satellites is key in resolving the ambiguities reliably, there are still other adverse signal conditions like "signal fade-away", "in-band" and "out-of-band interference", and "brownout" that we need to deal with. The receivers used in RTK applications, in particular, need to have higher tracking capabilities and be able to handle these conditions since after any loss of lock we have to wait to resolve the integers again.

In the next chapter, we will describe advancements in GPS technology that improve receiver performance under these conditions, along with the mitigation of multipath errors that we deferred earlier.

Summary

In this chapter, we learned:

- *The accurate knowledge of the base position directly impacts the accuracy of the position computed by the rover.*
- *Differential mode will remove most of all errors except multipath and receiver errors. If multipath errors could be removed, millimeter level accuracy could be obtained with the carrier phase and decimeter level accuracy with the code phase.*
- *DGPS is based on measuring distances to satellites with code phase. We can measure distances instantly but not accurately.*
- *Resolving the integers correctly is the key in RTK. The big question is how long it will take to resolve the integers reliably after satellites are locked.*
- *Dual frequency GPS+GLONASS is the only RTK system that can always work for short and long baselines.*

Chapter 5. Advanced Features

Recent innovations in GPS receiver technology have introduced ways to enhance precision, availability, and reliability in advanced receivers. These innovations address mitigation of multipath errors, receiver (antenna center) stability, and data reliability under adverse signal conditions.

Adverse conditions arise from partial obstructions like canopies and light trees that make satellite signals weak, interference from other communication systems (radio and TV broadcasts, radar, communication satellites, etc.), interference from power lines and transformers, high levels of ionospheric activities that make tracking of L2 signal harder, abrupt motions of the receiver, or momentary signal interruptions like going through a short tunnel that temporarily blocks all satellites. Some of the methods used to deal with these problems are unique to specific receivers, the details of which are beyond the scope of our discussion but may be found in the manufacturers' literature.

Multipath

There are two techniques available to mitigate the effects of multipath: a) Signal processing technique and b) multipath rejection choke rings.

Signal processing technique — In this method, the data is analyzed to separate the direct signal from the indirect signal(s). You can imagine the echo of your voice in a canyon. If the indirect path is substantially longer than the direct path, then you may be able to distinguish the two and concentrate on the direct path. But if the difference is small, the echoed signal may be so close to the direct signal that you cannot separate them. With GPS signals, the signal processing technique is ineffective if the difference between the direct path and the indirect path is less than a few meters. Removing a multipath signal comes at the expense of removing part of the direct signal too, which in turn increases noise. Oftentimes, the more we try to remove the short distance multipath, the more noise we add.

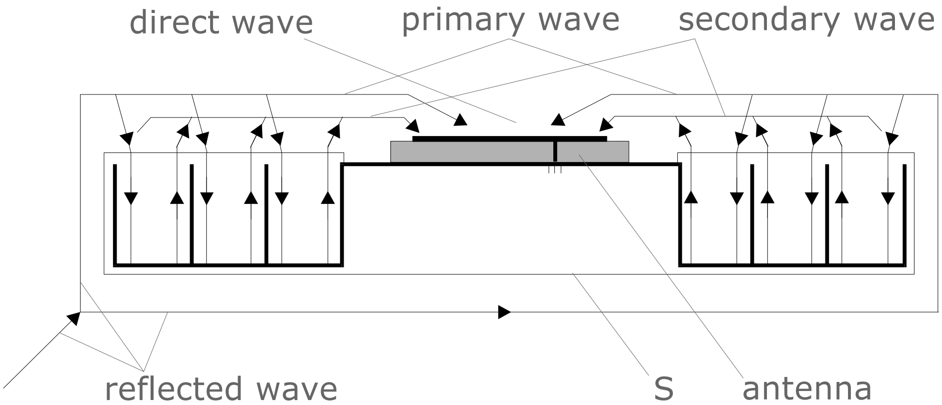
Multipath rejection choke rings — This technique works only for the multipath signals reflected from objects below the antenna. The reflected signal that hits the bottom side of the antenna can be rejected. This technique will do nothing for the reflected signals that hit the antenna on top, for example a signal that is reflected off a building above the antenna.

Fortunately most of the times the signals that are reflected from objects above the antenna have a multipath distance of more than 10 meters and signal processing techniques can mitigate them. For signals that are reflected off the ground, the multipath distance is in the order of a few meters and signal processing techniques cannot do much to address them, but choke rings can. Because of the complimentary nature of the two techniques, we can mitigate both ***"near***" and ***"far" multipaths***.

Choke Ring

A choke ring ground plane consists of several concentric thin walls, or rings, around the center where the antenna element is located. The area between the rings creates ***"grooves"***. The principle of the operation of choke ring ground planes is as follows.

The signal received by the antenna is composed of two components: ***Direct*** and ***reflected.*** The grooves have no effect on the direct signal other than decreasing the antenna gain at low elevation angles; for high elevation angles the choke ring ground plane works almost like a flat ground plane. But the grooves have much effect on reflected signal from underneath.



The electromagnetic field of the reflected signal in the vicinity of the choke ring ground plane can be viewed as the sum of a primary and a secondary field waves. The objective of the choke ring ground plane is for the primary and secondary reflected signals to substantially cancel each other and the direct signal to the antenna to remain as the dominant signal. If the amplitude of the primary and the secondary waves are equal and the phase difference between them is 180 degrees, then the two components of the reflected signal cancel each other at the antenna output and multipath is suppressed. So, a given choke ring has optimum effect only at the particular frequency that has resonance behavior.

For a given choke ring ground plane, the complete suppression of multipath only occurs at certain elevation angles; at other angles, the multipath is suppressed partially. The maximum suppression usually occurs at angles close to zenith, and minimal suppression at angles close to the horizon.

Choke rings are typically designed for one frequency. If a choke ring is designed for L1 then it has no effect on L2, while if it is designed for L2 then it has some benefits for L1. Recently, dual-frequency choke rings have been introduced that allow separate optimization for L1 and L2.

Antenna Phase Center Stability

The "electrical center" (also called the ***phase center***) of a GPS antenna is the point whose location we compute. To determine the coordinates of a point, we must co-locate the phase center of the antenna with that point. But usually the "physical center" of the antenna is positioned exactly above the point and the vertical offset is accurately measured. This vertical offset will then be taken into account when calculating the coordinates of the intended point. It is therefore assumed that the "physical center" of the antenna is the same as its phase center. This is true only in special advanced antennae designed for precision applications. The phase center of a typical antenna can change by many centimeters as the satellites' positions change. ***"Phase center stability"*** is the main characteristic of an antenna for precision applications.

In-Band and Out-of-Band Interference/Jamming Suppression

The operation of a GPS receiver can be severely limited or completely disrupted in the presence of in-band or out-of band interference and jamming signals. Most receivers filter out-of-band noise, but few suppress interference within the band.

The threat of the in-band interference and jamming signals increases daily as new communication systems are put in place and the radio frequency spectrum becomes more populated. The threat is not only from the interfering signals themselves but also from their harmonics that fall inside the GPS band.

Weak Signal/High Dynamics Tracking

Tracking satellites in an open field with no interfering signal and no partial obstruction is rather straightforward. The challenge comes when there are interfering signals that partially mask satellite signals, when there is a partial blockage that reduces the amount of signal energy that reaches the receiver, or when there are high dynamics, that increases noise. The merits that distinguish high performance receivers are in their ability to track satellites under all environmental conditions and dynamics, and in the type and quality of the measured data.

In tracking a satellite's carrier signals we have to track three main dynamics: 1) the dynamic due to the motion of the satellite, 2) the dynamic related to the motion of the receiver, and 3) the dynamic related to the oscillator (clock) of the receiver.

The motion of the satellite is the smoothest amongst the three. It can be predicted with the accuracy of a few meters per day. Even if we don't track the satellite for an hour and only rely on prediction, we will not be off by more than one meter. The motion of the receiver is unpredictable. The behavior of the clock of the receiver can also be unpredictable and abrupt, especially when it is subjected to shock, vibration, or temperature fluctuations. The change in the behavior of the receiver clock happens even when the receiver is not moving and, for example, it is sitting on a survey point.

All of the above three dynamics add together and result in a relative dynamic between the receiver and the satellite. And it is this total dynamic that the receiver must track.

There is a direct relationship between the ability to track a satellite and its signal strength. Consider the following analogy: You can easily observe a bright star in a clear night. If the brightness of the star decreases and starts to move randomly (forgive the moving star) you may have difficulty tracking it. Imagine that you are observing a weak star with a binocular. You may have no difficulty tracking its smooth motion. Now imagine you are following the same star with a binocular but this time you are sitting in a four wheel drive driven by a teenager in a bumpy road. Now you must track not only the motion of the star but also the motion of the car. If some imaginary instrument could track the motion of the car and compensate for it, you again may find it easy to follow the smooth motion of the star, perhaps even if the star became dimmer or if you drove under a tree which partially blocked your view. Even if you momentarily lost the star, you may not have much difficulty to quickly find it again when the obstruction disappears, since you can predict the smooth motion of the star.

A scheme developed by JPS (patent pending) performs the task of our imaginary instrument in the above analogy. In this scheme, the combination of all satellites is used to track the carrier dynamics of the receiver and its oscillator. Then, after the carrier of each satellite is compensated for the dynamics of the receiver and its oscillator, we have put ourselves back on a stable ground as in the above analogy. We can now track satellites under heavy foliage, or under high dynamic and interfering signals. The more satellites we track the more signals we have to add together and track the dynamics of the receiver and its clock. There are additional benefits derived from this scheme. For example, the accuracy of the output frequency signal is improved because it is tracked with the signals of all satellites, and the initial acquisition of low signal satellites is faster because the dynamics of the receiver and the oscillator are isolated from the task of acquiring and tracking satellites.

Summary

In this chapter, we learned:

- *Because of the complimentary nature of the choke ring and the signal processing methods, we can mitigate both "near" and "far" multipaths.*
- *Choke rings are typically designed for one frequency. If a choke ring is designed for L1 then it has no effect on L2, while if it is designed for L2 then it has some benefits for L1. New dual-frequency choke rings allow separate optimization for L1 and L2.*
- *"Phase center stability" is the main characteristic of an antenna for precision applications.*
- *The threat of the in-band interference and jamming signals increases daily as new communication systems are put in place and the radio frequency spectrum becomes more populated.*
- *The challenge in tracking satellites comes when there are interfering signals that partially mask satellite signals, when there is a partial blockage that reduces the amount of signal energy that reaches the receiver, or when there are high dynamics that increase noise. The merits that distinguish high performance receivers are in their ability to track satellites under all environmental conditions and dynamics, and in the type and quality of the measured data.*

Chapter 6. Radio Modems

As we discussed earlier, in real-time DGPS and RTK we transmit data from the base GPS receiver situated in a known position to the rover GPS receiver. The rover GPS receiver then takes the base data into account in order to compute its own position accurately.

Radio modems provide wireless communication between the GPS base receiver and the GPS rover receiver. We need a radio modem transmitter with the base GPS receiver and a radio modem receiver with the rover GPS receiver. When a base GPS receiver broadcasts data via its radio modem transmitter, an unlimited number of rover receivers can pick up the data via their radio modem receivers.

A radio modem transmitter consists of a radio ***modulator***, an ***amplifier***, and an ***antenna***. The radio modulator takes the GPS data from the GPS receiver and converts it to a radio signal that can be transmitted. The amplifier raises the power of the signal to a level that can reach the rover GPS. The farther the rover, the more signal power we need. The transmitter antenna then transmits the amplified signal. The power of the amplifier directly affects the distance that the signal can travel (the range) and the reliability of the communication. The range also depends on the terrain and the radio antenna setup.

The connection between the radio modem transmitter and the base GPS is via the serial ports of the GPS receiver and the radio modem transmitter. If the radio transmitter is integrated within the GPS electronics, then the connection is done internally.

A radio modem receiver consists of a radio receiver antenna and a radio ***demodulator***. The radio receiver antenna picks up the radio signal from the air and delivers it to the radio demodulator that converts the signal back to the form that can be delivered to the serial port of the rover GPS receiver.

The connection between the radio modem receiver and the rover GPS receiver is via the serial ports of the GPS receiver and the radio modem receiver. If the radio receiver is integrated with the GPS electronics then the connection is done internally.

There are varieties of radio modems in the market. The main characteristic of a radio modem is the form to which the data is converted for its transmission. ***UHF***, ***VHF***, and ***Spread Spectrum*** (frequency hopping or direct) are some examples. There are some advantages and some disadvantages to each form.

Recent introductions in GPS radio modems include the use of accurate timing of the GPS for data synchronization between base and rover(s) in order to enhance data integrity, and the use of direct and frequency hopping combination in Spread Spectrum radio to enhance communication reliability.

Government authorization may be required for using certain types of radio modems. There are international and national bodies that allocate frequency bands and issue authorization to transmit signals. In some countries, there are bands that are allocated for public use without the need for any special authorization. This is an important factor to consider when selecting a radio modem, since getting authorization to broadcast information is often not an easy task. The bands that are allocated for public use are of particular interest. The 900 MHz band in the United States and 2.4 GHz in most European countries are allowed for spread spectrum communication without any special authorization (but there are limitations on the amount of power that one can use to transmit signals).

UHF and spread spectrum radio modems are the most popular for DGPS and RTK applications. Spread spectrum radios (900 MHz and 2.4 GHz) have a range of about 20 kilometers (unless the antenna is installed in a very high location). UHF has a longer range. With a 35-Watt amplifier, a UHF radio can have a range of up to 45 kilometers, depending on terrain and antenna setup.

Summary

In this chapter, we learned:

- *There are varieties of radio modems in the market. UHF, VHF, and Spread Spectrum (frequency hopping or direct) are some examples.*
- *Recent introductions in GPS radio modems include the use of accurate timing of the GPS for data synchronization between base and rover(s) in order to enhance data integrity, and the use of direct and frequency hopping combination in Spread Spectrum radio to enhance communication reliability.*
- *It is important to consider public bands when selecting a radio modem, since getting authorization to broadcast information is often not an easy task. The 900 MHz band in the United States and 2.4 GHz in most European countries are allowed for spread spectrum communication without any special authorization.*

JAVAD®

POSITIONING SYSTEMS