Fundamentals of
Navigation and Inertial Sensors

AMITAVA BOSE
Guest Faculty
Aerospace Engineering and Applied Mechanics Department
Bengal Engineering and Science University, Shibpur
and
Former Director, ISRO Inertial Systems Unit, Trivandrum

K.N. BHAT
Emeritus Professor
Centre for Nano Science and Engineering (CeNSE)
Indian Institute of Science, Bangalore

THOMAS KURIAN
Dean and Head of the Department of Avionics
Indian Institute of Space Science and Technology, Trivandrum

PHI Learning Private Limited
Delhi-110092
2014

₹425.00

FUNDAMENTALS OF NAVIGATION AND INERTIAL SENSORS
Amitava Bose, K.N. Bhat, and Thomas Kurian

ISBN-978-81-203-4859-2

Published by Asoke K. Ghosh, PHI Learning Private Limited, Rimjhim House, 111, Patparganj Industrial Estate, Delhi-110092 and Printed by Star Print-O-Bind, F-31, Okhla Industrial Area Phase I, New Delhi-110020.

Contents

Preface

Navigation, which fundamentally provides information on position and direction, is needed in ocean, land, air and in space. This information has been extremely useful for the growth of civilisation through the ages. It is quite expected that myriad forms of navigation developed during this long period leading to current versions which are collectively called modern navigation or simply 'navigation'. Navigation for aerospace is a multi-disciplinary area primarily covering inertial navigation, satellite navigation, radio navigation, stellar navigation and integrated inertial navigation.

The book, *Fundamentals of Navigation and Inertial Sensors*, covers the topics of navigation primarily aiming for their application in aerospace and has focussed on topics related to inertial navigation, inertial sensors, MEMS based inertial sensors, satellite navigation, integrated inertial navigation, signal processing of inertial sensors and lastly their applications.

The book has aimed to meet the academic needs of undergraduate course on the subject of navigation and inertial sensors. The subject is taught in the branch of aerospace engineering as well as in the specialised branch of Avionics. An earlier version of the book, where the current authors contributed, is being taught at IIST, Trivandrum. However, that book is not available for students other than IIST. The current book is considerably revised and updated considering broad range of technical institutes in the country. The content of the book is likely to be more than what is taught in the undergraduate syllabus. Some of the chapters have extended coverage and could be taught at postgraduate level where research and design are emphasised. In particular, Chapter 7 on integrated navigation involving Kalman filter based design, is likely to have more relevance at the postgraduate level.

The subject of the book is specialised, involves multi-disciplinary branches of science and engineering, and these basic topics are normally taught at the initial semesters in undergraduate study. So, the course will suit the students if taught at the final year. Its utility as a textbook has been enhanced by incorporating 'worked examples' and problem assignments at the end of the chapter wherever found to be of relevance.

Chapter 1 deals with *Introduction to navigation*. It brings out the evolution of navigation in the last century and outlines the basic principles of modern navigation and inertial sensors.

Chapter 2 dwells on the working principle, analytical formulations and features of *autonomous strapdown inertial navigation* that has evolved during the last four decades and expanding in scope and application to include launch vehicles, missiles, aircrafts and host of land as well as sea going vehicles.

Chapters 3 and 4 discuss the principle of operation, features and some aspects of design and technology of macro-sized inertial sensors consisting of *gyro* and *accelerometers* respectively. These sensors constitute the backbone of modern autonomous inertial navigation.

Chapter 5 brings out salient features of the emerging technology of *micro-electromechanical inertial* sensors, popularly known as MEMS, the technology of which is becoming increasingly useful to emerging inertial sensors application where small size and mass along with improved reliability are considered essential.

Chapter 6 deals with satellite navigation, its operating principle, features, determination of navigation parameters, its error and error reduction schemes. While global positioning system is the primary system that is addressed, description also includes brief features of other satellite navigation schemes which have emerged.

Chapter 7, introduces the topic of integrated inertial navigation, through elaboration of the process of multi-sensor navigation and use of Kalman filter as an estimator. The principle of operation of various integrated schemes involving satellite navigation system and inertial navigation are described.

Chapter 8 deals with the basics of signal processing with particular relevance to inertial sensors whose output will be either analog or digital and associated with noise.

Chapter 9, which is the last chapter of the book, provides some interesting *applications of navigation and inertial sensors* to enable readers' to appreciate the diversity of current usage of this fascinating and growing technology.

Four appendices (A to D) have been added dealing on laser fundamentals, fibre-optics features, Q-factor and inertial sensor noise, which will aid in the understanding of the topic on gyros dealt in Chapter 3.

The subject of navigation, inertial sensors, MEMS-based inertial sensors and signal processing has made considerable use of increasing number of specific terminologies. Such terminologies are explained in the book as and when they appear and in addition these are listed in an alphabetical manner in Glossary.

The book has used certain symbols that carry common definition unless defined differently at some other place of the book. These common symbols have been presented at the beginning of the chapters.

The book has made considerable use of certain *acronyms* that appear at different places in the book and find considerable usage within inertial engineering. These acronyms have been listed alphabetically at the beginning of first chapter.

To the extent possible and needed, each chapter has been written to provide information on a common frame. For example, the equations, figures, worked examples and the tables, have been numbered fresh and sequentially from the beginning of each chapter. Problems are listed at the end of the chapters.

It is hoped that the book will be useful for the students and the faculty and stimulate further growth in this exciting field.

Amitava Bose
K.N. Bhat
Thomas Kurian

Acknowledgements

Authors gratefully acknowledge the following persons for the support and encouragement they have provided in realising this book:

We sincerely thank Dr. P.P. Mohanlal, ISRO Inertial Systems Unit for meticulously reviewing Chapter 7 and offering valuable comments.

We are grateful to Dr. David D. Lynch, Chief Scientist (retired), Litton Systems, Goleta, California, USA, for reviewing and offering valuable suggestions on the write up on Hemispherical Resonator Gyro and for allowing the use of materials from his published papers.

We thank Indian Institute of Space Science and Technology, Trivandrum for allowing the use of some of the materials from the book *Introduction to Navigation and Inertial Sensors.*

Amitava Bose
K.N. Bhat
Thomas Kurian

Acronyms

ADC	:	Analog-to-Digital Converter
ADOP	:	Attitude Dilution of Precision
AHRS	:	Attitude and Heading Reference System
ASIC	:	Application Specific Integrated Circuit
BITE	:	Built-In Test Equipment
C/A	:	Coarse Acquisition
CCW	:	CounterClockWise
CEP	:	Circular Error Probability
CMOS	:	Complementary Metal Oxide Semiconductor
CNT	:	Carbon Nano Tube
CVD	:	Chemical Vapour Deposition
CVG	:	Coriolis Vibratory Gyro
CW	:	ClockWise
DAC	:	Digital-to-Analog Converter
DCD	:	Direct Current Discharge
DCM	:	Direction Cosine Matrix
DETF	:	Double Ended Tuning Fork
DGPS	:	Differential Global Positioning System
DOF	:	Degrees of Freedom
DRIE	:	Deep Reactive Ion Etching
DTG	:	Dynamically Tuned Gyro
DSP	:	Digital Signal Processing
ECEF	:	Earth Centred Earth Frame
ECI	:	Earth Centred Inertial
ECM	:	Effective Centre of Mass
EDP	:	Ethylene Diamine Pyrocatacol solution

EMI	:	ElectroMagnetic Interference
ENOB	:	Effective Number Of Bits
ENU	:	East North Up
FDI	:	Failure Detection and Isolation
FIR	:	Finite Impulse Response
FMECA	:	Failure Mode Effect and Criticality Analysis
FOG	:	Fibre-Optic Gyro
GaAlAs	:	Gallium Aluminium Arsenide
GAGAN	:	Gps Aided Geo Augmented Navigation
GDOP	:	Geometric Dilution of Precision
GLONASS	:	GLobal Orbiting NAvigation Satellite System
GNSS	:	Global Navigation Satellite System
GPa	:	Giga Pascal
GPS	:	Global Positioning System
GSO	:	GeoSynchronous Orbit
GTO	:	Geostationary Transfer orbit
HDOP	:	Horizontal Dilution Of Precision
HFD	:	High Frequency Discharge
HMC	:	Hybrid Micro Circuit
HRG	:	Hemispherical Resonator Gyro
IARS	:	Inertial Attitude Reference System
IBE	:	Ion Beam Etching
IC	:	Integrated Circuit
ICP	:	Inductively Coupled Plasma
IFOG	:	Interferometric Fibre-Optic Gyro
IMU	:	Inertial Measurement Unit
INS	:	Inertial Navigation System
IRNSS	:	Indian Regional National Satellite System
KOH	:	Potassium hydroxide solution
LADGPS	:	Local Area DGPS
LCP	:	Left Circularly Polarised
LIGA	:	LIthographie Galvanoformung Abformung
LPF	:	Low Pass Filter
LPI	:	Launch Point Inertial
LSB	:	Least Significant Bit
LSE	:	Least Square Estimate
MCS	:	Master Control Station
MEMS	:	Micro ElectroMechanical Sensor or System
MMSE	:	Minimum Mean Squared Error
MSB	:	Most Significant Bit
NED	:	North East Down

NEMS	:	Nano ElectroMechanical Systems
NMPH	:	Nautical Mile Per Hour
PDF	:	Probability Density Function
PDOP	:	Position Dilution Of Precision
PLL	:	Phase Locked Loop
PPM	:	Parts Per Million
PPS	:	Precise Positioning Service
PSD	:	Power Spectral Density
PWM	:	Pulse Width Modulation
PZT	:	Lead Zirconate Titanate
QUEST	:	QUaternion ESTimation
RAIM	:	Receiver Autonomous Integrity Monitoring
RCP	:	Right Circularly Polarised
RF	:	Radio Frequency
RFOG	:	Resonant Fibre-Optic Gyro
RG	:	Rate Gyro
RIE	:	Reactive Ion Etching
RIG	:	Rate Integrating Gyro
RLG	:	Ring Laser Gyro
RPM	:	Revolutions Per Minute
SA	:	Selective Availability
SAR	:	Synthetic Aperture Radar
SFB	:	Silicon Fusion Bonding
SFS	:	Super-fluorescent Fibre Source
SiC	:	Silicon Carbide
SINS	:	Strapdown Inertial Navigation System
SLD	:	Super Luminescent Diode
SMPM	:	Single Mode Polarisation Maintaining
SMT	:	Surface Mount Technology
SNR	:	Signal-to-Noise Ratio
SNS	:	Satellite Navigation System
SOI	:	Silicon On Insulator
SOL	:	Safety Of Life
SPS	:	Standard Positioning Service
SSPO	:	Sun Synchronous Polar orbit
SV	:	Space Vehicles
TCR	:	Temperature Coefficient of Resistivity
TDOP	:	Time Dilution Of Precision
TMAH	:	Tetra Methyl Ammonium Hydroxide solution
TNEA	:	Total Noise Equivalent Acceleration
TRIAD	:	TRI-axis Attitude Determination

TRP	:	Total Reflecting Prism
UERE	:	User Equivalent Range Error
UTC	:	Universal Time Co-ordinated
VBA	:	Vibrating Beam Accelerometer
VCO	:	Voltage Controlled Oscillator
VDOP	:	Vertical Dilution Of Precision
VFC	:	Voltage-to-Frequency Converter
WAAS	:	Wide Area Augmentation System
WGN	:	White Gaussian Noise
WGS	:	World Geodetic System

Notations

λ	:	Wavelength of light
ω_e	:	Earth's inertial angular rate about polar axis
c	:	Velocity of light
c_0	:	Velocity of light in vacuum
Ω	:	Inertial input angular rate
ω	:	Frequency
h	:	Hour
f	:	Specific force
g	:	Earth's newtonian gravitational acceleration
µ	:	Micro (10^{-6})
n	:	Nano (10^{-9})
m	:	Unit of length in metre
km	:	Unit of length in kilometre
cm	:	Unit of length in centimetre
g_p	:	Earth's plumb bob gravity
θ, ψ, ϕ	:	Three-axis rotation defined in Euler angles
q_0, q_1, q_2, q_3	:	Three-axis rotation defined in quaternion
τ	:	Time constant
ξ	:	Damping ratio
Bold alphabet	:	Symbol for a three-dimensional vector in equation
M	:	Proof mass
K	:	Stiffness
Q	:	Quality factor
J_2	:	Earth's second gravitational constant
V	:	Velocity
F	:	Force
H	:	Angular momentum
°C	:	Unit of temperature in degree celsius
mA	:	Unit of current in milliampere
arc-s	:	Arc second

1

Introduction to Navigation

1.1 Navigation Definition

Navigation is the art and science of manoeuvring safely and efficiently from one point to another. The word navigation (In Latin 'navis' means boat 'agire' means guide) traditionally meant the art or science of conducting ships or watercrafts from one point to another. Navigation has many facets, many definitions and many subsets. Beginning by simply wanting to know 'where am I' and then expands upon the simple statement with information desired on how to get from 'where I am' to 'how do I get to my destination'? The question then arises 'reference to what' and the problem becomes more complex, involving co-ordinate systems, space and so on. In nutshell, we can define navigation as:

Navigation is the determination of a physical body's position and velocity and directing the course of the body, relative to some reference co-ordinate frame such that the destination is achieved.

The above definition implies three kinds of operations, namely

(a) Determination of the present position and velocity of the vehicle (or body) relative to a known reference system

(b) Modifying the course of the vehicle to achieve destination

(c) Stabilising the vehicle and staying on track

In the modern world, first operation is referred as 'Navigation', the operation of achieving destination is termed as 'Guidance' and the aspects of stabilisation and staying on track are termed as 'Control'.

1.2 Types of Navigation and Their Evolution

Celestial navigation is one of the earliest forms of navigation where the stars define a fixed reference frame in space, commonly known as inertial reference frame. For many centuries, sailors planned their courses, sailed and recorded their progress all with reference to a single star called North star or Pole star or Polaris, which is 300 light years away from earth. Polaris is in the same direction as the North Pole of the earth. So, by moving in the direction of the Pole star, leads one towards north. Its height above the horizon provides for the latitude of the place. Star sighting along with a plumb bob that provides direction towards the earth centre were used for ship's navigation. In the daytime, when the stars are not visible, the sun was used to define the direction by observing its rise in the east, setting in the west and defining the south at around mid-day.

In the modern days, these celestial observations have been instrumented to provide what is known as *Stellar navigation* where star sensors and sun sensors provide very accurate direction, normally called attitude, to multifarious users including satellites and space missions. The star sensor provides direction when only one star is sighted and by sighting three stars without ambiguity, three-axis inertial attitude determination is possible. The distance of stars from earth or the solar system is too large, as a result, the star sensor observations can be used for finding direction only.

Navigation using *X-ray pulsers* [**Graven et al., 2008**] provide an alternate means for interplanetary missions and for spacecraft tracking. The rapidly rotating neutron stars, known as pulsars, were discovered in the 1960s. These are called celestial lighthouses and are becoming useful for very deep space navigation. Their signals of periods from 1 second to 1000 seconds are of atomic clock quality. A comparison of the arrival time of their pulses at spacecraft and at the earth via an earth orbiting satellite gives 3D position of the spacecraft. A small on board X-ray detector using one day of data gives position accuracy of 150 km independent of range in contrast with techniques measuring angles leading to degradation in accuracy with distance. The Kalman filter fusing the spacecraft dynamics with pulsars signals, and further using inertial sensors, star trackers, sun and horizon sensors all help autonomous navigation in deep space. For a spacecraft using deep space network tracking, the range and range rate measurements from earth enhances accuracy and a synergy with pulsar signals is even better and the latter more attractive for very deep space applications.

In the 17th century, Newton provided laws of inertia and gravitation that eventually led to the fundamental principle of *Inertial navigation*. But it took nearly three centuries for the inertial navigation technologies to develop a viable system encountering enroute tremendous set backs even in the 1930s when experimental aircraft sorties were undertaken by Joaness Boykow's to prove the concept of inertial navigation. Sustained efforts thereafter proved the feasibility of inertial guidance during the Second World War, when V-2 rockets were guided with inertial system to hit the targets.

Contrary to other navigational aids, INS does not rely on external measurements. Instead, it utilises the inertial properties of sensors to provide self-contained, non-radiating, non-jammable and accurate determination of instantaneous navigation states. Two types of measurements are involved, namely, inertial rotation as measured by the gyros and the specific force as measured by the accelerometers. The term specific force is used to define acceleration due to inertial forces that are measured by accelerometers as per Newton's law. Since, Newtonian gravitational acceleration is not measured by the vehicle borne accelerometers; it is modelled and stored in the navigation computer. By combining the inertial measurements along with a gravity model, autonomous determination of position, velocity and direction, also called attitude, are carried out and as a result such a system is called an *Autonomous inertial navigation system*. Accurate modelling of the gravitational acceleration is an integral part of inertial navigation.

Radio navigation is an invention in the early part of the 20th century when electronic instruments supplemented many centuries old navigation technologies. In 1904, German inventor Christian Hu'lsmeyer introduced the first navigation devices that used radar technology. Radio direction finders were the first form of electronic navigation to come into use. In 1926, successful two-way radio air to ground communication began. The first radio equipped airport control tower was built in Cleveland, Ohio in 1930. Until World War II, pilots used the simplest form of electronic navigation called NDB (Non Directional Beacon), which relied heavily on low frequency transmitters. After the war, very high frequency transmitters, called the Very High Frequency Omni Directional Range (VOR) and Instrument Landing System (ILS) stations were developed and put into place. VOR provided azimuth information, and Distance Measuring Equipment (DME) was required for positioning.

Satellite navigation is the latest addition to the modern navigation where artificial earth satellites are used to provide continuous information on user location and velocity.

The era of artificial satellites started with the launching of Russian satellite *Sputnik* in 1957, which was a significant technological breakthrough. Two researchers, Guier and Wieffenbach at the Johns Hopkins Applied Physics Laboratory (APL) determined the entire Sputnik 1 orbit from Doppler shift data during a single pass of the satellite. McClure immediately recognised that this technique could be inverted. If the orbit of the satellites were already known, a radio receiver's unknown position could be determined accurately from the same Doppler measurements. Thus the world discovered that Doppler shift in the signal could be used in determining precise position anywhere. This concept was realised in 1960s with the US Navy's Navigation Satellite System, called TRANSIT satellite program.

TRANSIT was composed of six satellites at 1075 km altitudes in nearly six circular polar orbits, circling the earth every 107 minutes. This constellation of orbits formed a birdcage within which the earth rotates. Whenever a satellite passed above the horizon, the user had the opportunity to obtain a single horizontal position. TRANSIT satellites broadcast two continuous signals at 150 MHz and 400 MHz respectively. Dual frequencies were used to remove ionospheric delay. In the operational system, each TRANSIT satellite provided four to six position updates per day for every user. But there were two major drawbacks: Inherent system error and the error introduced by user's unknown motion during the satellite pass. TRANSIT system was not three-dimensional, provided only periodic updates, and had degraded accuracy when used by a moving user. However, this program proved several important points for satellite navigation.

(a) Satellites could be very reliable with satellite life as long as 5 years or more.

(b) Satellite positions could be predicted very accurately for navigation use.

(c) Ionospheric effects could be compensated for dual frequency signal design.

(d) Highly stable clocks could be utilised in orbit.

By 1972, another navy satellite system was extending the state of art by orbiting very precise clocks, known as *Timation*. It was based on providing accurate time and three-dimensional position to users based on the ranging, not on Doppler measurements. The ranging requires stable clocks on the satellite and must be synchronised to a master ground clock. Ranging was provided by a method known as *Side Tone Ranging* (STR) that allowed users to resolve phase ambiguities to the satellite. This program developed the first atomic clocks for space that proved critical for progress in later generation satellite navigation. The final precursor to current generation satellite navigation was the satellite program 621B, which introduced the concept of *Pseudo Random Noise* (PRN) codes **[Parkinson and Bradford, 1997]**. These codes modulate the signal with a repeatable sequence of ones and zeros. An appropriate family of codes can be used to broadcast multiple channels of information at the same carrier frequency since they can be chosen to be nearly orthogonal. Such technology developments had a profound effect and resulted in the emergence of Satellite Navigation System (SNS), providing continuous global positioning capability to users all over the world by orbiting sufficient number of satellites and also ensuring that at least four were electronically visible round the year and 24 hours a day. Global positioning system of USA and global orbiting navigation satellite system of former USSR were the outcome of this vision.

Integrated navigation is another form of modern navigation where data from complementary navigation sensors are used to improve the navigation accuracy and when one of the data source is an inertial system, it is often globally referred as integrated inertial navigation system or specifically an aided inertial system. The aim is to eliminate intrinsic time dependent error growth in INS for flights, which are of long duration. Integration of an INS with radio-based navigation aids such as LORAN and OMEGA had been used for decades. With the introduction of satellite navigation system worldwide, and the availability of cheaper and reliable receivers, the current trend is to integrate INS with SNS.

In an integrated inertial navigation system, the unbounded error growth with time in an autonomous inertial navigation system is corrected with the help of another external measurement. An external measurement can be of any quantity that duplicates a navigation parameter such as velocity, position or instrument orientation. Each external measurement is compared to the value computed by the INS and the difference is used in a filter, typically a Kalman filter, to reduce or bound the navigation error growth. Strategies for incorporating external measurements can be as simple as executing position or velocity resets to a more elaborate mixing of data. The integrated navigation system combines the navigation data provided by complementary sensors to obtain the final navigation solution. Navigation sensors are *Complementary* if they meet the following conditions:

(a) The set of sensors generate all the information required to compute a complete navigation solution.
(b) The sensors have complementary error dynamics.
(c) Error dynamics of the sensors are observable.

1.3 Features of Inertial Navigation

The basic features of inertial navigation encompass reference frames, inertial frame, non-inertial frame, gravitational acceleration and gravity, universal standard time and inertial systems instrumented with gyros and accelerometers. These are explained in the subsequent sections.

1.3.1 Reference Frame and Inertial Frame

In the different types of navigation working on different principles, a common aspect is the requirement and choosing a reference co-ordinate frame. The choice comes as several reference frames have evolved which suit a particular application. In three-dimensional measurements, all these frames are orthogonal and right handed. Figure 1.1 shows such an orthogonal frame XYZ with origin at P.

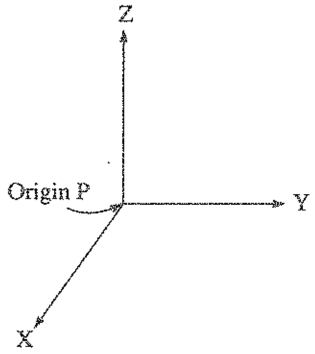


Figure 1.1 Orthogonal co-ordinate frame.

A *right handed orthogonal frame* in the sequence X, Y, Z can be defined by placing unit vectors $\mathbf{i}_1$, $\mathbf{i}_2$, $\mathbf{i}_3$ on these axes X, Y and Z respectively, then these unit vectors satisfy $\mathbf{i}_1 \times \mathbf{i}_2 = \mathbf{i}_3$; $\mathbf{i}_2 \times \mathbf{i}_3 = \mathbf{i}_1$; $\mathbf{i}_3 \times \mathbf{i}_1 = \mathbf{i}_2$. This can be further interpreted by saying if we choose the direction of first two orthogonal axes, then the third axis direction is governed by the right handed rule and is no more independent. The direction can also be represented by the convention of right hand thumb towards the X-axis, the index finger towards the Y-axis, then the middle finger points to the direction of the Z-axis.

The term 'Inertial frame' means a reference frame that is non-rotating and non-accelerating in inertial space. Newton's laws of motion are valid in this frame. It is a frame where navigation accelerometer measures the specific force without requiring any compensation for Coriolis as well as centrifugal forces. For a more practical appreciation, such a frame is considered stationary with respect to distanced stars. However, for majority of navigation application, the inertial frame should be earth based, and hence the emergence of Earth Centred Inertial (ECI) frame. Further explanations are provided to show how ECI frame is considered inertial. The frame orientation is as follows:

X: along the direction of the vernal equinox and lying on the equatorial plane
Z : along the earth's spin axis pointing celestial north
Y: completes the right handed system and lies on the equatorial plane

The origin of the frame is at the centre of earth. Since Newtonian gravitation is considered zero at the centre of earth, it is treated as inertial. The time of vernal equinox occurs around March 21–23 every year. The epoch, when the sun is crossing the equator from southern hemisphere to northern hemisphere, the direction of the line joining the centre of earth to the centre of sun at the time of vernal equinox is the direction of X-axis. The axis is treated as inertial.

Earth's spin axis, N–S axis, always points towards the pole star even though earth is travelling around the sun and thus the spin axis is treated as inertial. This aspect is shown in Figure 1.2.

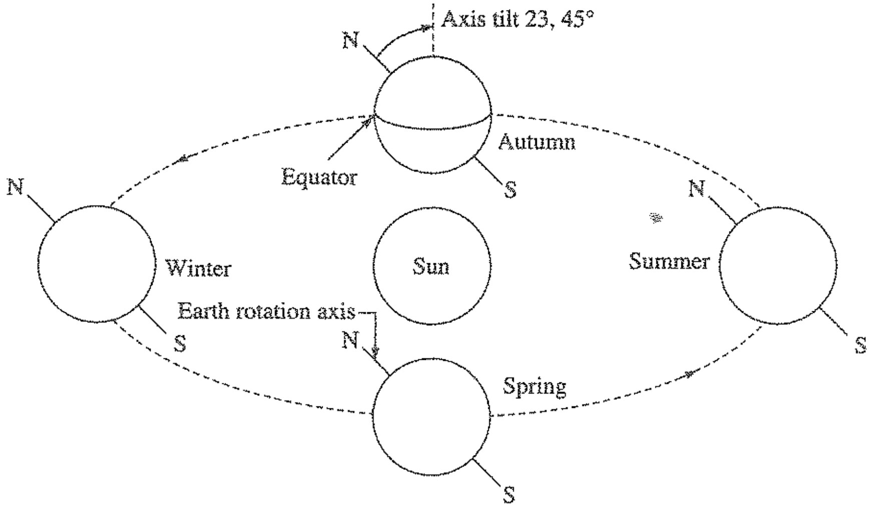


Figure 1.2 Inertial pointing of earth's spin axis while travelling around the sun.

Because of this inertial orientation of spin axis, motion of earth around the sun is treated as displacement and not a rotation. Thus it can be said that ECI-axes can be considered as non-rotating with respect to distanced stars and hence is inertial. Having defined ECI frame in this manner, it is appropriate to mention that such a frame is considered as quasi-inertial. This arises as the net gravitational force at the centre of earth is not absolutely zero. The second reason is that the orientation of the spin axis of earth is not truly inertial. Author **[Parvin, 1962]** has illustratively brought out various long-term motions of the spin axis. Effect of this deviation is generally neglected even for high accuracy navigation.

1.3.2 Sidereal Earth Rate

Spin rate of earth leads to day and night and is 15°/h taking sun as reference which results in the mean solar day of 24 hours. But taking distanced star as reference, the spin rate value is close to 15.04°/h. This rate is known as sidereal rate of earth that is considered as the inertial earth rate and this angular rate leads to sidereal day of 23 h 56 m 04 s. Figure 1.3 shows the occurrence of the solar day and the sidereal day showing that the latter is slightly shorter. The sidereal earth rate is a vector. Its magnitude changes with cosine function of latitude on the horizontal plane of earth and as a sine of latitude along the vertical axis. Knowledge of these values is important for calibration of gyros and in navigation computation.

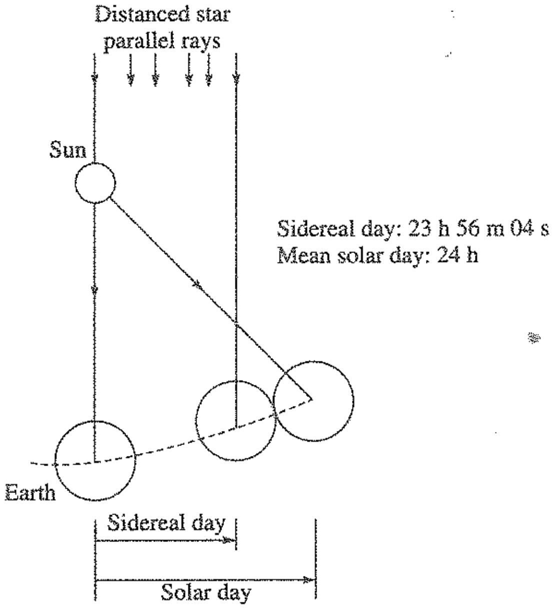


Figure 1.3 Occurrence of solar day and sidereal day.

1.3.3 Shape of Earth

For all types of navigation in the vicinity of earth, it becomes a necessity to define the shape of earth. For inertial navigation, the necessity additionally arises for accurately computing gravitational acceleration which in turn demands a model to define the shape of earth. Shape of earth can be defined as sphere where low accuracy navigation is involved; otherwise it is normally defined as ellipsoid which is shown in Figure 1.4.

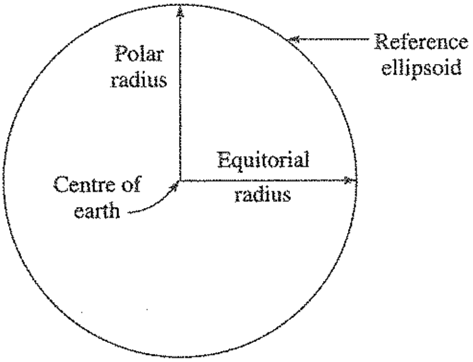


Figure 1.4 Ellipsoid shape of earth.

Due to rotation of earth about the polar axis, a bulging occurred at the equator resulting in a non-spherical shape where the equatorial radius is slightly greater than the polar radius. The difference between equatorial radius and the polar radius is around 22 km. However, the actual shape of earth is irregular with hills and valleys, and for navigation computation a model is necessary which closely approximates this shape. Ellipsoid shape of earth provides the mathematical model as well as the close approximation to the non-spherical shape. Various parameters define this shape and universally, the parameters and their values provided in World Geodetic Survey (WGS)-84 and their later versions, are accepted for use. Centre of earth is then the centre of this ellipsoid earth.

1.3.4 Co-ordinated Universal Time (UTC) and Unit of Time

As sun rises at different times over the globe, each country tries to follow its own standard time. For world wide navigation, it then becomes necessary to adopt a time scale that is universally accepted and this time is called Co-ordinated Universal Time (UTC). UTC has its historical origin in Greenwich Mean Time (GMT) where the prime meridian is passing through the Greenwich and the zero hour starts at midnight.

Earlier, the unit of time, the *second*, was defined as 1/86400 of a mean solar day as determined by the *rotation* of the earth around its axis. Since the rotation rate is not sufficiently stable, a correction is required to be made in UTC at certain intervals known as 'Leap second' where 1 s is either added or subtracted. Later on, with the availability of cesium atomic clock, a more accurate definition of *second* has emerged. As per this atomic clock, a second is defined by the time for 9,192, 631,770 ticks of the atomic clock. This happens to be the frequency of a cesium atom. Located onboard the navigation satellite, such ultra high accuracy clock, provides accurate time with 10 ns (nano-second) accuracy, and is made available to the SNS receiver user anywhere in the world. This provides a facility of synchronisation of clocks across the globe.

1.3.5 Earth Reference Navigation Frames

Earth Centred Inertial (ECI) frame is not suitable for varieties of earth based navigation application. This happens as ECI is an inertial reference frame. This means that a stationary object on earth with zero relative velocity with respect to earth will indicate displacement with time in ECI frame as earth is spinning. Similarly, as the location of a place on earth is mapped with latitude and longitude, an aircraft pilot needs to know the corresponding values in the aircraft navigation system while flying over a terrain. So, the pilot needs that the navigation frame used in the aircraft provides this mapped information. As a result quite a few earth based navigation frames have emerged and are in use. But such a frame is not inertial due to rotation of earth in inertial space. The terminology of inertial navigation is used for all such navigation systems where the primary sensors are gyros and accelerometers.

Reference Frame Relationship and Co-ordinate Transformation

Navigation information computed in one reference frame can be expressed in another reference frame using a mathematical process called co-ordinate transformation. This is possible as such mathematical transformations are well defined. This is extremely useful and a widely used process. This means that a system navigating in ECI frame can provide data in an earth based frame and vice versa, assuming that the onboard computer has adequate computing capacity. In the modern time, onboard computer speed is not a limitation which was the case three decades ago.

1.3.6 Inertial Navigation System Description

In inertial navigation, the first requirement is the choice of a reference frame for navigation. The choice of the frame can be inertial or non-inertial as discussed earlier. A brief description is given below for a system where the choice is made for a reference frame that is considered as inertial.

The instrumentation of the inertial reference frame is carried out using gyros. How the gyros provide the reference frame, is explained later in Chapter 3. The second requirement is to compute velocity and position information with respect to the inertial reference frame. Accelerometers, when aligned to this reference frame, measure the inertial forces acting on the vehicle but not the earth's gravitational acceleration. In order to compute position, earth's gravitational acceleration is modelled in a computer. Now, there are two schemes for implementing the reference frame with gyros. The first and the earlier scheme used gyro stabilised gimbals. Later on, the scheme uses analytic gimbals, in lieu of the mechanical gimbals, to provide the reference frame using gyros. The analytic scheme is popularly called strapdown system and is explained further.

In strapdown inertial navigation (refer Figure 1.5), the vehicle, whose velocity and position are to be measured, is rigidly connected to an inertial sensing unit consisting of three numbers each of gyros and accelerometers which are aligned to an orthogonal frame with axes X, Y, Z.

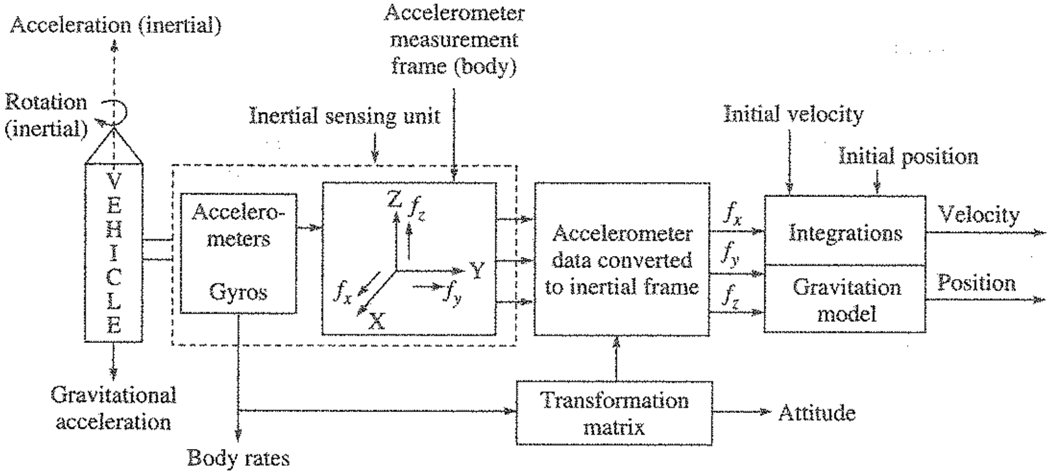


Figure 1.5 Principle of Strapdown (analytic) inertial navigation.

The vehicle is acted on by inertial forces (e.g. thrust, drag, aerodynamic), the gravitational force and will have three-axis rotational motions. The accelerometer measurements (f_x, f_y, f_z), called specific force, are made in body frame. Using the inertial rotation information measured by the gyros, transformation matrix is computed that is then further used to convert the accelerometer data to inertial frame. This is equivalent to saying that with the transformation, three accelerometers are providing specific force (f_x, f_y, f_z), in the reference inertial frame. Earth's gravitational acceleration is computed using appropriate gravitation model which resides in the computer. The resultant vehicle acceleration in inertial frame is then obtained by the algebraic summation of these two accelerations. Using successive integrations, the velocity and position are further computed. As two integrations are involved, initial values of these parameters are inserted at the start of navigation. This is the basic process of computing in inertial reference frame, while there will be several steps to be followed in a sequential manner to execute the computations. Attitudes are obtained from the transformation matrix, while the body rates are directly available from the gyro measurement.

So, strapdown inertial navigation system functions can be summarised as follows:

1. Instrument the reference frame for navigation. This is achieved by measuring the inertial rate in body frame with the body frame mounted gyros and then computationally deriving the transformation matrix.
2. Measure specific force using accelerometers in body frame and transform it to reference navigation frame.
3. Model gravitational acceleration in a computer.
4. Sum the specific force and gravitational acceleration and time integrate the resultant acceleration information to obtain velocity and position after incorporating appropriate initial values of them at the start of integration.
5. Derive attitude information from the computed data on transformation matrix.
6. All the computations to be updated at frequencies that depend on the vehicle dynamics.

1.3.7 Navigation Errors

Like any other instrumented systems, computed navigation information has errors in them due to errors in gyros and accelerometers. The problem is that the navigation errors increase with time due to integrations involved in the computation. For aircraft flights, the error is often expressed in Nautical Miles Per Hour (NMPH) and typical specified navigation accuracy is around 1 NMPH. It means that after one hour of flight, the navigation system is expected to indicate position which will lie within a radius of one nautical mile (1.852 km) from the true position. This indicates that the navigation error specification is closely linked to elapsed time in navigation or flight time.

As a result of this integration of error, it becomes absolutely necessary to reduce the sensor errors as much as possible. A sensor can be modelled with different types of error and the model varies from one type of sensor to another. One common error is defined as the bias error. Bias defines a sensor output when input to the sensor is not applied.

In gyro, the term 'drift' is used to describe its error and its unit is expressed in °/h rather than in the SI unit of rad/s. In a simplified explanation for error, it means that the gyro is no more maintaining the inertial orientation and causing the instrumented frame to drift away from the reference frame. This drifted frame then leads to error in navigation. For a navigation system with one NMPH performance, the gyro should have drift rate below 0.01°/h and accelerometer bias below 100 μg and these represent very good measurement precision. Difficulty in such precision can be appreciated when we realise that the dynamic operating range of such gyros are greater than 10^7, and for accelerometer, it is more than 10^6. A dynamic operating range is defined as the ratio of maximum operating range to its detectable threshold. Further, such performance is to be ensured under hosts of adverse operating environments which increase drastically, the sensor cost.

1.3.8 Inertial Sensors

Inertial sensors, consisting of gyros and accelerometers, are the essential instruments in inertial navigation where gyro data provide information leading to instrumenting the reference frame, while the accelerometers provide the measurements of specific force. Besides the use in inertial navigation, these inertial sensors provide other useful information, e.g., vehicle body rate and body frame acceleration for stabilising and control of the vehicle. Different concepts and branches of physics are involved in their principle of operation, and over the last few decades, technological refinements have replaced the older versions with the new one.

Not long ago and still in use but declining, successful navigation grade gyros used principle that is based on conservation of angular momentum. There were various types in them and the most widely used types are called rate gyro and rate integrating gyro. A view of such a gyro is shown in Figure 1.6 with various features. Inertial angular rate about the input axis generates precession about the output axis, and a measurement of precession provides the desired information on the input angular rate.

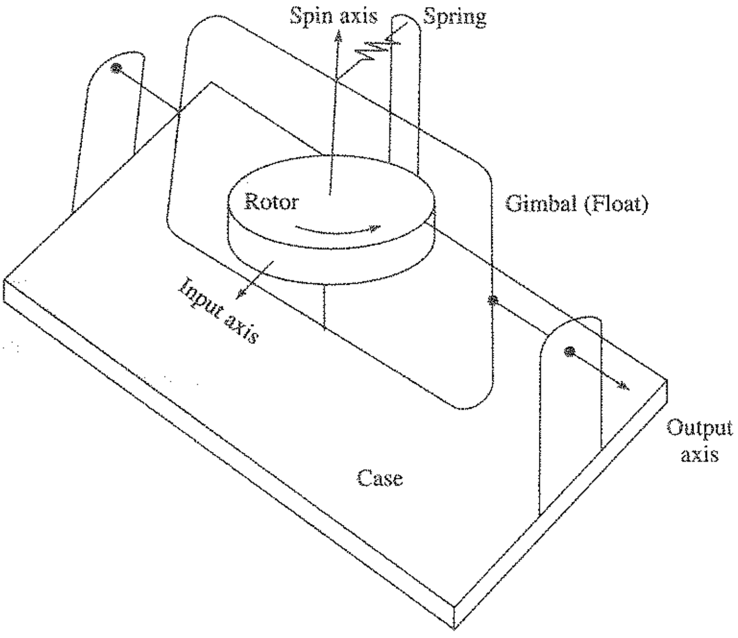


Figure 1.6 Gyro based on conservation of angular momentum.

Another later version in this group is called dynamically tuned gyro, where, as the name implies, the speed of the gyro is set against some gyro parameters, called tuning, to realise very good performance.

Later on, with the advent of laser, Sagnac effect gyros like ring laser gyro and fiber optic gyro became operational. Collectively, these gyros are also called optical gyros and since no rotor rotation is involved, such gyros are also referred as solid state gyros. A view of a ring laser gyro is shown in Figure 1.7, with various functional features. In this gyro, which is shown to be configured with three-mirror based optical resonator, the inertial rotation about the plane perpendicular to the resonator plane, produces a difference in frequency between the CW and the CCW laser beams and the readout system detects this frequency shift and then converts it into an electrical output using the photodiode.

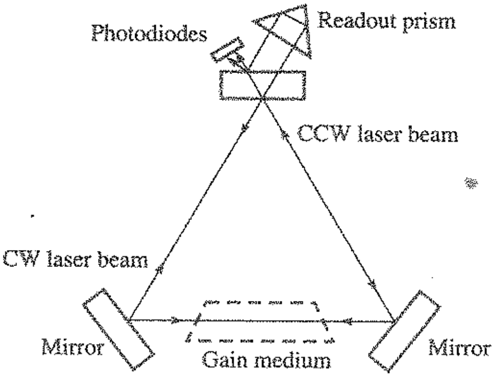


Figure 1.7 Ring laser gyro based on Sagnac effect.

Fibre optic gyro emerged as it was thought that use of optical fibre would dramatically reduce the cost of the gyro in comparison to ring laser gyro where expensive materials were used in addition to critical manufacturing process. The view of a fibre optic gyro is shown in Figure 1.8 displaying some of its functional elements. This gyro is configured as an integration of these functional elements like laser source, coupler, fibre coil, modulator and detector. The coupler splits the laser beam into CW and CCW beams and then recombines. Inertial rate, perpendicular to the plane defined by the fibre coil, causes a phase shift in the recombined laser beam, and the phase shift is detected by the detector. The functions of the other elements are required to suppress or eliminate some errors in this gyro. The laser source being outside, the gyro is also called passive fibre optic gyro as against the active laser gyro shown in Figure 1.7.

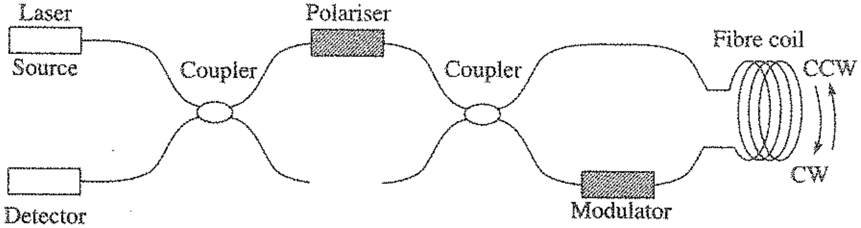


Figure 1.8 Fibre optic gyro based on Sagnac effect.

Another gyro concept has emerged that is based on the physics of Coriolis force which was propounded by the French scientist, GG Coriolis and so named after him. Such gyros, that have become operational, are known as 'Coriolis vibratory gyro' and a schematic of such a gyro is shown in Figure 1.9. In this gyro, the proof mass is made to resonate at its natural frequency, and when an inertial rate acts in direction perpendicular to the velocity of the proof mass, a Coriolis force is produced that is perpendicular to both the vibration velocity as well as the inertial rate. Detection of this oscillating force is then made to measure the input rate.

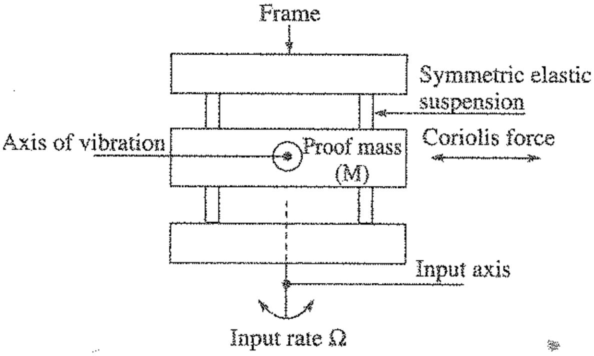


Figure 1.9 Illustration of a Coriolis vibratory gyro.

The most accurate gyro in this type is called hemispherical resonator gyro. However, much bigger impact is provided by such type of Coriolis gyros which are produced using MEMS technology, the technology that enables large scale production under low cost yet the sensors are highly reliable.

Accelerometers primarily use concept where acceleration acting on a proof mass produces reaction force, which is detected by various schemes. The schematic diagram of an accelerometer is shown in Figure 1.10.

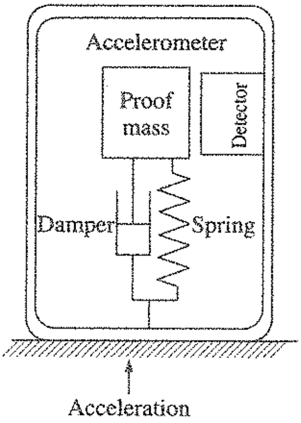


Figure 1.10 Spring mass accelerometer schematic.

Here, the inertial force on the proof mass due to the input acceleration is opposed by the mass, damper and the spring. The resultant steady state displacement of the proof mass is then detected by the detector as a measure of acceleration. Such an accelerometer is normally called open loop as the detector directly provides a measure of acceleration. However, such accelerometers, incorporating force feedback, are very accurate where the feedback current is a measure of acceleration and are normally used in navigation application.

Another version of accelerometer, which came out much later, uses change of frequency of a quartz resonant beam structure when subjected to an acceleration induced force. Such accelerometers are called 'vibrating beam accelerometer'. The schematic view of this type of accelerometer is shown in Figure 1.11. In this scheme, a thin section beam made of high Q material is attached at one end to proof mass and the other end to a fixed end support. This beam is made to resonate in lateral direction with electrical forcing at a frequency which is called no load beam frequency. When acceleration of magnitude a in the direction of the length of the beam is applied, the no load resonant frequency of the beam changes.

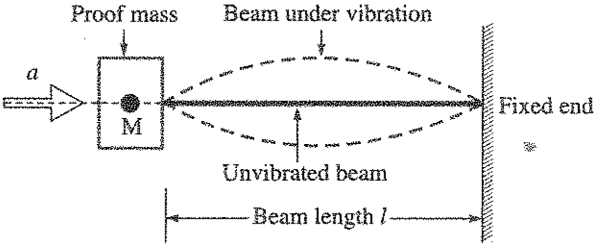


Figure 1.11 Accelerometer based on force to frequency conversion.

The detection of the change in the resonant frequency becomes a measure of the acceleration. Currently, silicon, in lieu of quartz, is preferred due to exponential growth in silicon fabrication process.

Research on Cold atom interferometry based gyros and accelerometers is at an advanced phase of development. Atoms behave like wave and the relationship, between the atom, which is a matter, and the wave, is governed by 'De Broglie' equation. This wave like behaviour of atom is further exploited using interferometric scheme after cooling close to absolute zero, to measure inertial rotation or acceleration. A basic scheme depicting an accelerometer operation with a Mach-Zehnder type of interferometer is shown in Figure 1.12. In the presence of acceleration, the recombined atomic wave exhibits a phase shift, that leads to interference fringes which is a measure of acceleration. Similaraly, for gyro, a Sagnac interferometer is instrumented to measure the rotation. These inertial sensors are capable of providing extremely high accuracy and are targeted for use where satellite aided navigation is not feasible.

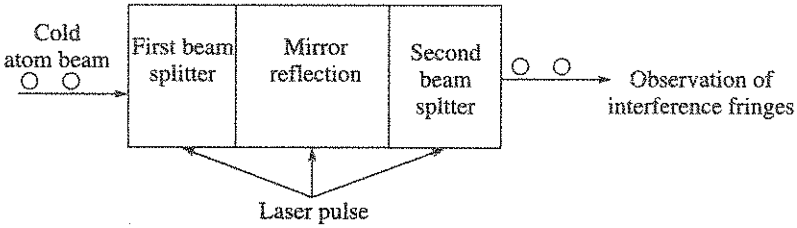


Figure 1.12 Cold atom accelerometer with interferometric detection.

There are other versions of inertial sensors which are under operation, but for application where precision accuracy is needed, the types described earlier, constitute the majority.

While minimisation of errors like drift in gyro or bias in an accelerometer are essential for defining a precision inertial sensor, parameter like high dynamic operating range also describes the characteristic of the precision sensors. A host of terminologies have evolved over time to describe the inertial sensor features and all these are listed at the end of the book.

1.4 MEMS based Inertial Sensors

The micro-fabrication technology, which has been the key to the success of microchips and microelectronics, is now revolutionising the microsystems involving both microelectronics and micro-mechanical components. The basis for this revolution has been the excellent mechanical properties of silicon and its suitability for batch processing miniaturised mechanical devices using the already well established processing techniques for microelectronics devices and a few additional processes which are today referred to as micromachining. Etching out portions of silicon or any other material to realise the miniaturised mechanical structure is one of the several micromachining techniques. Miniature systems involving one or more micromachined microscale devices are generally referred to as MEMS which stands for Micro-Electro-Mechanical Systems or simply the microsystems. While MEMS technology based sensors are numerous in types with ever increasing applications, the technology is used in the realisation of critical inertial sensors with improvement in performance seen with sustained thrust in R&D. Coriolis vibratory gyro concept is the backbone to this technology while realising gyros, but accelerometers of all types have come under this technology. The current spurt in low cost integrated inertial navigation in aerospace is primarily due to availability of miniaturised GPS receivers and MEMS based inertial systems.

1.5 Satellite Navigation

Satellite Navigation System (SNS) provides extremely accurate three-dimensional position and velocity information to users anywhere in the world. By trilateration, position determination is based on the measurement of the transit time of RF signals by the speed of light from a minimum of three satellites, which are orbiting around the globe. A receiver is used to measure this transit time, and when this measurement is multiplied with the velocity of light, the range to the particular satellite is obtained. A minimum of three such error free range measurements of three different satellites, whose positions are accurately known, can provide a non-ambiguous position of the receiver. This is explained along in Figure 1.13.

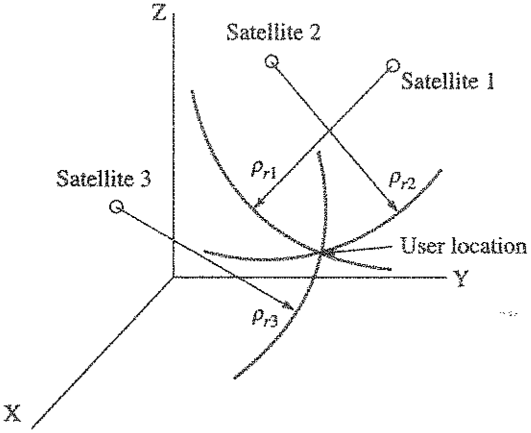


Figure 1.13 Satellite ranging and user position determination.

The three range measurements ρ_{r1}, ρ_{r2}, ρ_{r3} are actually the equations of three spheres with centres at the corresponding satellite locations, and they intersect at two locations only. One intersection gives the true user position, while the other intersection, far away from the true, is eliminated by software technique.

Receiver clock is not an accurate clock, unlike the highly accurate atomic clock used in satellites, as it has bias error. Just imagine that even 10 μs clock error will lead to range error of 3000 m. So clock bias leads to error in the range measurements and the observed ranges are then called *pseudo ranges*. The *pseudo ranges* in turn lead to uncertainty in the position determination. To circumvent this problem, the clock bias is estimated by taking a fourth *pseudo range* measurement from a fourth satellite. Thus visibility of four satellites and range measurements from these four different satellites become essential in three-dimensional position determination. Satellite position, called satellite ephemeris, is sent by the satellite to the receiver.

Satellite navigation system has shown very high position accuracy in the range of 30–40 m. Typically, the navigation frame used is known as earth centred earth fixed frame, and as a result, the shape of earth is to be accurately defined. It is important to note that the computed position error is a random noise and does not grow with time.

1.6 Integrated Inertial Navigation

Integrated navigation is another form of modern navigation where data from complementary navigation sensors are used to improve the navigation accuracy, and when one of the data source is an inertial system, it is often globally referred as integrated inertial navigation system or specifically an aided inertial system.

In an integrated inertial navigation system, the unbounded error growth with time in an autonomous inertial navigation system is corrected with the help of another external measurement. An external measurement can be of any quantity that duplicates a navigation parameter such as velocity, position or orientation, often referred as 'states'. Each external measurement, provided by the Aiding System, is compared to the value computed by the INS and the difference is used in an estimator, typically a Kalman filter, to estimate the error in the INS, which is further used to provide a corrected INS output. The feature of the scheme is shown in Figure 1.14.

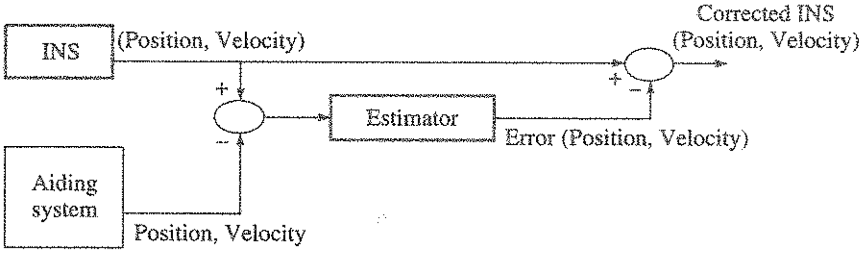


Figure 1.14 Feature of integrated inertial navigation.

The biggest benefit in an integrated inertial navigation is the emergence of cost effective system without compromise on performance and reliability. Emergence of low cost MEMS based inertial sensors is generating lots of thrust on integrated navigation.

1.7 Signal Processing of Inertial Sensors

Dynamic operating range, for high precision inertial sensors used for navigation, is very high and touches ten million. Normally, the sensors operate over a single range and as a result extremely low level of signal is encountered at low angular rate and acceleration. The analog signal detection level is around nano volt, while the capacitance variation to be detected is around femto Farad. In amplitude control, the amplitude variation is maintained to within a few parts per million. These signals are processed to suit the various applications for which they are to be used. These operations include, filtering, enhancement, digitisation, compression, and many other methods. Although enhancement and filtering can be carried out in analog domain, the need for digitising the signal becomes necessary because it provides scope to use more complex algorithm to improve signal-to-noise ratio (S/N), provides compression, provides emphasis and de-emphasis, all of which can be carried out easily using a digital signal processor.

Digital Signal Processing (DSP) is the mathematical operation that uses different algorithms and various other techniques which are necessary to manipulate the signals after they have been converted to a digital form. Statistics and probability play an important role in digital signal processing applications. Low level signals received from sensors are normally submerged in noise or interference. These corrupted signals can be processed, using statistical methods and probability, to remove the noise components without losing the intelligible information.

SUMMARY

The chapter has introduced various forms of modern navigation and provided an historical evolution on these systems. The chapter has also introduced the basics of modern navigation involving reference frames, shape of earth, gravitation and the time standard. The basic operation of strapdown navigation system and the navigation error propagation with the sensor errors are introduced. Various types of modern gyros and accelerometers are brought out which have primary application in aerospace. It introduces the emerging field of MEMS based inertial sensors. The basic concept in determining position with satellite navigation is described. The scheme of satellite navigation integrated inertial navigation, to reduce time dependent error in inertial navigation, is introduced, which is largely exploited towards making low cost navigation system. The significance of signal processing of inertial sensor to enable it to meet its performance in navigation is brought out.

EXERCISES

1.1 Explain the term sidereal earth rate and what is its northerly direction magnitude at the equator and at the north pole? [**Ans:** 15.04°/h; 0]

1.2 (a) Shape of earth defines its gravitation model, is it true or false? [**Ans:** True]

(b) In INS as well as in satellite based navigation system, precise knowledge of the shape of earth as well as the gravitational acceleration is necessary. True or false? Briefly explain the correct answer. [**Ans:** For INS: **True**; For satellite based navigation: **False**]

1.3 (a) Explain the difference in the characteristics of error between INS and satellite based navigation system.

(b) In an INS, there is only one error in the X-channel which is the bias in the accelerometer. If the error magnitude is 1 mg, what will be the velocity error and the position error in that channel after 100 s of navigation?

[**Ans:** Velocity error: 0.98 m/s; Position srror: 49 m]

REFERENCES

Graven, P.H., Collins, J.T., Sheikh, S.I., and Hanson, J.E., Spacecraft Navigation using X-ray Pulsars, 7th International ESA Conference on Guidance, Navigation and Control Systems, Tralee, Ireland, 2008.

Parkinson, W. and Bradford, B., Origins, Evaluation and Future of Satellite Navigation, *Journal of Guidance Control and Dynamics*, **20**(1), 1997.

Parvin, Richard H., *Inertial Navigation*, D. Van Nostrand Company, New York, 1962.

2

Autonomous Strapdown Inertial Navigation

Navigation is the determination of a physical body's position and velocity, and attitude expressed in some reference co-ordinate frame. To express the above states of a body, a suitable reference frame is to be chosen where the choice depends on the particular application. In order to implement an inertial navigation scheme, we need inertial sensors called gyros and accelerometers. The word 'Autonomous' has been used to describe a navigation system which uses only gyros and accelerometers to derive the measurements in order to compute position, velocity and attitude. Currently, two types of INS are in operation, and they are called 'gimbaled platform navigation' and 'strapdown navigation' respectively. However, majority of modern navigation systems are strapdown in nature and will be discussed in the subsequent sections.

2.1 Inertial Acceleration and Gravitational Acceleration

Newton's laws of motion describe the acceleration of objects with finite size. The *modern* understanding of Newton's first two laws of motion is as follows:

First Law

When observed from an inertial reference frame, if the particle moves without any change in velocity, then the force acting on the particle is zero.

Second Law

Observed from an inertial reference frame, the net force on a particle is proportional to the time rate of change of its linear momentum. Momentum is the product of mass and velocity. This law is often stated as $F = ma$ (the force on an object is equal to its mass multiplied by its acceleration).

It is to be noted that both the laws only hold when the observation is made from an inertial reference frame. Newton also formulated the law of gravitational attraction between the two masses and also computed the magnitude of the gravitational acceleration for a freely falling body on earth.

Later on, Einstein in his '*Principle of relativity and equivalence principle*' has explained that an observer on the surface of the earth could not find any difference between the gravitational attraction of earth and the inertial force that he feels when he is in a rocket in outer space that accelerates upwards (from the standpoint of the observer). In other words, he may regard any inertial force as a gravitational force. The concept of equivalence principle is described further to derive the principle of inertial navigation using illustration which is known as Einstein lift experiment.

Consider a mass M that is hanging (Figure 2.1) from the roof of an elevator using a spring and the mass is at rest. The downward force of gravitation is Mg, where g is the gravitational acceleration. Under this condition, the spring is in tension.

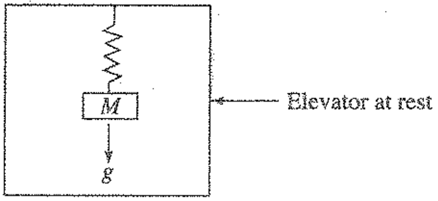


Figure 2.1 Spring–mass accelerometer with elevator at rest.

When the lift is accelerating upward with acceleration A, then also the force experienced by the mass is a downward pull so that the spring is again in tension (Figure 2.2).

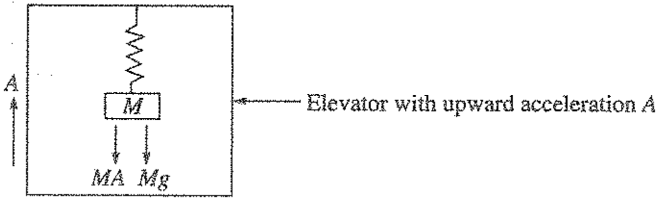


Figure 2.2 Elevator under upward acceleration.

Hence, under this condition and assuming upward acceleration as positive, the spring deflection measures the difference between the two forces, which is $(MA - Mg)$. So, the accelerometer measurement f is given by

$$F = A - g \tag{2.1}$$

Inertial navigation is based on this observation and formulation.

2.2 Inertial Navigation

In this section, the mathematical formulation on the principle of inertial navigation will be developed that provides information of position and velocity of a vehicle with respect to an orthogonal co-ordinate frame that is considered as inertial. Figure 2.3 shows an inertial and orthogonal co-ordinate frame with axes XYZ. At a point of time, an accelerating vehicle is at a position described by the position vector **R** from the origin of the frame.

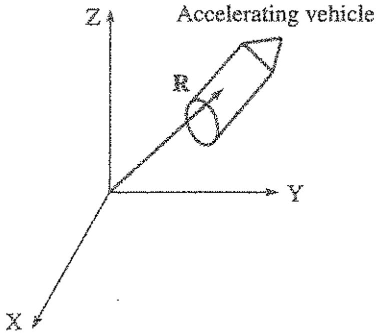


Figure 2.3 Navigation in inertial frame.

Noting that acceleration and displacements are vectors, the relationship between the acceleration $\mathbf{A}^{\mathrm{I}}$ and the displacement vector $\mathbf{R}^{\mathrm{I}}$ of the vehicle in inertial frame can be written as:

$$\mathbf{A}^{\mathrm{I}} = \frac{d^2\mathbf{R}^{\mathrm{I}}}{dt^2} \tag{2.2}$$

Accelerometers mounted on the vehicle provide information on specific force vector $\mathbf{f}^{\mathrm{I}}$ only. Specific force, that is measured by accelerometer, is defined as the force per unit mass and consists of any or combination of inertial forces such as thrust, drag and aerodynamic, which act on a flying vehicle but excludes measurement of Newtonian gravitational acceleration. Extending the inertial navigation formulation shown in Eq. (2.1), Eq. (2.2) can be expressed as:

$$\mathbf{A}^{\mathrm{I}} = \frac{d^2\mathbf{R}^{\mathrm{I}}}{dt^2} = \mathbf{f}^{\mathrm{I}} + \mathbf{g}^{\mathrm{I}} \tag{2.3}$$

where $\mathbf{g}^{\mathrm{I}}$ is gravitational acceleration vector and the superscript I refers to an inertial frame.

The first integral of the equation gives the inertial velocity, and it can be written as:

$$\dot{\mathbf{R}} = \int_0^t \mathbf{f}dt + \int_0^t \mathbf{g}dt + \dot{\mathbf{R}}(0) \tag{2.4}$$

where

$\dot{\mathbf{R}}$ = inertial velocity vector

$\dot{\mathbf{R}}(0)$ = initial velocity at $t = t_0$

Similarly, the second integral gives the position with appropriate initial condition.

Now there are two methods of implementation of *I-frame* using gyros. In earlier days, gyros implemented the frame using servo driven gimbals to isolate the cluster from vehicle rotational motion. Thus accelerometers, which are mounted on the cluster, are actually measuring specific force in *I-frame* as shown in Eq. (2.3). Currently, inertial systems dispense with such gimbals and allow accelerometers to measure the specific force in body frame, and it is no more inertial. Transforming the specific force measured in body frame to inertial frame is done analytically using information provided by the gyros and the system is called analytic gimbaled system, or more popularly, called strapdown system, which is discussed further in Section 2.7.

Since the vehicle borne accelerometers provide the measurement of specific force $\mathbf{f}^I$ only, the gravitational acceleration $\mathbf{g}^I$ is modelled and computed. Equation (2.3) is perfectly generalised and is implementable in a computer so long as the gravitational acceleration can be modelled. In the vicinity of earth, gravitational field of earth dominates over the gravitational forces of sun and moon, and hence their effects are neglected for near earth inertial navigation. So, modelling of earth's gravitational acceleration $\mathbf{g}^I$ as a function of vehicle position becomes a necessity. If the vehicle is predominantly in the gravitational field of moon, then appropriate model of $\mathbf{g}^I$ (moon) is required for moon-based inertial navigation. This brings out the singular importance of gravitational acceleration model in the context of inertial navigation and is discussed further in Section 2.3.

2.3 Geometry of Earth, Gravitation and Gravity

Earth's gravitational field at different places on and above earth is closely linked to the geometry of the earth. So, it becomes necessary to accurately model the geometrical aspects of the earth. The assumption of spherical earth model (refer Figure 2.4) provides the simplest Newtonian gravitational model that could be used in applications where accuracy requirement is not high.

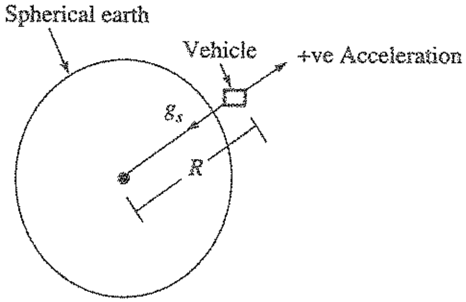


Figure 2.4 Spherical and homogeneous earth model gravitation.

For a vehicle at a distance R from the centre of a homogeneous and spherical earth, the gravitational acceleration vector $\mathbf{g}_s$ directed towards the centre of earth is given by

$$\mathbf{g}_s = -\frac{\mu}{R^3}\mathbf{R} \tag{2.5}$$

where μ is called gravitational constant that is defined as the product of earth mass and the universal gravitational constant. The negative sign is used where, by choice, acceleration away from earth is considered as positive. The term 'homogeneous' is used if the mass distribution of earth is totally uniform so that it behaves like a homogeneous body.

If, x, y, z are the co-ordinates of the vehicle from the centre of earth as origin, the resolved components of the gravit tional acceleration components can be expressed as $(g)x/R$, $(g)y/R$ and $(g)z/R$, where $(g) = \mu/R^2$. On substitution of (g), we get the components as:

$$g_x = -\mu \frac{x}{R^3}$$
$$g_y = -\mu \frac{y}{R^3} \qquad (2.6)$$
$$g_z = -\mu \frac{z}{R^3}$$

$$R = \sqrt{x^2 + y^2 + z^2} \qquad (2.7)$$

Over millions of years, due to rotation of earth about the earth's polar axis, a bulging had occurred around the equatorial region whereby the equatorial radius is ~22 km more than the polar radius, and owing to this slight flattening of the earth at the poles, spherical model assumption of earth is not sufficient to meet high accuracy navigation. It is customary to mathematically model the earth as a reference ellipsoid, approximating closely to the true geometry.

In the ellipsoid geometry of earth system, the earth is approximated as an ellipsoid, which is generated when the ellipse is rotated about its minor axis passing through true north and south poles. Figure 2.5 shows the planar view of the ellipsoid for a section taken along the meridian plane. This shape of earth is a complex mathematical model, can be appreciated from the undulation of the terrain with numerous hills and valleys on the surface of an actual earth.

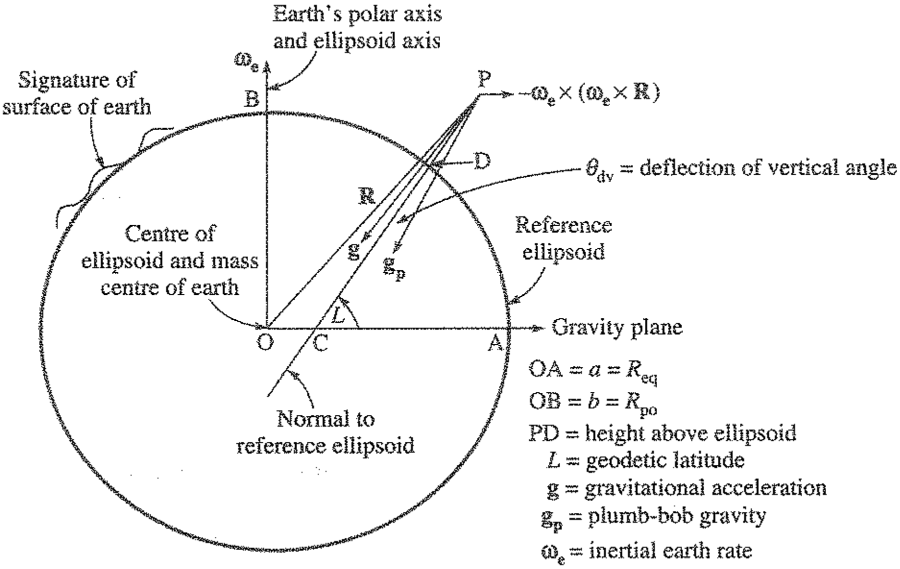

Earth's polar axis and ellipsoid axis
ω_e
Signature of surface of earth
B
P
$-\omega_e \times (\omega_e \times \mathbf{R})$
D
θ_{dv} = deflection of vertical angle
R
Centre of ellipsoid and mass centre of earth
g
g_p
Reference ellipsoid
L
O
C
A
Gravity plane
Normal to reference ellipsoid
OA = $a = R_{eq}$
OB = $b = R_{po}$
PD = height above ellipsoid
L = geodetic latitude
g = gravitational acceleration
g_p = plumb-bob gravity
ω_e = inertial earth rate

Figure 2.5 Reference ellipsoid, gravitational acceleration and gravity.

Based on enormous amount of experimental observations, the WGS-84 (World Geodetic System-84) has defined the reference ellipsoid for world wide navigation. Some of the parameters that define this reference ellipsoid are as follows:

- Semi-major axis length a (equatorial radius, R_{eq})
- Semi-minor axis length b (polar radius, R_{po})

- Ellipticity e
- Eccentricity ε

Ellipticity of the ellipsoid and eccentricity are defined as:

$$e = \frac{R_{\text{eq}} - R_{\text{po}}}{R_{\text{eq}}}; \quad \varepsilon = \sqrt{1 - \frac{R_{\text{po}}^2}{R_{\text{eq}}^2}} \tag{2.8}$$

2.3.1 Gravitation Ellipsoid

Gravitation ellipsoid is a mathematical model of earth's gravitational acceleration arising out of ellipsoid geometry model of earth, where the gravitation vector does not pass through the centre of ellipsoid. The gravitation ellipsoid model is complex due to the incorporation of the higher order terms like second gravitational constant J_2. A typical *gravitation ellipsoid* model **[Britting, 1971]** with J_2 term and in component form is shown below at a point with co-ordinates (x, y, z) where the co-ordinates are defined in ECI frame.

$$g_x = -\frac{\mu}{R^2}\left[1 + \frac{3}{2}J_2\left(\frac{R_{\text{eq}}}{R}\right)^2\left(1 - 5\frac{z}{R}\right)^2\right]\frac{x}{R} \tag{2.9}$$

$$g_y = -\frac{\mu}{R^2}\left[1 + \frac{3}{2}J_2\left(\frac{R_{\text{eq}}}{R}\right)^2\left(1 - 5\frac{z}{R}\right)^2\right]\frac{y}{R} \tag{2.10}$$

$$g_z = -\frac{\mu}{R^2}\left[1 + \frac{3}{2}J_2\left(\frac{R_{\text{eq}}}{R}\right)^2\left(3 - 5\frac{z}{R}\right)^2\right]\frac{z}{R} \tag{2.11}$$

It is appropriate to mention here that WGS-84 is the largely used geodetic model for inertial navigation and Table 2.1 **[US Defense Mapping Agency, World Geodetic System, 1984]** provides values of various constants for use in the gravitation ellipsoid model. For the ellipsoid model, R is measured from the centre of the ellipsoid earth. The position components (x, y, z) are defined in earth centred inertial frame. While with J_2, an accurate ellipsoid gravitational model is available, for more navigation accuracy with precision inertial sensors, additional higher order terms, like J_3 and J_4, are recommended **[Chatfield, 1997]**.

Table 2.1 WGS-84 Geodetic and Gravitational Parameters

Parameter	Symbol	Unit	Value
Defining parameters			
Equatorial radius	$R_{\text{eq}}(a)$	m	6378137.0
Angular velocity	ω_e	rad/s	7.292115×10^{-5}
Earth's gravitational constant	μ	m^3/s^2	$3.98600448 \times 10^{+14}$
Second gravitational constant	J_2	m^3/s^2	1.0826268×10^{-3}
Polar radius	$R_{\text{po}}(b)$	m	6356752.3
Derived constants			
Ellipticity	e	—	1/298.2572
First eccentricity	ε	—	0.081819190
Gravity at equator	g_{eq}	m/s^2	9.7803267714
Mean value normal gravity	g_n	m/s^2	9.7976446561

2.3.2 Gravity

Gravity is defined as the combined effect, vector sum, of the gravitational acceleration vector $\mathbf{g}$ of the earth mass and the inward centripetal acceleration due to earth's angular rate vector $\boldsymbol{\omega}_e$ in inertial space The direction of the gravity vector is the direction pointed by a plumb bob that is suspended above the earth, and hence it is also termed as plumb-bob gravity $\mathbf{g}_p$ and is given by:

$$\mathbf{g}_p = \mathbf{g} - \boldsymbol{\omega}_e \times (\boldsymbol{\omega}_e \times \mathbf{R}) \tag{2.12}$$

Gravity $\mathbf{g}_p$ is shown in Figure 2.5 for a mass suspended at a point P and at a distance $\mathbf{R}$ from the centre of the ellipsoid. On the surface of earth, it is the combined effect that is visible. For an ideal and homogeneous ellipsoid, the gravity vector is considered perpendicular to the reference ellipsoid, which means that it has no horizontal components on the ellipsoid surface. In reality, this is not so and this aspect is discussed further in Section 2.3.3. The line normal to reference ellipsoid intersects the equatorial plane at point C and the included angle is called the geodetic latitude L.

On the surface of ellipsoid earth and at sea level, gravity model is a function of latitude L. This has been accurately modelled by **Helmert's** equation **[International Gravity Formula]** as:

$$g(L) = 9.780326\lfloor(1 + 0.0053\sin^2 L - 5.9\times 10^{-6}\sin^2 2L)\rfloor \text{m/s}^2 \tag{2.13}$$

Since gravity decreases with altitude, a correction factor with a negative sign is employed for height h in metres above the sea and is usually given by -3.086×10^{-6} h and Eq. (2.13) to be used with this correction term where altitude is involved.

For high altitude, the gravity magnitude g_h is expressed as:

$$g_h = g_0\left(\frac{R_{\text{av}}}{R_{\text{av}} + h}\right)^2 \tag{2.14}$$

where R_{av} is defined as the average radius of the ellipsoid earth with the accepted magnitude of 6371 km, and g_0 is the mean value of normal gravity on the surface of earth having the accepted magnitude of 9.797 m/s^2 (refer Table 2.1).

EXAMPLE 2.1 To find gravity at the pole and gravitational acceleration at the equator.

Solution: This can be found out using Eq. (2.12), Eq. (2.13) and Table 2.1.

Equation (2.13) can be used to find out the gravity at the pole using $L = 90°$. We get:

$$g_{(\text{pole})} = 9.780\ [1 + 0.0053] = 9.8321 \text{ m/s}^2$$

The gravitational acceleration g at equator can be found out using Eq. (2.12) and Table 2.1.

By doing the vector multiplication, Eq. (2.12) can be reduced to its component form as:

$$g_p = g - \omega_e\omega_e R_{\text{eq}}$$

Noting the gravity magnitude at equator and earth rate magnitude ω_e from Table 2.1, we write

$$9.780 = g - \omega_e\omega_e R_{\text{eq}}$$

where

$\omega_e = 7.292 \times 10^{-5}$ rad/s, and
$R_{\text{eq}} = 6.378 \times 10^6$ m

$$9.780 = g - 7.292 \times 7.292 \times 10^{-10} \times 6.378 \times 10^6$$

Solving for the gravitational acceleration at equator, we get

$$g = 9.780 + 0.0339 = 9.8139 \text{ m/s}^2$$

The computation shows that the sea level gravity increases from 9.780 m/s^2 at the equator to 9.832 m/s^2 at the poles. The increase at the pole is due to absence of centripetal acceleration and shorter polar radius, both of which increase the gravity magnitude.

EXAMPLE 2.2 To find the earth rate components at the equator and at the pole.

Solution: The inertial earth rate vector ω_e, as shown in Figure 2.5, can be resolved on an earth based frame, e.g., north, east and vertical at a point on earth whose latitude is L. The resolved earth rate components are as follows:

Along north	: $\omega_e \cos L$
Along east	: 0
Along vertical	: $\omega_e \sin L$

Therefore, at equator ($L = 0$) the components along [north, east, vertical] are [15.04, 0, 0]°/h. At the pole ($L = 90$), the similar components are [0, 0, 15.04]°/h.

2.3.3 Gravity Anomaly

For earth conforming to a conceptualised homogeneous ellipsoid model, the plumb-bob gravity would have been perpendicular to reference ellipsoid. In reality, this is not so, as the deflection of vertical occurs due to difference between actual shape of earth and the reference ellipsoid model and due to inhomogeneous distribution of earth mass. The direction of plumb-bob gravity is then perpendicular to the gravity equipotential surface, which at mean sea level is termed as *geoid*. The angle θ_{dv} (refer Figure 2.5), between the plumb-bob gravity vector and the line normal to reference ellipsoid is called *deflection of vertical* whose value is variable over the earth and typically lying between 10 arc-secs to 40 arc-secs. This is called gravity anomaly and constitutes error in gravity computation as horizontal components of gravity are not accounted in the reference ellipsoid model described earlier.

2.4 Reference Frames

This section will further explore the features of various types of reference frames which are used in inertial navigation. We can primarily categorise the reference frames as under:

1. Inertial reference frame
2. Non inertial reference frame (rotating reference frame)

The navigation in an inertial frame has been discussed earlier, and the relationship between the derivative of a position vector **R** in a rotating frame E and the derivative of **R** in a non rotating frame I, is governed by Coriolis formulation. It can be stated as:

$$\frac{d\mathbf{R}^{\mathrm{I}}}{dt} = \frac{d\mathbf{R}^{\mathrm{E}}}{dt} + \boldsymbol{\omega} \times \mathbf{R} \tag{2.15}$$

where ω is the rotation rate vector of *E-frame* with respect to *I-frame* and the cross (×) represents vector cross product. This formulation will be used subsequently for rotating earth frame navigation.

2.4.1 Right Handed Orthogonal Frame

All the frames of reference, used for navigation, are right handed and orthogonal. The matrix inversion relationship in orthogonal frame is simple and is given by its transpose. For example, $C_B^I = [C_I^B]^T$, where C_I^B is a transformation matrix relating *I-frame* to *B-frame,* while C_B^I is the inverse matrix obtained simply by the transpose operation. This useful property of orthogonal matrix will find frequent application in navigation.

2.4.2 Inertial Reference Frames in Operation

A very important inertial reference frame of use is called Earth Centred Inertial (ECI) frame, which is shown with superscript (i) in Figure 2.6.

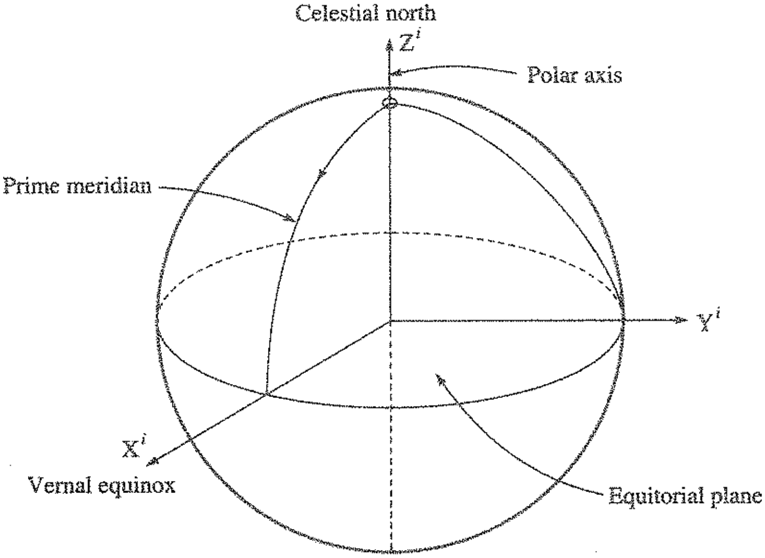


Figure 2.6 Earth Centred Inertial frame of reference.

It has its origin at the centre of ellipsoidal earth. The frame orientation is as follows:

X^i: along the direction of the mean vernal equinox at t_0 and lying on the equatorial plane.

Z^i: along the earth's spin axis pointing celestial north at epoch t_0.

Y^i: completes the right handed system and lies on the equitorial plane.

However, such a frame is treated as quasi inertial where the measurement inaccuracy is within 2 × 10^{-7} g in acceleration and 5 × 10^{-5} °/h in angular rate [**Kayton, 1997**]. This error is attributed to certain motions of earth, which include daily rotation about the earth axis, annual revolution about sun, precession and nutation and wandering of the poles. For strapdown inertial navigation, a more convenient form of ECI frame is used where X-axis is the direction of the line joining earth centre with prime meridian at epoch $t = 0$. It is known that prime meridian rotates with earth and not inertial, hence the need for defining the epoch. The frame remains inertial from this epoch. There is no change with respect to the other axes. The ECI frame is important for navigation involving satellite launch, interplanetary missions and also for modelling the gravitation that involves measurement from centre of earth.

There is one more inertial frame of importance, and this is called *Launch Point* (*or Launch Centred*) *Inertial* (*LPI*) *frame*, which is shown with superscript I. It has its origin coinciding with the location of the satellite launcher. The frame orientation is as follows (refer Figure 2.7):

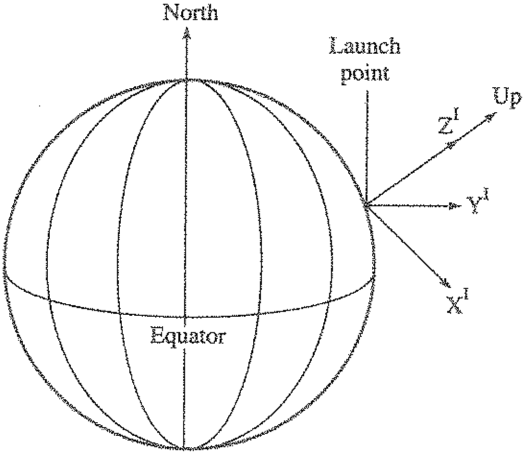


Figure 2.7 Launch Point Inertial frame.

X^I : along launch azimuth
Z^I : along local vertical and pointing outward
Y^I : completing the right handed system

The frame goes to inertial, at the vehicle lift off time T_0. At the navigation system level, the gyros maintain the inertial frame. The origin of the frame is not inherently inertial as in the case of ECI frame due to earth's rotation effect requiring appropriate velocity initialisation at T_0.

EXAMPLE 2.3 (on right handed frame): In Figure 2.7, if launch azimuth is 135°, show the direction of Y^I?

Solution: Now, X^I and Z^I are m utually perpendicular as azimuth axis X^I is defined to lie on earth's horizontal surface, and Z^I is perpendicular to local horizon. If we define the co-ordinate axes sequence as XZY, then using right handed frame property, Y^I axis will be horizontal at an azimuth of 225°. On the other hand, if the sequence of axes is defined as XYZ, then Y^I axis will be at an azimuth of 45°.

EXAMPLE 2.4 Compute velocity initialisation at lift off time in LPI frame.

Solution: The navigation system incorporating LPI frame (refer Figure 2.7), goes to inertial at lift off time $t = t_0$. Due to inertial rotation of earth, the system fixed to the vehicle body frame on earth, experiences surface velocity with respect to LPI frame at $t = t_0$. The direction of this velocity is eastward and its magnitude is given by $\omega_e \cos LR_e$ where R_e is the radius of earth at the launch point, and L is the geodetic latitude of the launch point.

When the launch azimuth (X direction) is a_z, the resolved components of the surface velocity at the launch point along the inertial frame axes are given by

$$V_x = \omega_e \cos L \cos(a_z - 90) R_e$$

$$V_y = \omega_e \cos L \sin(a_z - 90) R_e$$

$$V_z = 0$$

These equations provide the magnitudes of velocities to which the system needs to be initialised at $t = t_0$.

Considering, as an example, $L = 12°$; $a_z = 135°$; the surface velocities can be computed using the constants shown in Table 2.1. The computed velocities are:

$$V_x = [7.29 \times 10^{-5} \times \cos 12 \times \cos (135 - 90) \times 6.37 \times 10^{+6}] = +321 \text{ m/s}$$

Similarly, $V_y = +321$ m/s

These are the magnitudes of velocities which will go as input at the start of navigation computation in LPI frame. If the navigation system is sitting on earth at this location, it will indicate displacements of +321 m along X-axis and +321 m along Y-axis respectively, at the end of 1 s of navigation.

2.4.3 Rotating Frame of Reference

There are quite a few of such rotating reference frame in operation depending on application. A few of them are discussed as follows.

Earth Centred Earth Fixed (ECEF) frame (E-frame) (refer Figure 2.8), has its origin at the centre of earth. The Z^E-axis is along the earth's spin axis, while the remaining axes (X^E, Y^E) are fixed to the earth on the equatorial plane. X^E-axis intersects the sphere of the earth at 0 degree latitude and 0 degree longitude. Y^E-completes the right handed triad.

The axes X^E and Y^E rotate at the earth's spin rate magnitude of 15.041°/h with respect to the ECI frame. Since the axes are fixed to the rotating earth which thus provides the name of the frame. In Figure 2.8, $\omega_e t$ shows the rotation of the *E-frame* axis from ECI axis due to inertial rotation of earth. ECEF frame is right handed and orthogonal and finds use in satellite navigation.

Geographic frame (G-frame) is a navigation frame with its origin located at the INS location. Its axes (Z^G, Y^G, X^G) are aligned with the local vertical up (U), direction of north (N) and east (E) respectively. It is a right handed co-ordinate system when defined as ENU. The north axis Y^G is in the direction of the earth's angular inertial velocity vector projected on the local horizontal plane, and being a right handed frame, the X^G axis also lie on the horizontal plane.

The frame is often referred as local level north pointing. Another variation to geographic frame is as north, east and vertical down, in short NED. A third variation is defined as north, west and vertical (up), in short NWV. All these frames conform to right-handed orthogonal co-ordinate system. The word geographic is used as the horizontal position components are computed as geodetic latitude and longitude and the vertical component as altitude. The turn rate of this frame with respect to ECEF frame is often called transport rate. In this frame, the velocity is defined with respect to the earth model at the given instant. For an aircraft pilot, the navigation information in this frame is easy to understand, and as a result, these versions of geographic frames are widely used for terrestrial navigation involving air, land and sea.

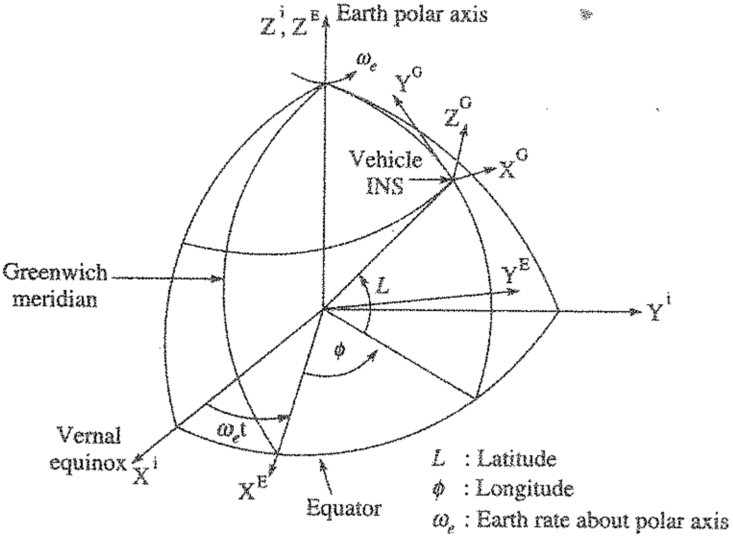


Figure 2.8 Earth Centred Earth Fixed and geographic navigation frames.

EXAMPLE 2.5 (on geographic frame navigation):

(i) An aircraft is stationary at a latitude of 0° and longitude of 100° east and has a INS navigating in geographic frame. What will be its indicated velocity?

(ii) If the aircraft takes off and flies east at 100 m/s for one hour on a level flight, what is the horizontal position INS will show?

Solution:

(i) The indicated velocity is relative to earth frame. Hence, it will be zero when stationary on earth.

(ii) Since velocity is relative to earth frame, INS indicated distance travel will be 100 × 3600 m along equitorial east. Assuming a spherical earth, taking the equatorial radius from Table 2.1, and neglecting the travel altitude relative to the equitorial earth radius magnitude 6378 km, longitude covered is 360/6378 rad = 3.23°. Since, longitude is positive in east direction, the indicated longitude will be 100 + 3.23 = 103.23° and the latitude remains at 0°. Geographic frame navigation is covered later in Section 2.9.5.

2.4.4 Reference Frame Relationship and Co-ordinate Transformation

Navigation information computed in one reference frame can be computed in another reference frame using a mathematical process called co-ordinate transformation. The choice of a reference frame is application driven such as the use of geographic frame in terrestrial navigation, inertial frame in space launches and ECEF co-ordinates for satellite navigation. All these reference frames are related to each other through suitable co-ordinate transformation between the two frames, so that navigation data available in one frame can be transformed using these relations. The basis of this transformation is explained initially.

Let a vector **V** has components, which are known, in an arbitrary reference frame, called *I-frame*, as follows:

$$[\mathbf{V}]^I = \begin{bmatrix} V_X \\ V_Y \\ V_Z \end{bmatrix}^I \tag{2.16}$$

It is desired to calculate the vector components in another frame called *B-frame*.

The *B-frame* components of vector $[\mathbf{V}]^I$ is computed from the relationship given by

$$[\mathbf{V}]^B = \left[C_I^B\right][\mathbf{V}]^I \tag{2.17}$$

where C_I^B is the direction cosine matrix that transforms a given physical vector represented in *I-frame* to the same physical vector [**V**] represented in *B-frame*. The direction cosine matrix will have a form as follows:

$$C_I^B = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{bmatrix} \tag{2.18}$$

The element in the *i*th row and *j*th column represents the cosine of the angle between the i-axis of the *B-frame* and the j-axis of the *I-frame*, and hence the name direction cosine. This method is used to generate transformation matrix between two frames. A demonstration on the use of the method is shown later in Section 2.7.3. Using this method, some useable transformation matrices are shown below:

(a) Geographic *G-frame* (ENU) to *E-frame* (ECEF) transformation matrix: The transformation matrix C_G^E is given by:

$$C_G^E = \begin{bmatrix} -\sin\phi & -\sin L\cos\phi & \cos L\cos\phi \\ \cos\phi & -\sin L\sin\phi & \cos L\sin\phi \\ 0 & \cos L & \sin L \end{bmatrix} \tag{2.19}$$

Here, L is the latitude and ϕ is the longitude.

(b) ECEF frame (*E-frame*) to ECI (*i-frame*) frame transformation matrix C_E^i relation: Assuming, $\phi_0 = 0$, where ϕ_0 is the initial longitude, the transformation matrix C_E^i is given by

$$C_E^i = \begin{bmatrix} \cos\omega_e t & -\sin\omega_e t & 0 \\ \sin\omega_e t & \cos\omega_e t & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.20}$$

EXAMPLE 2.6 (on navigation frame transformation): Position of a vehicle in geographic co-ordinate is on equator and at a longitude of 90° east. Compute the transformation matrix C_G^E.

Solution: The transformation matrix is given by Eq. (2.19). We have $L = 0$ and $\phi = 90°$ east. The computed C_G^E is given by

$$\begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

If the vehicle travels along east and the longitude becomes 180°, the new C_G^E matrix becomes:

$$\begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

2.5 Navigation in Rotating Earth Frame

Navigation in rotating earth frame is widely used for navigation in the vicinity of earth for aircrafts, ships and submarines. Using the fundamental navigation equation in inertial frame along with the Coriolis transformation method to a rotating frame, the derivation of navigation equation in rotating earth frame is shown hereafter.

The relationship between a vector measured in inertial frame I and the earth fixed rotating frame E is given by Coriolis equation [refer Eq. (2.15)]. This equation can be expressed as:

$$\left[\frac{d\mathbf{R}}{dt}\right]^I = \mathbf{V} + \boldsymbol{\omega}_e \times \mathbf{R} \tag{2.21}$$

where $\mathbf{V}$ is the velocity with respect to the earth fixed frame and $\boldsymbol{\omega}_e$ is the angular velocity of earth about inertial space. Differentiating Eq. (2.21), we get

$$\left[\frac{d^2\mathbf{R}}{dt^2}\right]^I = \left[\frac{d\mathbf{V}}{dt}\right]^I + \boldsymbol{\omega}_e \times \frac{d\mathbf{R}}{dt}\bigg|^I \tag{2.22}$$

Using Coriolis equation, we can further write

$$\left[\frac{d\mathbf{V}}{dt}\right]^I = \left[\frac{d\mathbf{V}}{dt}\right]^N + \boldsymbol{\omega}_N \times \mathbf{V} \tag{2.23}$$

where $\boldsymbol{\omega}_N$ is the navigation frame angular velocity with respect to inertial space.

Also,

$$\underline{\omega_e \times \frac{d\mathbf{R}}{dt}}\Bigg|^{I} = \omega_e \times \mathbf{V} + \omega_e \times (\omega_e \times \mathbf{R}) \tag{2.24}$$

Substituting Eqs. (2.23) and (2.24) in Eq. (2.22), we get

$$\left[\frac{d^2\mathbf{R}}{dt^2}\right]^{I} = \left[\frac{d\mathbf{V}}{dt}\right]^{N} + \omega_N \times \mathbf{V} + \omega_e \times \mathbf{V} + \omega_e \times (\omega_e \times \mathbf{R}) \tag{2.25}$$

Substituting Eq. (2.3) in Eq. (2.25) and rearranging gives:

$$\left[\frac{d\mathbf{V}}{dt}\right]^{N} = \mathbf{f} + \mathbf{g} - \omega_e \times (\omega_e \times \mathbf{R}) - (\omega_N + \omega_e) \times V \tag{2.26}$$

Substituting $\mathbf{g}_p$ from Eq. (2.12) in Eq. (2.26) gives:

$$\left[\frac{d\mathbf{V}}{dt}\right]^{N} = \mathbf{f} + \mathbf{g}_p - (\omega_N + \omega_e) \times \mathbf{V} \tag{2.27}$$

where $\mathbf{f}$ is the specific force transformed to the appropriate navigation frame using the transformation $C_B^N \mathbf{f}^B$ where $\mathbf{f}^B$ is the sensed specific force in body frame in a strapdown system.

Equation (2.27) is a generalised navigation equation in earth fixed rotating frame. The third term, in Eq. (2.27), is called the *Coriolis correction term* and like the gravity modelling, Coriolis correction is also modelled in computer. The difference between the inertial frame navigation represented by Eq. (2.3) and the earth fixed rotating frame navigation becomes clearer when one notes the following features in the rotating frame:

(a) Use of plumb-bob gravity model which reduces Eq. (2.26) to Eq. (2.27).

(b) Implementation of correction for Coriolis acceleration for the specific force that has been transformed to the navigation frame as needed in a strapdown system.

(c) If the navigation system is fixed to earth and allowed to navigate, ideally, it will continue to show zero velocity and no further change in position. This matches with the intuitive feeling of a terrestrial observer.

EXAMPLE 2.7 Derive what the vertical accelerometer senses when the navigation system is fixed to earth.

Solution: It is assumed that one accelerometer is in vertical orientation, whereas the remaining two are on the horizontal planes. Since the system is fixed on an earth rotating frame, Eq. (2.27) can be used to find out what the vertical accelerometer measures. As the system is stationary with respect to earth, $V = 0$ and $dV/dt = 0$. Equation (2.27) reduces to:

$$0 = f + g_p \quad \text{or} \quad f = -g_p$$

Hence the vertical accelerometer senses specific force f whose magnitude is g_p. But the system will not indicate integrated position or altitude, as the gravity model in the computer provides gravity compensation whose magnitude is same as that of the magnitude of specific force sensed by the accelerometer. However, when the system is in flight, the accelerometer measures only inertial acceleration. Similarly, when the system is fixed on a free falling vehicle, the accelerometer will not measure the gravitational acceleration.

2.6 Inertial Systems

So far, the basic aspects of autonomous navigation involving various types of reference frames and the modelling of gravitational acceleration and/or gravity have been covered. In order to instrument a navigation system, gyros and accelerometers are necessary. The instrumented system is quite often termed as *inertial system*. Now, there are primarily two types of inertial systems defined as:

1. Gimbaled platform system
2. Strapdown system

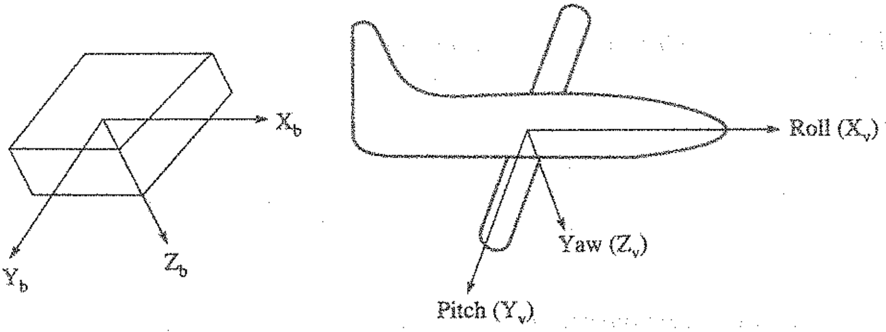
Historically, gimbaled platform system preceded the development of the strapdown system, but the majority of modern system is strapdown in nature and is described further in the subsequent sections.

2.6.1 Orthogonal Frames in Inertial System

Various navigation frames have been explained in the earlier sections. This section will explains additional co-ordinate frames in the context of functioning of strapdown inertial systems.

Inertial system body frame (B-frame) is defined as an orthogonal frame [refer Figure 2.9(a)] on which the measurement of inertial sensors are defined. Physically, the platform cluster provides accurate and physical orthogonal reference axes for the gyros and the accelerometers with the origin at the cluster centre of gravity. A cluster is an accurately machined three-dimensional structure. Hereafter in the book, the cluster frame has been used to define the strapdown system body frame (*B-frame*).

Vehicle body frame is an orthogonal frame fixed to the vehicle body [refer Figure 2.9(b)], and the axes conventionally termed as roll (X_v), pitch (Y_v) and yaw (Z_v), with the origin located at the centre of gravity of the vehicle.



(a) INS body frame (b) Vehicle body frame

Figure 2.9 Inertial system body frame and vehicle body frame.

We will assume for further description and analysis that the inertial system body frame axes are parallel with the vehicle roll, pitch and yaw axes with coincident origin. Even when the origins are not coincident, it is possible to mathematically tackle the problem of non coincident origin with suitable linear transformation that is applicable for a rigid body.

Sensor frame (S-frame) is defined by the input axes (sensitive axes) of the sensors. In reality, there are two sensor frames: one for the set of three accelerometers and the other for three gyro set, and none of these are usually orthogonal. The sensor input axes are required to get aligned with the cluster orthogonal axes either physically or computationally. When the sensor input axes are truly aligned with the cluster, then it represents the sensor orthogonalised frame with the origin coinciding with the cluster.

Computational frame (^) is often used in strapdown system to define an orthogonal frame that is slightly misaligned from the true navigation frame due to gyro drift errors. The small misalignment angles between the two frames define the error in attitude measurement.

2.6.2 Relationship between the Sensor Frame and the System Body Frame

The relationship between the sensor frame and the system body frame is a skew symmetric matrix that usually defines misalignment error between the two frames. For example, the angular rates sensed by the gyros in *S-frame* are transformed to *B-frame*, using the following transformation which assumes small angle approximation:

$$\begin{bmatrix} \dot{\theta}_X \\ \dot{\theta}_Y \\ \dot{\theta}_Z \end{bmatrix}^B = \begin{bmatrix} 1 & \Delta\theta_{XY} & \Delta\theta_{XZ} \\ \Delta\theta_{YX} & 1 & \Delta\theta_{YZ} \\ \Delta\theta_{ZX} & \Delta\theta_{ZY} & 1 \end{bmatrix} \begin{bmatrix} \dot{\theta}_X \\ \dot{\theta}_Y \\ \dot{\theta}_Z \end{bmatrix}^S \tag{2.28}$$

Here, $\Delta\theta_{XY}$, $\Delta\theta_{XZ}$, etc., are the small misalignment angles. Initially, these misalignment angles are measured by calibration process and compensated in the navigation software. For ideal calibration and with proper compensation, two frames get aligned to each other. In reality, residual misalignments remain between the two frames, which cause error in navigation. Figure 2.10 explains the sensor misalignment error terms. If (X, Y, Z) co-ordinates define the orthogonal body frame in which a sensor input axis has a composite but small misalignment angle $\Delta\theta$ with respect to the reference X-axis, then this misalignment can be expressed as the

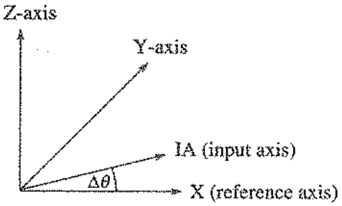


Figure 2.10 Sensor misalignment angles in body frame.

result of two independent rotations about the two orthogonal body axes. The first rotation is about Z-axis leading to misalignment angle $\Delta\theta_{XY}$ and then the second rotation is about Y-axis leading to misalignment angle $\Delta\theta_{XZ}$. So, two misalignment angles define the input axis (IA) error for one single axis sensor. Thus three sensors will have six misalignment angles as shown in Eq. (2.28).

2.7 Strapdown Navigation

Strapdown navigation uses inertial sensors, which are mounted on the vehicle body frame and hence the name. This means accelerometers are experiencing the rotation effect of the vehicle and hence they are no more in inertial frame. Functioning of this system is discussed further with respect to navigation in inertial reference frame.

When the gyros and the accelerometers are directly mounted on the vehicle body frame, the gyros measure inertial rotation of the vehicle in vehicle body frame, while the accelerometers measure the specific force vector in vehicle body frame. It is necessary that the specific force vector is transformed to the reference inertial frame. This is achieved by pre-multiplying the specific force vector $\mathbf{f}^{\mathrm{B}}$ in body frame with the body to inertial frame transformation matrix C_B^I, and the strapdown navigation equation becomes:

$$\frac{d^2\mathbf{R}^{\mathrm{I}}}{dt^2} = C_B^I\,\mathbf{f}^{\mathrm{B}} + \mathbf{g}^{\mathrm{I}}\,;\ \text{Since } \mathbf{f}^{\mathrm{I}} = C_B^I\,\mathbf{f}^{\mathrm{B}}$$

We write,

$$\frac{d^2\mathbf{R}^{\mathrm{I}}}{dt^2} = \mathbf{f}^{\mathrm{I}} + \mathbf{g}^{\mathrm{I}} \tag{2.29}$$

The transformation matrix is computed using the gyro output, which provides instantaneous inertial angular rate data sensed in body frame, which is then used to compute the rotation of the body frame with respect to the inertial frame. Once this incremental rotation angle is computed, the data is used to compute the transformation matrix C_B^I. The transformation of the specific force $\mathbf{f}^{\mathrm{B}}$ to $\mathbf{f}^{\mathrm{I}}$ makes the navigation formulation matching with Eq. (2.3) which, in the earlier days, was mechanised by gyro-driven gimbaled platform system.

2.7.1 Strapdown System Features

In Figure 2.11, an orthogonal body frame with axes (X, Y, Z) is defined on a sensor block where three numbers each of gyros and accelerometers are mounted and their input axes are properly aligned to this frame using calibration and software modelling.

In Figure 2.12, this sensor block is shown to be acted on by inertial acceleration and angular rate. The orthogonal accelerometer triad measures the body frame specific force components of $\mathbf{f}^{\mathrm{B}}\,(f_X, f_Y, f_Z)^B$, while the orthogonal triad of gyros provides body frame inertial angular rate components $(\omega_X, \omega_Y, \omega_Z)^B$ of angular rate vector $\mathbf{W}_{\mathbf{IB}}^{\mathbf{B}}$. The data is initially processed to remove the sensor errors by using the sensor model as appropriate to the sensor. The error coefficients are obtained from the sensor calibration. From the gyro data and using suitable scheme that is discussed in the subsequent sections later, the transformation matrix C_B^I is computed. Once the C_B^I matrix is available, the specific force components of $\mathbf{f}^{\mathrm{B}}$ are transformed to *I-frame*

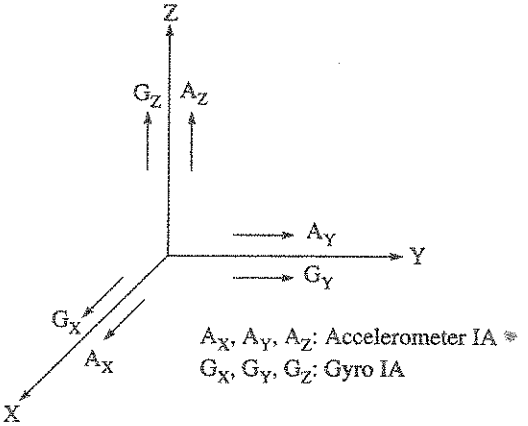

Z
G_Z
A_Z
A_Y
Y
G_Y
G_X
A_X
X
A_X, A_Y, A_Z: Accelerometer IA
G_X, G_Y, G_Z: Gyro IA

Figure 2.11 Gyros and accelerometers aligned to an orthogonal cluster frame.

components of $\mathbf{f}^{\mathbf{I}}$. The gravitation model is used to compute the gravitational acceleration components in *ECI-frame* and thereafter, transformed them to *I-frame*. This enables implementation of Eq. (2.3) and computation of the components of vector $\mathbf{A}^{\mathbf{I}}$. Thereafter, using suitable initial values and successive integrations, the inertial velocity and position are computed. From the elements of C_B^I matrix, it is possible to compute the Euler angle attitudes θ, ψ, ϕ that is described later. Thus the system outputs the position, the velocity and the attitude information for further use in flight navigation and guidance. Additional data provided to vehicle autopilot are body frame vehicle angular rate and acceleration. The data, as mentioned above, are computed and outputted at certain periodicity using onboard digital computer. The alignment block executes the essential preflight function of estimating the angular misalignment between the *B-frame* and the navigation frame. These misalignment angles, called initial attitudes, along with initial values of velocity and position are needed to initiate the navigation computation from the start of the flight time T_0.

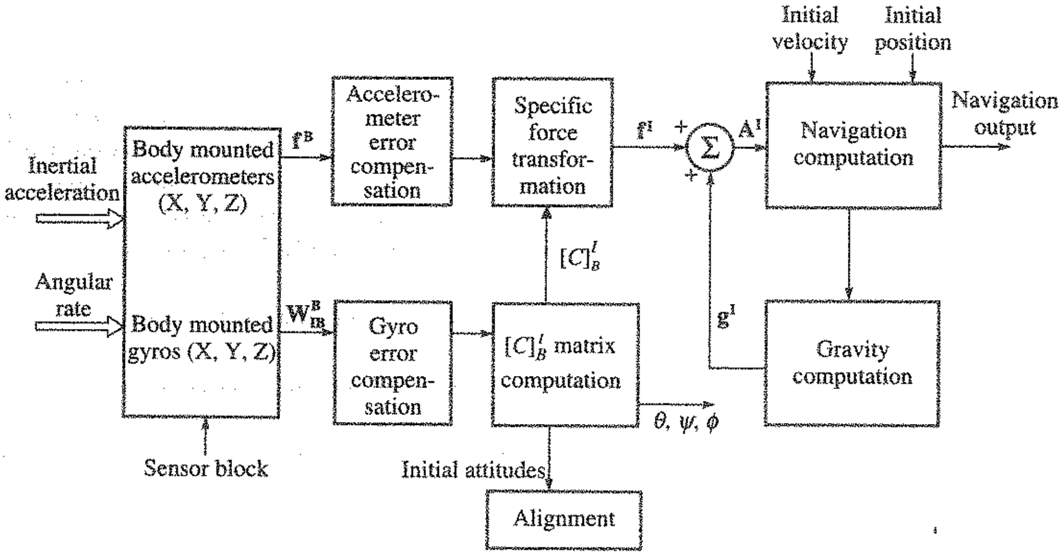

Initial velocity
Initial position
Navigation output
Inertial acceleration
Body mounted accelerometers (X, Y, Z)
$\mathbf{f}^{\mathbf{B}}$
Accelero-meter error compen-sation
Specific force transfor-mation
$\mathbf{f}^{\mathbf{I}}$
+
Σ
+
$\mathbf{A}^{\mathbf{I}}$
Navigation computation
$[C]_B^I$
$\mathbf{g}^{\mathbf{I}}$
Angular rate
Body mounted gyros (X, Y, Z)
$\mathbf{W}_{\mathbf{IB}}^{\mathbf{B}}$
Gyro error compen-sation
$[C]_B^I$ matrix computation
Gravity computation
θ, ψ, ϕ
Sensor block
Initial attitudes
Alignment

Figure 2.12 A strapdown navigation system schematic implementing inertial frame.

Question may arise on the requirement of minimum number of sensors for three-dimensional navigation. Three single axis sensors, consisting of three gyros and three accelerometers, are minimum needed and theoretically their sensitive axis must not lie in one plane. However, there is an optimal arrangement of the sensor layout with respect to the body fixed co-ordinates X, Y, Z. The optimal sensor layout, based on navigation performance, has been defined by a terminology called GDOP (discussed further in Section 2.10) and as per this criterion, three orthogonally mounted sensors, aligned to the body frame, provide the optimal sensor arrangement. As a result, the orthogonal sensor arrangement, depicted in Figure 2.11, is normally used as a primary non-redundant configuration.

2.7.2 Co-ordinate Transformation of Accelerometer Data

The method adopted, for co-ordinate transformation of accelerometer data from body frame to navigation frame, is vitally important in strapdown system. In the case of strapdown system, the accelerometer transformation matrix requires fast update with time, and as a first step, the transformation matrix itself is to be computed at equally high speed using gyro data. There are three evolved schemes for the computation of the transformation matrix using gyro data:

(a) Euler angle
(b) Direction cosine
(c) Quaternion

Principles of each of these schemes are discussed in the subsequent sections.

2.7.3 Euler Angle Transformation

The *Euler angle* representation of rotation is perhaps one of the simplest techniques in terms of physical appreciation. A transformation from one co-ordinate frame to another can be carried out as three successive rotations about certain sequence of axes. For the strapdown system, $X^BY^BZ^B$ represents one set of orthogonal frame that is called body frame, while $X^IY^IZ^I$ represents another set of orthogonal frame that is called reference inertial frame, which are shown in Figure 2.13(a). Measurement of specific force carried out in body frame can be transformed to inertial frame by using sequential rotations involving Euler angles. The sequential rotations are not unique and to some extent depend on application. The derivation of transformation matrix shown below assumes a typical rotation convention that is used in satellite launch vehicle. A transformation from reference axes (*I-frame*) to body *B-frame* can be executed through the following sequential positive rotations of magnitudes θ, ψ, ϕ, called Euler angles, assuming that the frames are initially aligned to each other.

Rotate through angle θ about reference Y-axis to obtain the orthogonal set $X_1Y_1Z_1$ [Figure 2.13(b)].

Rotate through angle ψ about the new X_1-axis to obtain the orthogonal set $X_2Y_2Z_2$ [Figure 2.13(c)].

Similarly, rotate through angle ϕ about the new Z_2-axis to obtain the desired orthogonal set $X^BY^BZ^B$.

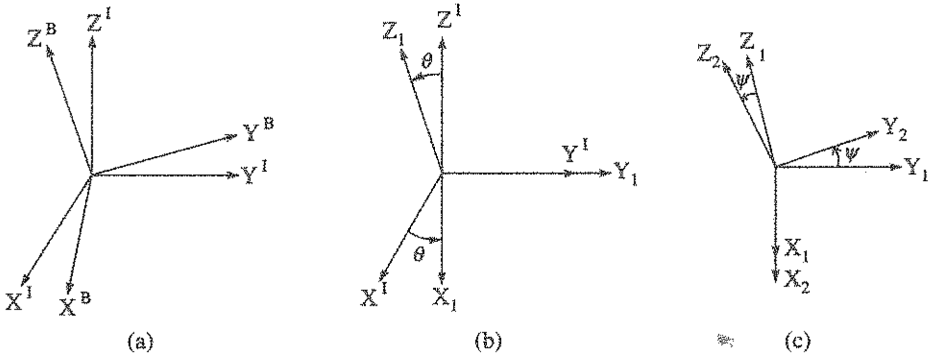


Figure 2.13 Representation of sequential rotation of two orthogonal frames.

where θ, ψ and ϕ are referred to as the Euler rotation angles. The three rotations may be expressed mathematically as three separate direction cosine matrices C_1, C_2 and C_3.

The rotation from reference inertial frame to the body frame can be expressed as the product of three transformations as:

$$C_I^B = [C_3 \times C_2 \times C_1] \tag{2.30}$$

Writing the transformation matrices, we get:

$$C_I^B = \begin{bmatrix} C\phi & S\phi & 0 \\ -S\phi & C\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & C\psi & S\psi \\ 0 & -S\psi & C\psi \end{bmatrix} \begin{bmatrix} C\theta & 0 & -S\theta \\ 0 & 1 & 0 \\ S\theta & 0 & C\theta \end{bmatrix} \tag{2.31}$$

where

C = cosine

S = sine

After performing the matrix multiplication, since, $C_B^I = [C_I^B]^T$, which is valid for orthogonal frame, Eq. (2.31) can be expressed as:

$$C_B^I = \begin{bmatrix} C\phi C\theta + S\phi S\psi S\theta & C\phi S\psi S\theta - S\phi C\theta & C\psi S\theta \\ S\phi C\psi & C\phi C\psi & -S\psi \\ S\phi S\psi C\theta - C\phi S\theta & S\phi S\theta + C\phi S\psi C\theta & C\psi C\theta \end{bmatrix} \tag{2.32}$$

Vehicle rotation sequence can be different, and as a result, the C_B^I matrix will be different from that shown in Eq. (2.32), but the derivation methodology is similar. For small angle rotation, the trigonometrical terms in Eq. (2.32) can be reduced suitably, which means (sin $\theta \approx \theta$ and cos $\theta \approx 1$).

Euler angle rate equation defines its propagation equation, which can be further used to derive the Euler angles. **Fernandez and Macomber, 1962** have shown that the Euler angle rates $(\dot{\phi}, \dot{\psi}, \dot{\theta})$ can be related to the body rates $(\omega_z, \omega_y, \omega_x)$ as:

$$\begin{bmatrix} \dot{\phi} \\ \dot{\psi} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} 1 & \cos\phi\tan\psi & \sin\phi\tan\psi \\ 0 & -\sin\phi & \cos\phi \\ 0 & \cos\phi\sec\psi & \sin\phi\sec\psi \end{bmatrix} \begin{bmatrix} \omega_z \\ \omega_y \\ \omega_x \end{bmatrix} \tag{2.33}$$

From Eq. (2.33), the Euler angles can be derived through integration as follows:

$$\phi = \int_0^t \dot{\phi}\, dt + \phi;\ \psi = \int_0^t \dot{\psi}\, dt + \psi_0; \theta = \int_0^t \dot{\theta}\, dt + \theta_0 \tag{2.34}$$

Suitable numerical integration scheme is used to solve Eq. (2.34). Once the Euler angles are computed, these are further used to compute C_B^I matrix as shown in Eq. (2.32). Observe singularity in Eq. (2.33) for ψ approaching 90° or 270°. This makes the scheme not at all suitable for all attitude vehicle manoeuvres. This is a big limitation on the use of this scheme for all attitude strapdown navigation, and as a result, this method of transformation is not currently used. However, representation of vehicle attitudes with Euler angles is still in vogue.

EXAMPLE 2.8 (on transformation): At a time $t = t_0$, two orthogonal frames, the body frame axes $X^B Y^B Z^B$ and the reference frame axes $X^I Y^I Z^I$ are aligned to each other as shown in Figure E2.1. Assume coincident origins. After a time, the body frame rotates by an angle θ with respect to reference frame about the common Y-axis.

(i) Sketch the rotated frames

(ii) If steady acceleration measured in the body frame axes is 20 m/s^2 at $t = 0$; find the acceleration in the reference frame for a rotation of 30°.

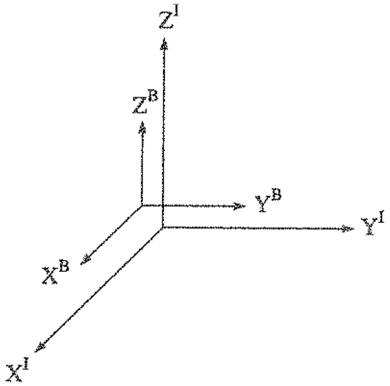


Figure E2.1 Initial mutual alignment of the two frames.

Solution: (i) Figure E2.2 shows the sketch of the two co-ordinate systems after a single rotation.

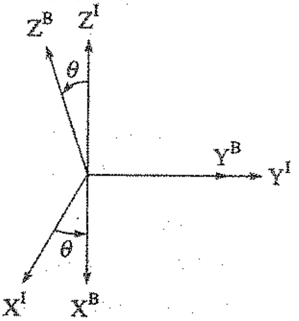


Figure E2.2 Rotated frame.

(ii) The transformation matrix after rotation is given by

$$C_I^B = \begin{bmatrix} C\theta & 0 & -S\theta \\ 0 & 1 & 0 \\ S\theta & 0 & C\theta \end{bmatrix}$$

For orthogonal frames, $C_B^I = [C_I^B]^T$, and the C_B^I matrix is given by:

$$C_B^I = \begin{bmatrix} C\theta & 0 & S\theta \\ 0 & 1 & 0 \\ -S\theta & 0 & C\theta \end{bmatrix}$$

We know that $[\mathbf{f}]^{\mathrm{I}} = C_B^I[\mathbf{f}]^{\mathrm{B}}$, so for $\theta = 30°$ and the acceleration vector ($[\mathbf{f}]^{\mathrm{B}}$) having components 20 m/s^2, the transformation matrix to covert to reference frame components will be given by

$$\begin{bmatrix} f_X^I \\ f_Y^I \\ f_Z^I \end{bmatrix} = \begin{bmatrix} \cos 30 & 0 & \sin 30 \\ 0 & 1 & 0 \\ -\sin 30 & 0 & \cos 30 \end{bmatrix} \begin{bmatrix} 20 \\ 20 \\ 20 \end{bmatrix}$$

Solving we get reference frame accelerations as: $f\,[XYZ]^{\mathrm{I}} = [\ 27.2, 20, 7.2]$ m/s^2

2.7.4 Direction Cosine Matrix Method

The direction cosine matrix (DCM) is a 3×3 transformation matrix. When the transformation of the vector is needed from body to inertial, the DCM is written in short form as C_B^I, the columns of which represent unit vectors in body axes (B-axes) projected along the reference axes (I-axes) and is written in component form as:

$$C_B^I = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \tag{2.35}$$

The element in the *i*th row and *j*th column represents the cosine of the angle between the i-axis of the reference frame and the j-axis of the body frame.

A vector $\mathbf{R}^{\mathbf{B}}$ defined in body axes, may be expressed in a reference frame (*I-frame*) as follows:

$$\mathbf{R}^{\mathbf{I}} = C_B^I \mathbf{R}^{\mathbf{B}} \tag{2.36}$$

As the C_B^I matrix is changing with time, its solution needs evaluation of the time derivative, also known as propagation of DCM. The propagation of the direction cosine matrix with time is obtained by differentiation, and this can be expressed as:

$$\dot{C}_B^I = \lim_{\delta t \to 0} \frac{C_B^I(t+\delta t) - C_B^I(t)}{\delta t} \tag{2.37}$$

where $C_B^I(t)$ represents the DCM at a time instant t, and $C_B^I(t + \delta t)$ represents the new DCM due to rotation about the body frame. Using matrix methods and small angle approximation, it has been shown [**Britting, 1971**] that Eq. (2.37) finally leads to:

$$\dot{C}_B^I = C_B^I \Omega_{\mathbf{IB}}^{\mathbf{B}} \tag{2.38}$$

where
$$\Omega_{\mathbf{IB}}^{\mathbf{B}} = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix} \tag{2.39}$$

$\Omega_{\mathbf{IB}}^{\mathbf{B}}$ is defined as a skew symmetric matrix. The skew symmetric matrix $\Omega_{\mathbf{IB}}^{\mathbf{B}}$ is evaluated from gyro measurements of the angular rate vector $\mathbf{W}_{\mathbf{IB}}^{\mathbf{B}} = [\omega_x, \omega_y, \omega_z]^T$ where $\mathbf{W}_{\mathbf{IB}}^{\mathbf{B}}$ is expressed as the *B-frame* angular rate relative to the reference frame *(I-frame)* co-ordinatised in body frame axes. Substituting Eqs. (2.35) and (2.39) in Eq. (2.38), we get the direction cosine rate equation, and from which the matrix elements can be derived using the integration equation shown below.

$$\dot{C}_B^I = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix}$$

$$C_{ij} = \int_0^t \dot{C}_{ij}\, dt + C_{ij(0)} \qquad \text{For } i, j = 1, 2, 3 \tag{2.40}$$

Equation (2.40) can be expressed in component form by matrix multiplication, and then solved in onboard computer. The DCM scheme does not have singularity and is suitable for all attitude manoeuvres. In general, the transformation from body to any navigation frame (*N-frame*) can be written as C_B^N.

2.7.5 Quaternion and Rotation Vector Method

In the concept of attitude quaternion and rotation vector, the attitude quaternion is a four-parameter representation such that any transformation from one co-ordinate frame to another may be effected by a single rotation of magnitude ϕ about a suitably chosen axis whose unit vector is $\mathbf{u}$. This is shown in Figure 2.14. The figure shows the reference frame (X^I, Y^I, Z^I) that has been rotated to a new frame (X^B, Y^B, Z^B) and $\mathbf{u} = [u_x, u_y, u_z]^T$ is the unit vector that defines the axis of rotation and ϕ is the rotation magnitude about the axis of $\mathbf{u}$.

The quaternion vector $\mathbf{q}$ is a four-parameter entity containing a scalar and a 3D vector part, which are functions of $\boldsymbol{\phi}$ and $\mathbf{u}$, and defined as:

$$\mathbf{q} = \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} \cos(\phi/2) \\ (u_x)\sin(\phi/2) \\ (u_y)\sin(\phi/2) \\ (u_z)\sin(\phi/2) \end{bmatrix} \tag{2.41}$$

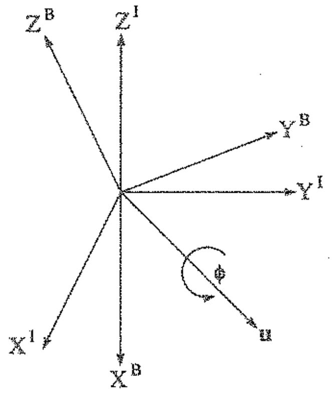


Figure 2.14 Quaternion and rotation vector concept.

The vector part $[q_1\ q_2\ q_3]^T$ still represents the axis of rotation (multiplying u_x, u_y and u_z with common factor sin ($\phi/2$) will not change the direction of the vector which is the axis of rotation). In principle, the scalar must contain the angle information directly or as a convenient function of the angle. Here cos ($\phi/2$) is chosen for convenience. Once cos ($\phi/2$) is chosen for the scalar part, automatically the common factor, sin ($\phi/2$) is used to multiply the vector part in particular for forcing the norm of the quaternion equal to 1.

The magnitude ϕ and direction (axis) **u** are defined in order that the reference frame ($X^I Y^I Z^I$) may be rotated into coincidence with the new frame ($X^B Y^B Z^B$) by the single rotation ϕ about axis **u**.

In the rotation vector concept, vector $\boldsymbol{\phi}$ can be defined as:

$$\boldsymbol{\phi} = \begin{bmatrix} u_x\phi \\ u_y\phi \\ u_z\phi \end{bmatrix} = \begin{bmatrix} \phi_x \\ \phi_y \\ \phi_z \end{bmatrix} \tag{2.42}$$

So, quaternion elements can also be expressed in terms of rotation vector $\boldsymbol{\phi}$ as:

$$\begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} \cos\left(\dfrac{|\phi|}{2}\right) \\ \dfrac{\phi_x}{|\phi|}\sin\left(\dfrac{|\phi|}{2}\right) \\ \dfrac{\phi_y}{|\phi|}\sin\left(\dfrac{|\phi|}{2}\right) \\ \dfrac{\phi_z}{|\phi|}\sin\left(\dfrac{|\phi|}{2}\right) \end{bmatrix} \tag{2.43}$$

where $|\phi| = \sqrt{\phi_x^2 + \phi_y^2 + \phi_z^2}$

The quaternion can also be viewed as an extension of the more usual two-parameter complex number form with one real component and three imaginary components. With this representation, the quaternion vector $\mathbf{q}$ is defined as hyper-complex number [**Salychev, 2004**] as:

$$\mathbf{q} = q_0 + \mathbf{i}q_1 + \mathbf{j}q_2 + \mathbf{k}q_3 \tag{2.44}$$

where (q_0, q_1, q_2, q_3) are real numbers and $(1, \mathbf{i}, \mathbf{j}, \mathbf{k})$ constitute the vector space.

Properties of quaternions

Quaternions exhibit some typical properties for executing mathematical operations such as additions, multiplications and inverse operation.

The multiplication of two quaternions ($\mathbf{q} = q_0 + \mathbf{i}q_1 + \mathbf{j}q_2 + \mathbf{k}q_3$) and ($\mathbf{p} = p_0 + \mathbf{i}p_1 + \mathbf{j}p_2 + \mathbf{k}p_3$) are defined by the following rules on complex numbers:

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = -1$$

where $\mathbf{i} = \mathbf{j} = \mathbf{k} = \sqrt{-1}$

$$\mathbf{ij} = \mathbf{k} = -\mathbf{ji}; \quad \mathbf{jk} = \mathbf{i} = -\mathbf{ki}; \quad \mathbf{ki} = \mathbf{j} = -\mathbf{ik}$$

where the order of unit vector is preserved.

It may be noted that the quaternion multiplication is not commutative. The product can be expressed using the vector dot and cross products as:

$$\mathbf{q} * \mathbf{p} = \mathbf{q}_0\mathbf{p} + \mathbf{p}_0\mathbf{q} - \mathbf{q} \cdot \mathbf{p} + \mathbf{q} \times \mathbf{p} \tag{2.45}$$

Alternatively, the quaternion product may be expressed in matrix form as:

$$[\mathbf{q} * \mathbf{p}] = \begin{bmatrix} q_0 & -q_1 & -q_2 & -q_3 \\ q_1 & q_0 & -q_3 & q_2 \\ q_2 & q_3 & q_0 & -q_1 \\ q_3 & -q_2 & q_1 & q_0 \end{bmatrix} \begin{bmatrix} p_0 \\ p_1 \\ p_2 \\ p_3 \end{bmatrix} \tag{2.46}$$

The addition and scalar multiplication are defined component wise as for usual four-dimensional vectors. Quaternion propagation with time, also called quaternion rate equation, is given by

$$\dot{\mathbf{q}} = \frac{1}{2}\mathbf{q} * \boldsymbol{\omega} \tag{2.47}$$

where $\boldsymbol{\omega} = [0, \omega_x, \omega_y, \omega_z]^T$.

With the quaternion multiplication shown in Eq. (2.46), the quaternion rate equation can be expressed in component form as:

$$\begin{bmatrix} \dot{q}_0 \\ \dot{q}_1 \\ \dot{q}_2 \\ \dot{q}_3 \end{bmatrix} = \left[\frac{1}{2}\right] \begin{bmatrix} q_0 & -q_1 & -q_2 & -q_3 \\ q_1 & q_0 & -q_3 & q_2 \\ q_2 & q_3 & q_0 & -q_1 \\ q_3 & -q_2 & q_1 & q_0 \end{bmatrix} \begin{bmatrix} 0 \\ \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} \tag{2.48}$$

Using suitable quaternion rate integration scheme, the four elements of the quaternion can be derived from the gyro data. Once the quaternion elements are obtained, the transformation matrix C_B^I can be expressed as:

$$C_B^I = \begin{bmatrix} 1-2(q_2^2+q_3^2) & 2(q_1q_2-q_0q_3) & 2(q_1q_3+q_0q_3) \\ 2(q_1q_2+q_0q_3) & 1-2(q_1^2+q_3^2) & 2(q_2q_3-q_0q_1) \\ 2(q_1q_3-q_0q_2) & 2(q_2q_3+q_0q_1) & 1-2(q_1^2+q_2^2) \end{bmatrix} \tag{2.49}$$

It is to be noted that $(q_0^2+q_1^2+q_2^2+q_3^2)^{1/2}=1$ is defined as the *norm* of quaternion, and this is used for quaternion normalisation. Also, quaternion is free from singularity.

Attitude, if required to be expressed in Euler angles, can also be computed by noting the equivalence of the transformation matrix elements given by Eqs. (2.32) and (2.49).

For example, attitude angle ϕ can be computed using the relation:

$$\phi = \tan^{-1}\frac{C_{21}}{C_{22}} = \tan^{-1}\frac{2(q_1q_2+q_0q_3)}{1-2(q_1^2+q_3^2)} \tag{2.50}$$

2.7.6 Comparison between the Three Transformation Schemes

Parameters: Euler angle scheme has three parameters only as against nine parameters in direction cosine and four in quaternion.

Singularity: Euler angle scheme suffers from singularity at 90° and 270° for the middle angle (ψ), making all attitude navigation not possible. Other two schemes are free from this singularity.

Rate equations: In Euler angle scheme, the rate equations are cumbersome due to handling of trigonometric functions. In other two schemes, it is not so but require periodic correction for orthogonality and normalisation.

Transformation matrix: Both in the Euler angle scheme as well as in the quaternion scheme, it is necessary to compute the direction cosine matrix. It is not necessary in direction cosine scheme.

Attitude computation: When attitude information is needed in Euler angle terms, no separate Euler angle transformation is needed in the Euler angle scheme. Other two schemes need separate Euler angle transformation.

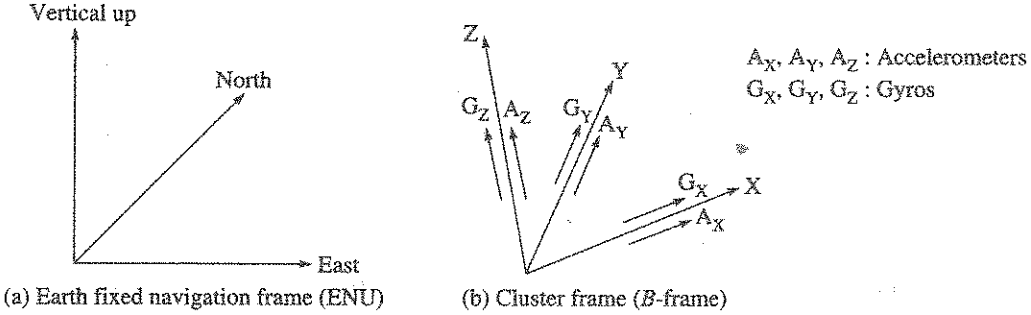
2.7.7 Principle of Self Alignment

In the navigation equation described earlier in the text, it is implied that the orientation of the orthogonal body frame to which the inertial sensors are aligned, is accurately known with respect to the reference navigation frame prior to initiation of navigation. Usually, in the service condition of an INS, the relative orientation between the two frames is unknown and the process of measurement of this unknown parameter is called initial alignment. Self alignment is a process that uses inertial sensors, meant for navigation, to determine the unknown orientation. In a strapdown system, the scheme is called analytic, as the process involves determination of initial attitudes, either in quaternion or in direction cosine, to enable computation of the initial transformation matrix between the body frame and the reference frame at the start of navigation.

Self alignment normally uses an earth fixed reference frame, such as, east, north and vertical up (acronym ENU). Even where the navigation frame is non geographic such as ECI or

LPI, it is convenient to choose for self alignment an earth fixed frame as it permits realisation of an autonomous alignment process for navigation initiated from the earth fixed location.

We can define the two orthogonal frames, the earth fixed navigation frame (ENU) called *G-frame* [refer Figure 2.15(a)], and the cluster frame or the *B-frame* [refer Figure 2.15(b)]. It is required that the orientation between these frames is to be determined by self alignment process.


Vertical up
North
East
Z
Y
X
G_Z
A_Z
G_Y
A_Y
G_X
A_X
A_X, A_Y, A_Z : Accelerometers
G_X, G_Y, G_Z : Gyros

(a) Earth fixed navigation frame (ENU) (b) Cluster frame (*B*-frame)

Figure 2.15 Principle of self alignment.

In the self alignment process, we choose two non-collinear and independent vectors that are known as local gravity vector $\mathbf{g_p}$ and the earth angular rate vector $\boldsymbol{\omega}_e$ whose components can be resolved in both the frames. Just recall from our earlier discussion that in the earth fixed frame, $\mathbf{g_p}$ is along the vertical axis and $\boldsymbol{\omega}_e$ is along north axis.

The resolved components of these vectors in the ENU frame, defined as *G-frame*, are as follows:

$$\mathbf{g}_\mathbf{p}^G = \begin{bmatrix} 0 & 0 & -g_p \end{bmatrix}^T ; \; \boldsymbol{\omega}_e^G = \begin{bmatrix} 0 & \omega_e \cos L & \omega_e \sin L \end{bmatrix}^T$$

Thus the components of these two vectors are well defined in the earth fixed frame and only latitude L, where the self alignment is to take place, should be known. But in the *B-frame*, the components of $\mathbf{g_p}$ and $\boldsymbol{\omega}_e$ are measured by the inertial sensors where the measurements will depend on the orientation of the *B-frame* axes (X, Y, Z) relative to $\mathbf{g_p}$ and $\boldsymbol{\omega}_e$. In this case, for simplicity, it is assumed that the input axis of the gyros and the accelerometers are truly aligned, no residual errors, with the corresponding *B-frame* axes XYZ, which is shown in Figure 2.15(b).

However, we need to have a third vector besides the two known vectors. Using the local gravity vector and the earth angular rate vector, a third vector $\mathbf{p}$ can be created by their cross product relation, which is $\mathbf{p} = \mathbf{g_p} \times \boldsymbol{\omega}_e$. Only thing to be noted that $\mathbf{p}$, thus generated, is not an independent vector.

The components of the cross product vector $\mathbf{p}^G$ in *G-frame* can be written as:

$$\mathbf{p}^G = \begin{bmatrix} g_p \omega_e \cos L & 0 & 0 \end{bmatrix}^T$$

Now, the vectors in the body frame $[\mathbf{g_p}, \boldsymbol{\omega}_e, \mathbf{p}]^B$ and the corresponding earth reference frame vectors $[\mathbf{g_p}, \boldsymbol{\omega}_e, \mathbf{p}]^G$ can be related using a co-ordinate transformation matrix. This transformation matrix C_G^B between the two frames can be written as follows:

$$\begin{aligned} \mathbf{g}_p^B &= C_G^B \mathbf{g}_p^G \\ \boldsymbol{\omega}_e^B &= C_G^B \boldsymbol{\omega}_e^G \\ \mathbf{p}^B &= C_G^B \mathbf{p}^G \end{aligned} \tag{2.51}$$

Using these derived relations, the vectors $\mathbf{g}_p^G$, $\boldsymbol{\omega}_e^G$, and $\mathbf{p}^G$ can be expressed in their component form as:

$$\begin{bmatrix} (\mathbf{g}_p^G)^T \\ (\boldsymbol{\omega}_e^G)^T \\ (\mathbf{p}^G)^T \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 & -g_p \\ 0 & \omega_e \cos L & \omega_e \sin L \\ g_p \omega_e \cos L & 0 & 0 \end{bmatrix}^{-1} \tag{2.52}$$

Similarly, $\mathbf{g}_p^B$ and $\boldsymbol{\omega}_e^B$ can be measured along the body axes X, Y and Z as:

$$\mathbf{g}_p^B = [g_X \quad g_Y \quad g_Z]^T; \quad \boldsymbol{\omega}_e^B = [\omega_{e(X)} \quad \omega_{e(Y)} \quad \omega_{e(Z)}]^T$$

From these measurements, components of $\mathbf{p}^B$ are obtained.

Using these information, it will be possible to compute the nine elements of the transformation matrix C_G^B and also its inverse that provides for the elements of C_B^G. This method of alignment is called *strapdown self alignment principle*. The method works so long the matrix inverse operation to compute C_B^G exists and the inverse computation exists so long the independent vectors $\mathbf{g}_p$ and $\boldsymbol{\omega}_e$are not collinear. Only, at the poles, this non-collinearity condition is violated and self alignment scheme is not possible.

At the end of the analytic self alignment process and using the elements of the C_B^G matrix, three computed misalignment angles in the form of Euler angles ε_N, ε_E and ε_V will be available where the first two are normally termed as the *level misalignment angles* and the third one is termed as *azimuth misalignment angle*. When the desired navigation frame is different from the *G-frame*, then these angles can be transformed to the needed frame. The method, just described, is suitable when the navigation system is stationary with respect to the earth fixed frame.

Error in alignment

The inertial sensors are not error free and the sensor errors lead to error in alignment normally termed as the *level errors* and the *azimuth error*. In the presence of these errors, the computed misalignment angles will also be in error. If we consider the simplest type of error, (Δf_x, Δf_Y, $\Delta\omega_x$), called bias error in the sensors, the magnitude of the alignment errors can be expressed for small errors as:

$$\begin{aligned} \Delta f_x + g_P \Delta\varepsilon_N &= 0 \\ \Delta f_y - g_P \Delta\varepsilon_E &= 0 \\ \Delta\omega_x - \omega_e \cos L \Delta\varepsilon_V &= 0 \end{aligned} \tag{2.53}$$

where

Δf_x, Δf_y = bias error in the accelerometers in X-axis and Y-axis respectively

$\Delta\omega_x$ = uncompensated drift in X-axis gyro (nominally east gyro)

$\Delta\varepsilon_N$, $\Delta\varepsilon_E$, $\Delta\varepsilon_V$ = error in alignment about north, east and vertical respectively

The first two equations define *level errors*, and the third equation defines the *azimuth error*. Now, azimuth error due to bias gyro drift $\Delta\omega_x$ depends on the drift characteristic. Gyro exhibiting random walk (refer Appendix D) type of noise, has error that depends on the square root of alignment time and is expressed as:

$$\varepsilon_V = \frac{R_w}{\omega_e \cos L\sqrt{t}} \tag{2.54}$$

where

ε_V = standard deviation of error in azimuth alignment
R_w = gyro angular random walk in $°/\sqrt{h}$
t = alignment time

Equation (2.54) conveys that if the allowed alignment time is short, then the azimuth alignment accuracy will be dominated by R_w.

EXAMPLE 2.9 (on alignment error): Suppose the two-level accelerometers have bias error of 0.1 mg and the east gyro has uncompensated drift of 0.01°/h and the INS is operating at latitude of 30°. What is the level error and the azimuth error?

Solution: We have Δf_x, $\Delta f_y = 0.1$ mg, $\Delta\omega_x = 0.01°/\text{h}$, $g_p = 1$

Level error magnitude $\Delta\varepsilon_N$, $\Delta\varepsilon_V = (0.1 \times 10^{-3}/1)$ rad $= 0.1 \times 10^{-3}$ rad = 20 arc-second

Azimuth error $\Delta\varepsilon_V = 0.01/(15.04 \times \cos 30) = 0.00076$ rad = 158.6 arc-second.

For gyro exhibiting random walk type of noise of magnitude $0.002°/\sqrt{h}$, if the requirement is to limit the error to 1 m-rad at a latitude of 45°, then the estimated alignment time needed would be (from Eq. 2.54).

$$1 \times 10^{-3} = \frac{0.002}{15.04 \times \cos 45 \times \sqrt{t}}$$

Solving for t, we get $t = 0.0353$ h = 127 s

The example shows that precision azimuth alignment determination with self alignment scheme, needs very accurate gyro.

For situation where the INS experiences base level vibration due to wind and gust acting on the vehicle as well as noise in the sensor output, optimal filtering techniques have been used to improve upon the accuracy. In such scheme, the analytic scheme provides for the coarse estimate of the C_B^G matrix, followed by a refined estimate using Kalman filter.

There are applications when the vehicle will be in relative motion, where different schemes need to be applied other than self alignment. Velocity matching and transfer alignment from another INS are some typical schemes which are under use. In certain application, where the spacecraft in orbit is to be brought back to earth, sighting of stars is used to derive the misalignment angles prior to initiating the return manouevre operations.

2.7.8 Optical Azimuth Alignment

As very accurate gyros are needed to meet high accuracy azimuth alignment using self alignment scheme, often optical azimuth alignment has been used in missions that need high accuracy and where optical alignment is feasible to implement. A scheme is described below which is feasible in a satellite launch vehicle.

Azimuth is normally defined as the clockwise rotation from the true north. In this definition, an axis pointing along the east direction has an azimuth of 90°. In LPI reference frame navigation, the reference frame axis X^I (refer Figure 2.7), is chosen to point towards the desired azimuth direction, and the angle is specified by the mission requirement.

We need to find out what is the angular relation between the INS body frame axis X^B and X^I axis of the reference frame. It should be noted that both the frames are orthogonal.

In the principle of operation of optical alignment, refer Figure 2.16, the orientation of the body frame X^B-axis of the strapdown system from the true north is optically measured using auto collimating laser theodolite.

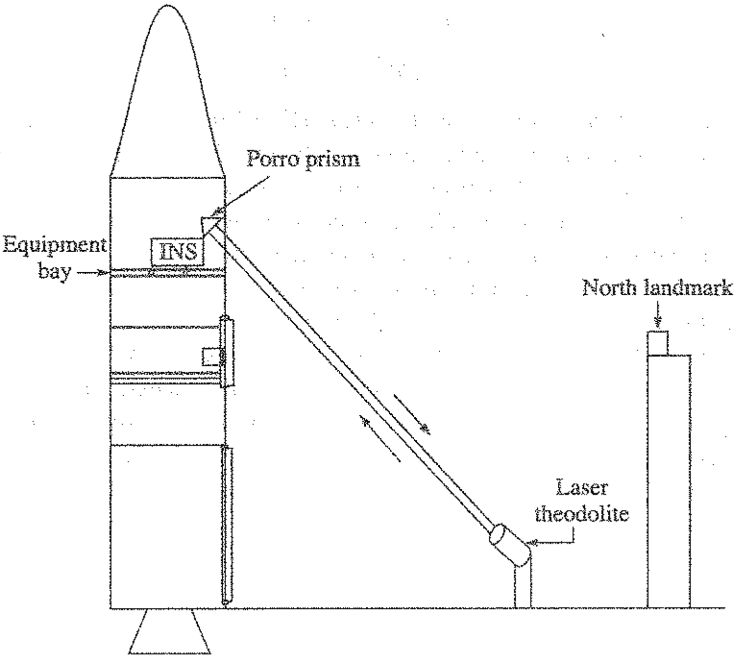


Figure 2.16 Optical azimuth determination at launch pad.

To facilitate the measurement of X^B-axis, a porro prism is mounted on the X^B-axis of the system. The porro prism provides an advantage where the incident and the reflected laser beams are parallel to each other thereby permitting easier sighting. A true north landmark is separately established at a nearby location of the laser theodolite benchmark. Sighting the porro prism with the laser theodolite from a bench mark location and then sighting the north landmark from the same theodolite bench mark location, the azimuth of the X^B-axis can be determined. The method also has some errors, such as tilt of prism roof edge if any and that should be calibrated and compensated for azimuth error. However, the method consumes considerable time and the preferable sighting time is after sunset and when wind disturbance is low. Level alignment continues with the accelerometers in self alignment mode.

2.8 Error Propagation in Strapdown INS

The output of an INS is erroneous primarily due to sensor errors. These sensor errors are normally categorised as systematic and random. While the systematic part is modelled and normally compensated, the random error part leads to INS error. The other source of error is in modelling the gravity, although in general, it is of lesser magnitude. INS error is not steady, but propagates with time due to successive integrations involved in computing velocity and position.

Working from the first principle, it is possible to derive analytically the error propagation in navigation outputs defined in terms of position, velocity and attitude. Such a derivation is described in the following paragraphs for a strapdown system operating in inertial reference frame.

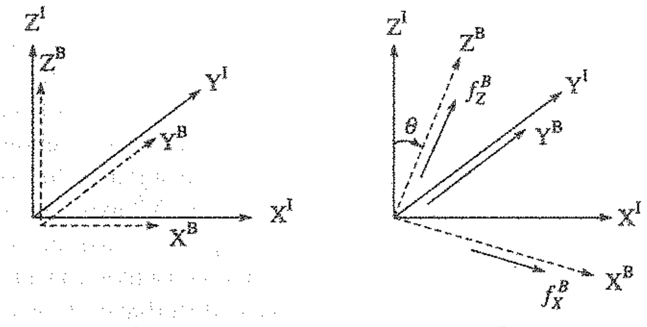
Equation (2.29) describes the strapdown navigation where, if the error in gravitation is neglected, the vector form of error equation can be written as follows:

$$\delta \mathbf{f}^{I} = C_B^I \delta \mathbf{f}^{B} + \delta C_B^I \mathbf{f}^{B} \tag{2.55}$$

Equation (2.55) shows that the error in specific force $\delta \mathbf{f}^I$ in *I-frame* results from the combination of error $\delta \mathbf{f}^B$ in the accelerometer output along with transformation matrix error δC_B^I. The error in C_B^I is caused due to combination of error in initial alignment and gyro drift. These errors alone on further integrations result in velocity as well as position error propagation. The transformation matrix C_B^I couples accelerometer error of one axis onto the other axis. This aspect of coupling and error propagation can be explained further for a simple case of strapdown navigation where there is single rotation of θ about Y-axis.

Initially, at the start of navigation, the body frame axes X^B, Y^B, Z^B is aligned with the corresponding inertial frame axes X^I, Y^I, Z^I using alignment process, and such aligned frames are shown in Figure 2.17(a). Thereafter, during flight, the body frame rotates relative to the inertial frame. Figure 2.17(b) shows the single rotation of the body fixed co-ordinates about the originally aligned Y-axis for an angle θ. The specific force sensed in body frame can then be resolved in inertial frame. The resolved equations of the specific force in body frame to the inertial frame are given as:

$$\begin{aligned} f_X^I &= f_X^B \cos\theta + f_Z^B \sin\theta \\ f_Z^I &= -f_X^B \sin\theta + f_Z^B \cos\theta \end{aligned} \tag{2.56}$$



(a) (b)

Figure 2.17 Rotation of body fixed frame relative to the inertial frame.

In matrix form, Eq. (2.56) can be rewritten as:

$$\begin{bmatrix} f_X^I \\ f_Z^I \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} f_X^B \\ f_Z^B \end{bmatrix} \tag{2.57}$$

For a single rotation, the transformation matrix C_B^I is thus a 2 × 2 matrix.

Equation (2.55) can be redefined by considering errors which will cause errors in the sensed specific force and in the rotation angle θ. For example, accelerometer bias causes error in specific force, while gyro bias drift causes error in C_B^I. Thus combining Eqs. (2.55) and (2.56), and simplifying, the error equations can be written as:

$$\begin{aligned} \Delta f_X^I &= f_Z^I \Delta\theta + B_X^B \cos\theta + B_Z^B \sin\theta \\ \Delta f_Z^I &= -f_X^I \Delta\theta - B_X^B \sin\theta + B_Z^B \cos\theta \end{aligned} \tag{2.58}$$

where

$\Delta(f_x, f_z)^I$ = error in specific force components in inertial frame

$(f_x, f_z)^I$ = specific force components resolved along X and Z axes respectively in inertial frame

θ = rotation about Y-axis

$\Delta\theta$ = error in measuring vehicle rotation, which is contributed due to Y-gyro drift

$(B_X, B_Z)^B$ = accelerometer measurement errors, such as bias, in body frame

$\Delta\omega_{IB}^B$ = Y-gyro drift in body frame

The propagated error in position, assuming one error at a time, is tabulated in Table 2.2 [**Titterton and Weston, 2004**]

Table 2.2 Navigation Error Sensitivity in Inertial Reference Frame

Error sources		Position error ΔX	Position error ΔZ
Accelerometer bias error,	B_X^B	$B_X^B \cos\,\theta t^2/2$	$B_X^B \sin\theta t^2/2$
	B_Z^B	$B_Z^B \sin\,\theta t^2/2$	$B_Z^B \cos\,\theta t^2/2$
Y gyro drift,	$\Delta\omega_{IB}^B$	$\Delta\omega_{IB}^B f_Z^I t^3/6$	$-\Delta\omega_{IB}^B f_X^I t^3/6$

Table 2.2 shows that the navigation error in position due to accelerometer bias error propagates as t^2, whereas navigation position error due to the gyro drift error propagates as t^3. This happens as gyro senses angular rate, and additional integration is necessary to derive the rotation angle for transformation matrix computation. Such a propagation characteristic in navigation error means that the listed error sources must be controlled to their low values which will be commensurate with the specified navigation error at the end of the flight time. If the individual error sources are considered independent and random, one method of computing the total navigation error is to use root sum square of the individual errors. However, the situation gets complicated if the error sources have acceleration sensitivity and/or rotation sensitivity.

A realistic error propagation estimate for a three-axes INS with sensor error model, is an involved exercise and normally carried out with techniques such as Monte Carlo and statistical covariance analysis (**Bose et al., 2008**). In an INS, the deterministic part of the error is compensated (refer Figure 2.12), which allows only the uncorrected part, normally considered random, to contribute to navigation errors. This permits use of statistical tools for navigation error estimation.

EXAMPLE 2.10 (on error propagation): An INS is mounted on a vehicle in a manner as shown in Figure 2.17. 100 s after take off, the vehicle turns quickly about Y-axis by 10°. If all the accelerometers have bias error of 200 × 10^{-6} g, calculate the position error along navigation axes for the two cases. Assume $g = 9.8$ m/s^2.

(i) During take off to 100 s

(ii) 100 s to 1000 s of flight duration

Solution: We can use the derivation shown in Table 2.2 to calculate the position errors.

(i) 0 to 100 s

The position error along the X, Y, Z axes (navigation) = $200 \times 10^{-6} \times 9.8 \times 100^2/2 =$ 9.8 m

(ii) 100 to 1000 s

The rotation of vehicle is at 100th second of the flight. So, we can use Table 2.2 to compute the errors.

Error ΔX_1 due to bias error

$$B_X^B = \frac{200 \times 10^{-6} \times 9.8 \times \cos 10 \times 900^2}{2} = 777.9 \text{ m}$$

Error ΔX_2 due to bias error

$$B_Z^B = \frac{200 \times 10^{-6} \times 9.8 \times \sin 10 \times 900^2}{2} = 134.9 \text{ m}$$

So,

$$\Delta X = [(\Delta X_1)^2 + (\Delta X_2)^2]^{1/2} = 789.5 \text{ m}$$

Similarly, ΔZ can be calculated and due to identical formulation, $\Delta Z = \Delta X = 789.5$ m

$$\Delta Y = \frac{200 \times 10^{-6} \times 9.8 \times 900^2}{2} = 793.7 \text{ m}$$

2.9 Strapdown System Technology

2.9.1 System Configuration and Functional Description

A strapdown navigation system configuration depends on several interlinked factors like type of sensors used, choice of navigation co-ordinate frame, algorithms, performance, interfaces and environmental specification. Figure 2.18 shows the functional block schematic of a basic strapdown system which is described further.

Three single axis inertial sensors are mounted on an accurately machined orthogonal cluster that is normally defined as the body frame. The cluster may have vibration isolation system to protect the sensors from environmental stresses and also an active temperature control for the sensor cluster. Currently, in lieu of the temperature control, sensor thermal model is used to compensate for the temperature related shift.

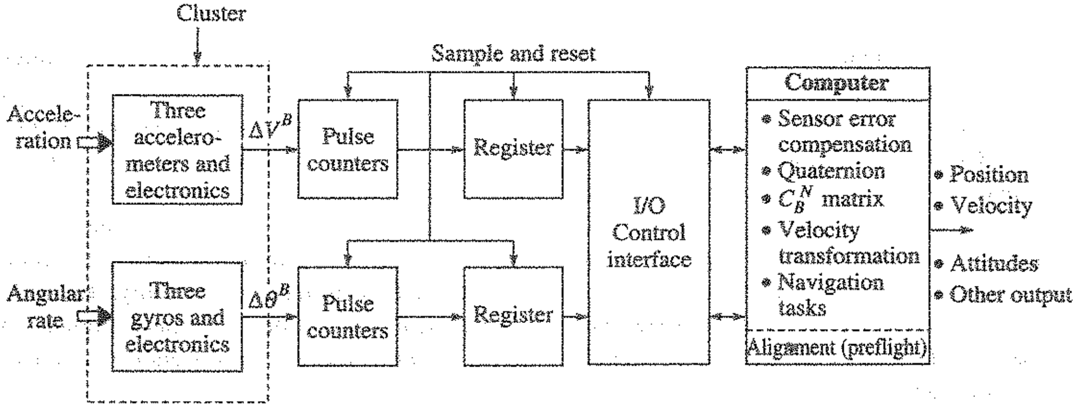


Figure 2.18 Typical strapdown system configuration.

The accelerometer output is in the form of velocity increment in body frame (ΔV^B) as it saves one integration in the computer. Same is true for gyro also, where the output is in the form of incremental angle in body frame ($\Delta\theta^B$). This also means that the subsequent software is formulated taking note of this feature of sensor output. The sensor output pulses are counted in up-down counters and stored in the registers. Under the control of an input-output (I/O) controller, increment of sensed velocity (integral of specific force) ΔV and increment of sensed rotation $\Delta\theta$ are accessed in a given sample time. The data is then used by the navigation computer to execute the listed tasks.

The software task structure is of particular importance in a strapdown system as it operates at two or more update frequencies. This is shown in Figure 2.19.

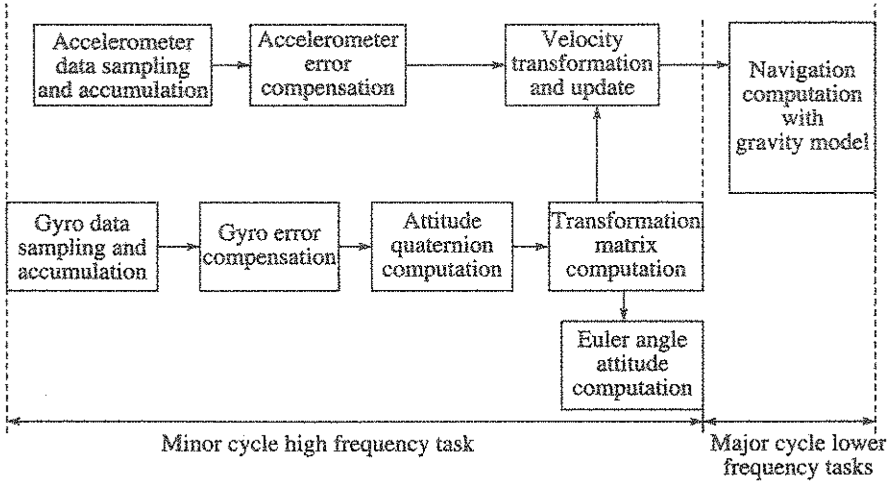

Accelerometer data sampling and accumulation
Accelerometer error compensation
Velocity transformation and update
Navigation computation with gravity model
Gyro data sampling and accumulation
Gyro error compensation
Attitude quaternion computation
Transformation matrix computation
Euler angle attitude computation
Minor cycle high frequency task
Major cycle lower frequency tasks

Figure 2.19 Strapdown software with multifrequency tasks.

The first major task is to transform the body frame ΔV^B to the desired reference inertial frame. This task, or rather a set of associated tasks, is performed at a high frequency which is sometimes called minor cycle update rate. Thereafter, the computed incremental velocity in inertial frame, ΔV^I, is further used to compute the navigation parameters, which are often done at a lower frequency called the major cycle update rate. Initially, the navigation parameter lower frequency update rate is decided based on application. Overall, the decision is linked to the bandwidth of the aerospace vehicle guidance and steering loop where navigation output is directly used. Navigation update frequency between 2 Hz and 20 Hz is seen to cover large number of aerospace applications. Normally, major cycle frequency to minor cycle frequency ratio is an integral multiple number k. This means there are k minor cycles in one major cycle. Once the navigation update frequency is fixed by the application requirement, factor k is decided and that goes as one of the inputs to fix up the minor cycle task frequency. Typically, this update frequency is around 50 Hz to 100 Hz and in some critical application, it is much higher. One factor that decides is the navigation computer speed. In some specific sensor choice, a separate processor handles the high speed data.

The computed attitude and the body rate are required by the vehicle control loop and normally executed as a minor cycle task. Self alignment task is completed before $t = t_0$, which provides the initial value of the transformation matrix.

2.9.2 Navigation Equation with Update

The analytical formulations imply use of continuous integration process. But the implementation scheme of the analytical formulations in a computer is discrete in nature and is normally executed with an update formulation, which is described further.

The velocity transformed to the inertial navigation frame in the interval $t_0 - t$ is given by the analytic relation in the continuous time domain as:

$$\mathbf{V}^{\mathbf{I}}(t) = \int_{t_0}^{t} C_B^I(t)\mathbf{f}^{\mathbf{B}}(t)\,dt \tag{2.59}$$

The approximate solution of Eq. (2.59), when expressed in iterative form, can be written as:

$$\mathbf{V}_{\mathbf{n}}^{\mathbf{I}} = \mathbf{V}_{\mathbf{n}-1}^{\mathbf{I}} + C_B^I(t_0 + n\,\delta t) \int_{t_0+(n-1)\delta t}^{t_0+n\delta t} \mathbf{f}^{\mathbf{B}}(t)\,dt \tag{2.60}$$

On further reduction, the equation becomes:

$$\mathbf{V}_{\mathbf{n}}^{\mathbf{I}} = \mathbf{V}_{\mathbf{n}-1}^{\mathbf{I}} + C_{Bn}^I \delta\mathbf{V}_{\mathbf{n}}^{\mathbf{B}} \tag{2.61}$$

where

$\delta\mathbf{V}_{\mathbf{n}}^{\mathbf{B}}$ = velocity increment vector measured by accelerometers in body axes during the nth update cycle

C_{Bn}^I = nth update cycle transformation matrix

$\mathbf{V}_{\mathbf{n}}^{\mathbf{I}}$, $\mathbf{V}_{\mathbf{n}-1}^{\mathbf{I}}$ = transformed sensed velocity vectors in inertial frame at nth cycle and (n-1)th cycle respectively

Equation (2.61) is computed at the high frequency minor cycle update rate for k number of cycles when the data is used for major cycle navigation computation. The major cycle navigation equation computes for inertial velocity and position and in iterative form, the equations can be expressed as:

$$\mathbf{V}_n^I = \mathbf{V}_{n-1}^I + \Delta\mathbf{V}_n^I + \Delta\mathbf{V}_{gn}^I \tag{2.62}$$

$$\mathbf{R}_n^I = \mathbf{R}_{n-1}^I + (\mathbf{V}_{n-1}^I + \mathbf{V}_n^I)\frac{\Delta t}{2} \tag{2.63}$$

where

$\mathbf{V}_n^I$ = inertial velocity after n major cycles

$\mathbf{V}_{n-1}^I$ = inertial velocity after $n-1$ major cycles

$\Delta\mathbf{V}_n^I$ = incremental component of velocity, from minor cycle task, in the nth cycle

$\Delta\mathbf{V}_{gn}^I$ = incremental component of velocity due to gravitational acceleration in the nth cycle

$\mathbf{R}_n^I$ = inertial position after n major cycles

$\mathbf{R}_{n-1}^I$ = inertial position after $n-1$ major cycles

Δt = integration step size

Equation (2.63) shows a trapezoidal form of integration. Depending on accuracy requirement, higher order integration can be used.

2.9.3 Strapdown Inertial Navigation Functional Summary

Functioning of strapdown navigation system has been discussed at length. It can be summarised as follows:

(a) Choose an appropriate reference frame for navigation.

(b) Minimum three single axis gyros and three single axis accelerometers are needed in orthogonal configuration.

(c) Accelerometer measurements, termed as specific force, are in body frame and needed to be transformed to reference frame.

(d) Gyros provide the information to compute the transformation matrix that will enable the transformation from body frame to reference frame.

(e) Various mathematical schemes are available for generating transformation matrix where quaternion method is currently a preferred option.

(f) Accelerometer measurements are transformed to reference frame when transformation matrix is realised with the data from gyros.

(g) These transformation matrices are required to be updated at high update rate.

(h) Vehicle attitudes are computed from the direction cosine elements or the quaternion elements of the transformation matrix.

(i) Navigation error grows with time. Since gyro and accelerometer errors contribute to error in navigation, it becomes absolutely necessary to reduce the sensor errors through appropriate model and through compensation in software.

(j) Gyro and accelerometer are also used for implementation of self alignment. In launch vehicle application, optical method of azimuth angle determination can be used to meet high accuracy.

(k) Appropriate gravitation model is used to maintain accuracy in navigation.

2.9.4 Specific Strapdown Errors

There are certain errors seen in a strapdown system due to vehicle dynamics and typical of such strapdown errors are as follows:

(a) Coning motion induced error
(b) Sculling error
(c) Size effect error

Coning motion induced error

Coning is characterised by certain typical rotation of the angular rotation vector where the axis of rotation is itself moving in space. In this motion, the axes of body trace a cone and a gyro, whose input axis is aligned to this axis, senses an average angular rate equal to the solid angle swept per unit time even though there is no net rotation-taking place about that axis.

Coning motion originates when there are angular oscillations at frequency $f = \omega/2\pi$ also called coning frequency, about two perpendicular axes with phase difference ϕ that leads to a coning motion about the third axis as shown in Figure 2.20.

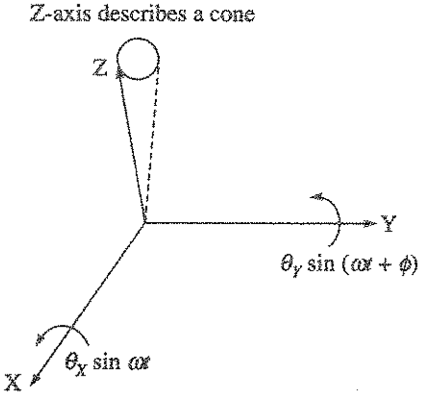


Figure 2.20 Coning motion illustration.

The oscillating angular motion about the X-axis and the Y-axis can be expressed as:

About X-axis: $\theta_X \sin \omega t$

About Y-axis: $\theta_Y \sin(\omega t + \phi)$

The average value of the motion about the third axis Z is then given as:

$$\left|\frac{\omega}{2}\right| \theta_X \theta_Y \sin\phi \tag{2.64}$$

Since the third axis describes a cone whenever the oscillating motions are out of phase, this kinematical phenomenon is called *coning*. The gyro aligned to the third axis will then respond to this average angular motion. In a strapdown system, the methodology by which such angular oscillation results in drift error is briefly explained. Under a typical situation when there is no coning motion and the angular rate ω is steady during the gyro update interval, the incremental angles, measured by the gyroscopes, are used to solve the attitude increment using the following approximate equation:

$$\Delta\phi \approx \Delta\theta = \int_{t-\Delta t}^{t} \omega \, dt \tag{2.65}$$

In the presence of oscillation in the other two axes having a phase difference, the above approximation leads to error in the computation of attitude update equation. This happens since the angular rate integration takes place at a finite computer sampling frequency, the oscillatory components in X-axis and Y-axis will not be faithfully reproduced, particularly if the coning frequency approaches or exceeds the computer sampling frequency, and a net error will be generated. This attitude computation error depends on coning frequency f, computer update rate interval Δt, coning motion angle magnitudes θ_X, θ_Y and phase difference ϕ. It is given by **[Kayton and Fried, 1997]**

$$\text{Drift error} = \pi f \theta_X \theta_Y \sin\phi \left(1 - \frac{\sin 2\pi f \Delta t}{2\pi f \Delta t}\right) \tag{2.66}$$

A straightforward approach may lie in increasing the transformation matrix update frequency, but this may not be always acceptable from high computer throughput requirement and instead preprocessing algorithms, known as coning algorithms **[Salychev, 2004]** have evolved which attempts to follow the rotation by acquiring and preprocessing the gyro data at higher rate.

The origin of such coning motion arises from 'S' turn manoeuvres in military aircrafts or locally from airframe vibration and also from rocking motion of the vibration isolator. Dither correction can also lead to coning error. Coning drift may require compensation to reduce the resulting drift within the acceptable level.

EXAMPLE 2.11 (on coning drift): Assume the coning motion frequency f is 50 Hz, the angular motion has an amplitude of 0.05° and a phase difference of 90°, and the attitude update rate is 0.02 s.

Solution: Given

Coning frequency $(f) = 50$ Hz, Phase difference $(\phi) = 90°$

$\phi_x = \phi_y = 0.05° = 0.05 \times \pi/180$ rad, Update rate $(\Delta t) = 0.02$ s

The coning drift is then computed using Eq. (2.66) as follows:

$$f\Delta t = 50 \times 0.02 = 1.0 \text{ implies } \sin(2\pi f \Delta t) = 0 \text{ and } \sin\phi = 1$$

So, coning drift error $= \pi \times 50 \times (0.05 \times \pi/180)^2 (1 - 0)$ rad/s

Simplifying and converting to °/h, we get

Coning drift error = 24.71°/h

The effect of high frequency attitude update rate can be visualised, if we change the update interval to 0.002 s with other parameters remaining same:

$$f\Delta t = 50 \times 0.002 = 0.1 \text{ implies } \sin(2\pi f\Delta t) = 0.587$$

Solving the bracketed term, we get

$$[1 - (0.587/2 \times 3.14 \times 0.1)] = [1 - 0.935] = 0.064$$

Therefore, the coning drift reduces by the factor 0.064, and the drift error becomes 1.58°/h.

Sculling error

Sculling phenomenon (refer Figure 2.21), is described when the strapdown system is experiencing a sinusoidal angular motion $\theta_0 \sin 2\pi ft$ about an axis-X_B in conjunction with a synchronised but linear oscillating acceleration $A_0 \sin 2\pi ft$ along a perpendicular axis-Y_B. Here f is called sculling frequency.

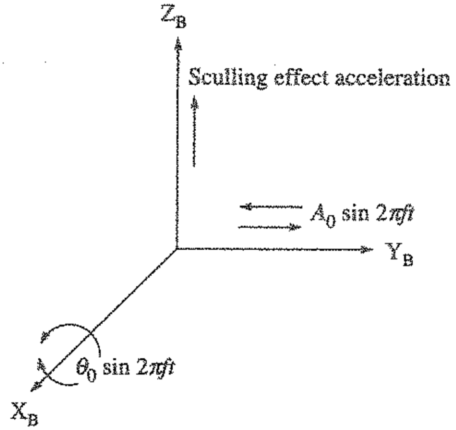


Figure 2.21 Sculling error illustration.

This situation can lead to an error that is characterised as a steady acceleration along the third axis-Z_B that is orthogonal to both rotation as well as acceleration. If the acceleration measured in body frame is transformed to the desired navigation frame continuously and integrated to velocity, no error is introduced and this is equivalent to analog mechanisation of specific force transformation equation. However, in an actual practice of an iterative form of typical digital computer mechanisation where integration and transformation are done at discrete intervals, the process can lead to error in the presence of sinusoidal rotation that is synchronised with an oscillating acceleration and when the rotation is not sampled at adequate frequency.

The discrete form of transformation mechanisation uses $\Delta\theta^B$ as gyro data and ΔV^B as accelerometer data respectively (refer Figure 2.18). We further note that oscillating body rate about X_B-axis can be expressed as:

$$(\dot{\theta}_0)^B(t) = \frac{d}{dt}[(\theta_0)^B(t)] = [2\pi f\theta_0 \cos 2\pi ft]$$

These sensors data can be expressed in the update interval Δt as:

$$\Delta\theta^B(X) = [2\pi f\theta_0 \cos 2\pi ft]\,dt$$

$$\Delta V^B(Y) = [A_0 \sin 2\pi ft]\Delta t \tag{2.67}$$

Since the effect of these motions is found on the Z_B-axis, the sculling effect acceleration can be represented **[Kayton, 1997]** as:

$$\Delta\dot{V}_{\mathrm{I}}(Z) = \frac{1}{2}\theta_0 \mathrm{A}_0\left(1 - \frac{\sin 2\pi f\Delta t}{2\pi f\Delta t}\right) \tag{2.68}$$

where $\Delta\dot{V}_I(Z)$ is the average sculling effect acceleration error along Z_B-axis.

Reduction of this error is achieved through much higher frequency computation of velocity transformation matrix relative to the sculling frequency, which is further explained through Example 2.12.

Example 2.12 (on sculling error reduction): Assume that amplitude $A_0 = 2g$ along the axis-Y_B for a sinusoidal acceleration of frequency 50 Hz which is in phase with an oscillating angular motion θ_0 of magnitude 500 × 10^{-6} rad about X_B-axis. Compute the sculling error when the velocity transformation is done at

(i) five times the sculling frequency

(ii) twice the sculling frequency

Solution: Given $A_0 = 2g$, $\theta_0 = 500 \times 10^{-6}$ rad; $f = 50$ Hz

(i) At 5 × 50 = 250 Hz transformation update frequency, $\Delta t = 0.004$ s

Sculling error $\Delta\dot{V}_I(Z)$ is estimated as:

$$\Delta\dot{V}_I(Z) = \frac{1}{2}500\times10^{-6}\times2\left(1 - \frac{\sin(2\times3.14\times50\times57.4\times0.004/2)}{(2\times3.14\times50\times0.004/2)}\right)g$$

$$= 500\left[1 - \left(\frac{0.588}{0.6280}\right)\right]\mu\mathrm{g}$$

$$= 31.8\ \mu\mathrm{g}$$

(ii) At 2 × 50 = 100 Hz update frequency, $\Delta t = 0.01$ s

The computed sculling error is:

$$= 500\left[1 - \left(\frac{0.732}{1.57}\right)\right]\mu\mathrm{g}$$

$$= 266\ \mu\mathrm{g}$$

The example clearly brings out the benefit of high frequency update rate in reducing the sculling error. Some special purpose sculling algorithms have also been developed **[Savage, 1999]** with the objective of reducing the load on computer.

Size effect error

All the three single axis accelerometers in a strapdown system cannot be collocated on the cluster. In the presence of angular motion, each accelerometer will sense centripetal and/or tangential acceleration. The magnitude of error due to centripetal acceleration is proportional to $\Omega^2 l$, whereas the error due to tangential acceleration is proportional to $\dot{\Omega} l$ where l is the lever arm distance of accelerometer proof mass from the axis of rotation, Ω is the vehicle angular rate and $\dot{\Omega}$ is the vehicle angular acceleration. Since, each accelerometer will have a different lever arm distance; the three accelerometer outputs will be different from each other for a given angular rate. Also, the angular rate vector can be resolved along the vehicle principal axes, and this in turn can be used to compute the magnitude of centripetal acceleration on each accelerometer. For a constant vehicle rate, the effect on navigation can average out to zero if adequate transformation update rate is chosen. However, when the vehicle rate is oscillatory, a constant bias is introduced due to rectification effect. Most of these errors due to size effect can be taken care in software design. In a gimbaled platform system, this error does not manifest, as the cluster is isolated from vehicle motion.

2.9.5 Strapdown Navigation in Geographic Frame

Earlier, Eq. (2.27) has defined the fundamental navigation equation in earth frame and where Geographic Frame, *G-frame*, is a subset of such earth-based frames, which finds considerable application in aircraft navigation. There are a few types of such *G-frame* which were discussed earlier in Section 2.4.3. Equation (2.27) can be expressed in the *G-frame* as:

$$\left[\frac{d\mathbf{V}^{G}}{dt}\right] = C_B^G \mathbf{f}^{B} + \mathbf{g}_p - (\boldsymbol{\omega}_G + \boldsymbol{\omega}_e) \times \mathbf{V}^{G} \tag{2.69}$$

Equation (2.69) conveys the following information:

(a) *G-frame* is ensuring a local vertical frame for a spherical earth or an ellipsoidal earth model. This means that $\mathbf{g}_p$ remains perpendicular to the surface of earth with no horizontal components.

(b) One navigation axis is always pointing to true north.

(c) $\boldsymbol{\omega}_G$ indicates *G-frame* angular rate vector relative to inertial space which incorporates the correction term due to motion of the vehicle over the earth.

(d) Sensed body frame acceleration vector $\mathbf{f}^{B}$ needs to be transformed to *G-frame*, and then compensated for the effect of Coriolis acceleration that is shown by the third term.

In component form, $\boldsymbol{\omega}_G = [\ \omega_X\ \omega_Y\ \omega_Z]^T$, and these components can be expressed as:

$$\omega_X = -\frac{V_Y}{R_m + h}$$

$$\omega_Y = \omega_e \cos L + \frac{V_X}{R_p + h}$$

$$\omega_Z = \omega_e \sin L + \frac{V_X}{R_p + h} \tan L \tag{2.70}$$

where

$\frac{V_Y}{R_m + h}, \frac{V_X}{R_p + h}$ = north and east position angular rates respectively with respect to earth frame and these terms are commonly known as *transport rate*

V_X = east velocity of the vehicle

V_Y = north velocity of the vehicle

R_m = radius of curvature of earth in meridional plane

R_p = radius of curvature of earth in the vertical east-west plane

h = vehicle altitude from the defined surface of the earth

L = geodetic latitude

After executing the vector cross product operation, the Coriolis acceleration compensation terms for the accelerometers can be expressed as:

For X accelerometer: $(\omega_Y + \omega_e \cos L)V_Z - (\omega_Y + \omega_e \sin L)V_Y$ (2.71)

For Y accelerometer: $(\omega_Z + \omega_e \sin L)V_X - (\omega_X V_Z)$ (2.72)

For Z accelerometer: $\omega_X V_Y - (\omega_Y + \omega_e \cos L)V_X$ (2.73)

After incorporating the gravity compensation and the Coriolis acceleration correction, the navigation frame acceleration $\dot{V}^G = [\dot{V}_X \ \dot{V}_Y \ \dot{V}_Z]^T$ is integrated to give velocity components V_X (east velocity), V_Y (north velocity) and V_Z (vertical velocity). Since horizontal position components are expressed in geodetic latitude L and longitude ϕ, these are obtained from the following expressions:

$$L = \int_0^t \frac{V_Y}{R_m + h} dt + L_0 \tag{2.74}$$

$$\phi = \frac{1}{\cos L} \int_0^t \frac{V_X}{R_P + h} dt + \phi_0 \tag{2.75}$$

The radius of curvature R_m is defined as the radius of the best fitting circle with respect to the reference ellipsoid meridian section. Similarly, R_p is defined as the radius of the best fitting circle with respect to reference ellipsoid vertical east-west section. These are approximated **[Kayton, 1997]** as:

$$R_m \approx R_e \left[1 + e^2 \left(\frac{3}{2} \sin^2 L - 1 \right) \right] \tag{2.76}$$

$$R_p \approx R_e \left(1 + \frac{e^2 \sin^2 L}{2} \right) \tag{2.77}$$

Equations (2.76) and (2.77) can also express these coefficients for a spherical earth model by putting $e = 0$.

Altitude is normally referenced to the mean sea level. The inertial computation of altitude involves use of specific force in *G-frame*, use of gravity model and correction for Coriolis acceleration component, all for Z-axis. This equation can be expressed using Eq. (2.69) in component form for Z-axis as follows:

$$\dot{V}_Z^G = f_Z^G + g_v^G - (\text{Coriolis terms}) \tag{2.78}$$

It has been shown by **[Salychev, 2004]**, that neglecting the Coriolis term and computing altitude h with the vertical accelerometer data in the presence of vertical accelerometer constant bias error B_Z, leads to an undamped error. This can be explained by assuming the gravity model, [Eq. (2.14)], in the form:

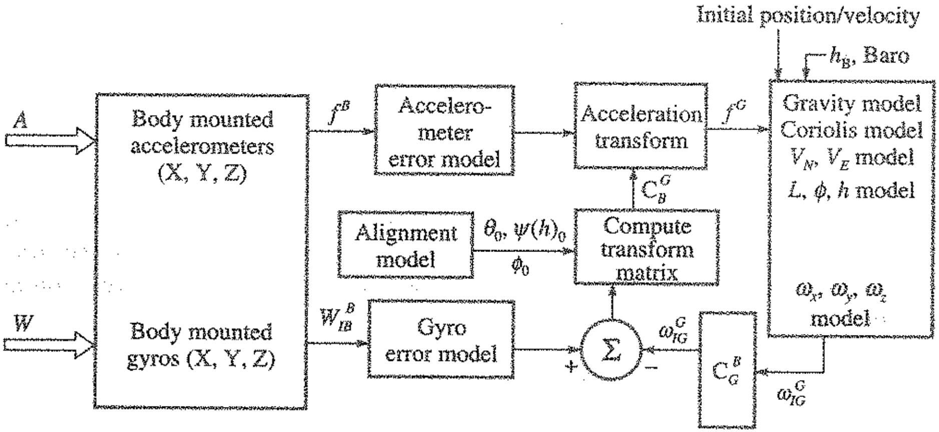
$$g_v = g_{\text{eq}} \left(\frac{R_{\text{eq}}}{R_{\text{eq}} + h} \right)^2$$

Expressing δg_v as $\delta g_v = -2\omega_s \delta h$ and then rewriting Eq. (2.78) with an altitude error term δh, results in

$$\delta \ddot{h} - 2\omega_s^2 \delta h = \delta f_z = B_Z \tag{2.79}$$

where, $\omega_s = \sqrt{g_e / R_e}$, the Schuler frequency.

The solution of Eq. (2.79) leads to an expression for h that has an exponential time growing property. As a result of this error growth with time, inertial computation of altitude needs to be suitably damped before any use. Prior to satellite navigation, the most convenient form of damping and altitude correction has been by baro-altimeter data that is still in use. Features of such a strapdown navigation system with Baro-damped altitude channel, are shown in Figure 2.22.



A : Inertial acceleration
W : Angular rate
f^B : Specific force sensed in body frame
f^G : Specific force transform to *G-frame*
ω_{IG}^G : Transport rate torquing in *G-frame*
$\theta_0, \psi(h)_0, \phi_0$: Initial attitudes
L, ϕ, h : Latitude, Longitude, Altitude

Figure 2.22 Strapdown system schematic implementing geographic frame navigation.

The first block on the left side is the sensor cluster where the sensors provide the information on specific force in body frame as well as inertial angular rate in body frame. All other blocks represent the tasks carried out in the navigation computer at least in two computational frequencies which have been shown earlier in Figure 2.19. The co-ordinate transformation matrix C_B^G is computed using the gyro data as well as the computed torquing data shown in Figure 2.22, where, $\omega_G = [\omega_X\ \omega_Y\ \omega_Z]^T$. The horizontal position components are computed to provide the latitude and longitude of the vehicle, while velocity components are computed as north velocity and east velocity.

The inertial altitude is provided as suitably baro altitude damped, which in turn results in accuracy, adequate bandwidth and low noise. This feature of altitude output is advantageous in quite a few airborne applications for the computation of inertial altitude, where the vertical gravity model can be avoided and replaced by a constant gravity.

2.9.6 Singularities and Their Avoidance

In the navigation system, described in Section 2.9.5, a few singularities are observed at $L = 90°$ and at 270°, which are known as polar crossing zones, where the system cannot compute longitude ϕ or provide the transport rate correction [refer Eqs. (2.70) and (2.75)].

Figure 2.23 describes the problem.

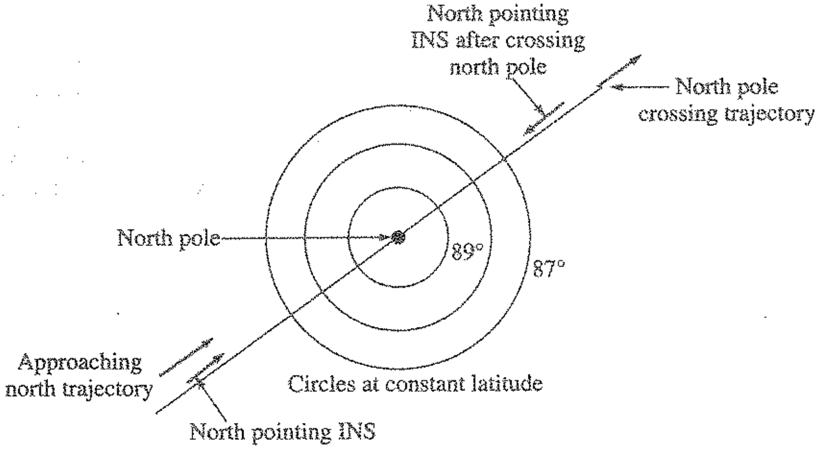

North pointing INS after crossing north pole
North pole crossing trajectory
North pole
89°
87°
Approaching north trajectory
Circles at constant latitude
North pointing INS

Figure 2.23 Geographic reference frame singularity near pole crossing.

A rapid change of longitude near the pole leads to high torquing rate and the direction of instrumented north suddenly changes by 180° as the pole is crossed. Alternate navigation frames have emerged towards solution to this problem, and a typical of such frames is called *wander azimuth*. In wander azimuth scheme, the azimuth axis is torqued at a rate $\dot{\alpha}$ so that the second term in Eq. (2.70) is cancelled. This means that the sign and magnitude of $\dot{\alpha}$ is given by

$$\dot{\alpha} = -\frac{V_X}{R_p + h}\tan L \tag{2.80}$$

The azimuth axis torquing equation becomes:

$$\omega_Z^N = \dot{\alpha} + \omega_e \sin L \tag{2.81}$$

with this implementation, singularity in the torquing equation over the poles is avoided.

2.10 Redundant Inertial Systems

Reliability in inertial system is essential to the success of any aerospace mission. Redundant inertial systems are in operation for various critical applications where the mission reliability needs to be enhanced by tolerating one or more random failures in the system. Typically, such redundant systems are considered for the following reasons:

1. When human beings are involved in the vehicle as in aircrafts or in reusable manned launch vehicles
2. Long duration mission like orbital satellites
3. When cost of the vehicle and the missions are high as in satellite launch vehicle and interplanetary missions

Such redundant systems can be inactive or active spare. In inactive scheme, the redundant hardware is activated only when the primary hardware starts malfunctioning. Active redundant systems are needed in applications where very little time is available to diagnose the faulty element, isolate the failed element and proceed with the critical mission functions. The engineering design of such active redundant systems has evolved over a long time undergoing considerable changes due to availability of improved algorithms, faster onboard computers and change over from gimbaled system to strapdown system. The challenge to the inertial system redundancy design has put emphasis in providing reliable and active redundancy schemes, which will work in minimum time during the time the system is in distress. Such active redundancy is further discussed.

However, redundancy is not a solution for improving reliability if there are weak elements in the system design. In the event of distress, the weak elements in both the main and redundantareas are most likely to fail together. This means that the redundancy management assumes that the nominal design of the sensors and the relevant electronics is sound and the failure in these can be treated as random.

2.10.1 Failure Detection and Isolation

Failure detection and isolation (FDI), is a method incorporated in the inertial system design by which the failed sensor element is identified, isolated automatically if necessary and reconfigure itself to maintain the continuity of system operation. The primary choice in strapdown inertial system FDI design is in the incorporation of algorithmic process and supported by the provision of adequate redundant hardware. In the context of redundancy in inertial systems, it means computer processing of the sensor output. The redundancy architecture ensures that all the sensor outputs, along with their associated electronics, are available up to the computer.

Generalised requirement for such software FDI scheme stipulates that the number of instruments must exceed the dimensions of the measured quantity by at least two. This means that at least three sensors are needed in one-dimensional space or at least five non-collinear sensors in three-dimensional space. It should be noted that inertial navigation deals with three-dimensional space.

Figure 2.24 depicts the configuration of an assumed inertial system scheme with nine single axis gyros and accelerometers which are mounted on an orthogonal cluster frame X, Y, Z.

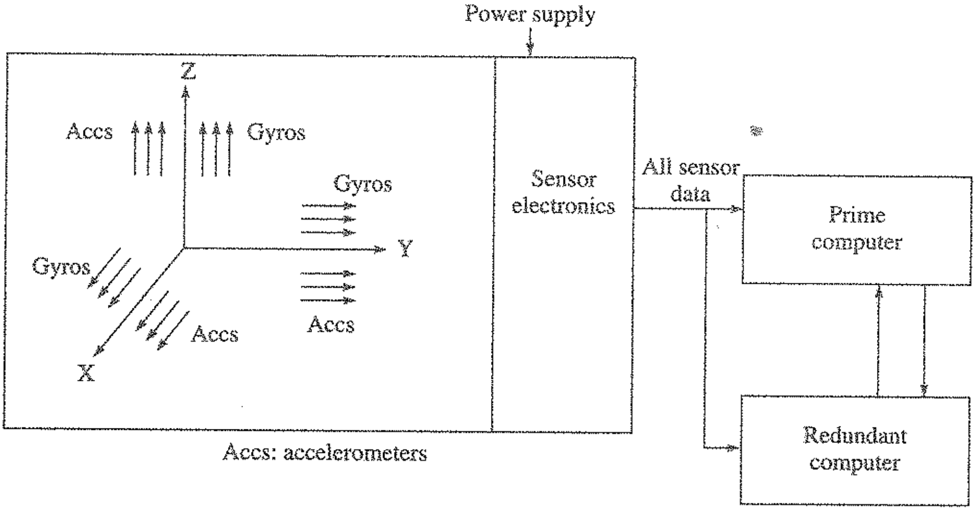


Figure 2.24 Orthogonal cluster frame with nine gyros and accelerometers.

The scheme ensures that there are three instruments, three each of gyros and accelerometers, along each of the body frame axes. It is assumed that each sensor is associated with its own electronics and powering scheme, and there is no electrical coupling between the sensor chains.

FDI schemes have been developed with different blends of complexity to provide efficient redundancy management coverage. Some such schemes are described along with the Figure 2.24

Mid value selection provides the simplest of algorithmic FDI scheme. In the mid value selection scheme, each sensor output data, aligned to a common axis, is arranged in ascending order of magnitude and then its mid value is selected for further use. The theoretical understanding is that, with any one failure, the mid value will be that of the healthy sensor only and the data can be used without the need to identify the failed sensor. Thus uninterrupted system operation is ensured. The scheme tolerates any one random failure of gyro and accelerometer from the available nine sensors in each. The scheme can tolerate two more sequential failures if they happen in the un-failed axes. But in FDI sense, these failures are termed as *conditional failures*.

While the scheme is algorithmically simple, it is a costly scheme as nine sensors along with their associated electronics are required in the system for any one random failure tolerance.

FDI using parity equations is another scheme which can be used for the configuration shown in Figure 2.24 and is discussed further as it can provide much better capability in the design as compared to mid value selection scheme. The output of the set of three gyros, aligned to X-axis, for example, can be written as follows when a common angular rate input ω_x is acting on them.

$$\begin{aligned} G_{x1} &= \omega_x + e_1 \\ G_{x2} &= \omega_x + e_2 \\ G_{x3} &= \omega_x + e_3 \end{aligned} \tag{2.82}$$

where

G_{x1}, G_{x2}, G_{x3} = gyro measurements

e_1, e_2, e_3 = measurement errors in the gyros

Since, all the gyro measurements are directly made about the X-axis, the following three parity equations can be formed:

$$\begin{aligned} G_{x1} - G_{x2} &= e_1 - e_2 = \text{parity residue} \\ G_{x2} - G_{x3} &= e_2 - e_3 = \text{parity residue} \\ G_{x3} - G_{x1} &= e_3 - e_1 = \text{parity residue} \end{aligned} \tag{2.83}$$

When all the three gyro outputs are error free, then all the three parity residues will be ideally zero indicating fault free operation. In the event one gyro is not error free, the parity residues, containing the faulty gyro, will not be zero, whereas it will be zero for the parity equation formed with the other two healthy gyros. From this, the failed gyro can be uniquely identified and isolated. This method of forming parity equations for the purpose of FDI indicates that it is necessary to have minimum three measurements along all the three axes of the orthogonal sensor body frame. The process of isolation is associated with reconfiguration as hereafter there will be only two healthy sensors to carry on the system operation. It is interesting to note that only measurement error, not the absolute value of vehicle rate, appears in the parity equations. Further, when two healthy sensors are only operating and if a problem develops in one of these two, the residue will be changed indicating the problem, but the sensor cannot be isolated as it is not clear which of the two is having the problem.

In reality, even a healthy sensor will have a permissible error associated with its normal output. To prevent such a healthy sensor getting declared as faulty and then isolated, the parity equations are redefined as follows:

$$\begin{aligned} e_1 - e_2 &> \text{Threshold} \\ e_2 - e_3 &> \text{Threshold} \\ e_3 - e_1 &> \text{Threshold} \end{aligned} \tag{2.84}$$

In this formulation, the parity residue has to exceed a finite threshold value to declare that a sensor has failed. The threshold value is selected from the permissible performance statistics of the sensors. Exactly, in a similar manner, the accelerometer parity equations can be formulated. In the case of accelerometers, acceleration will be the physical input. In mathematical sense, ω can be treated as a physical vector.

2.10.2 Hard and Soft Failure

Failure modes in inertial sensors have shown both 'hard' and 'soft' mode of failures. A 'hard' type of failure indicates gross performance change, which will reflect in large residue and comparatively easy to detect and isolate. A 'soft' failure is defined as a performance shift that is just beyond the permissible specification, and its real time detection and isolation become difficult. The soft failure detection poses interesting software design problem due to presence of noise in the sensor output, changing vehicle dynamics and for ascertaining if the degradation is temporary or permanent.

2.10.3 FDI with Optimal Skewed Configuration

Skewed sensor configuration, rather than orthogonal configuration, provides optimality between the number of sensors used and the failure tolerance it can provide. The earlier example of Figure 2.22 with nine sensors is not at all optimal because what is needed for parity equation formulation is minimum three measurements along each of the three orthogonal body frame axes and not necessarily three instruments along a particular axis. This means that proper geometrical skewing of the sensor input axis with respect to the orthogonal body frame can implement the parity equations with much less number of sensors. As a result, considerable research has gone in generating the most efficient geometrical sensor arrangements, and the study has shown that the efficient arrangements **[Pejas, 1971]** fall in one of the following two categories:

(a) Those with an odd number of single axis sensors with their input axes uniformly distributed around the surface of a cone.

(b) Those with an even number of single axis sensors with input axes uniformly distributed around a cone surface with one sensor input axis lying along the central axis of the cone.

The cone angle is selected to equalise the estimation error in all directions. For n odd numbered sensors, the cone angle is found to be 54.74°, while for the even category; the cone angle is given by

$$\cos^2 \alpha = \frac{n-3}{3n-3} \tag{2.85}$$

when $n = 6$, $\alpha = 63.43°$.

With five single axis sensors, Figure 2.25 shows a typical conical arrangement along with the orthogonal body frame axes X, Y, Z.

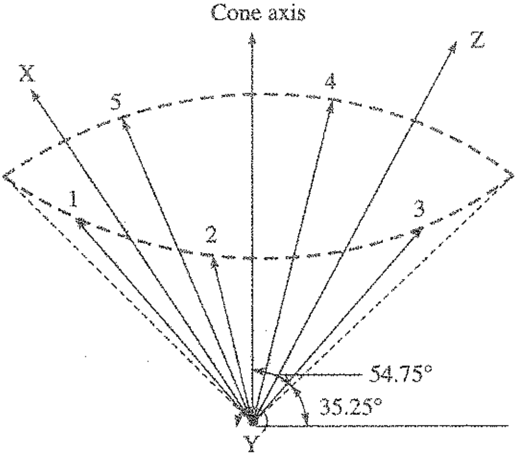


Figure 2.25 Five single axis inertial sensors arranged over conical configuration.

The skewing arrangement is so selected that even with any two sensor failure; the remaining three healthy sensors will not lie in a plane, which is required for three-dimensional navigation. For optimal skewing, GDOP is computed and should be low enough.

For a skewed array sensor configuration, Table 2.3 shows the FDI capability with different number of sensors.

Table 2.3 Skewed Array Sensors FDI Capability

No. of sensors	First failure	Second failure	Third failure	Fourth failure
3	—	—	—	—
4	D	—	—	—
5	DIR	D	—	—
6	DIR	DIR	D	—
7	DIR	DIR	DIR	D

D = detected only, DIR = detected, isolated and reconfigured.

Table 2.3 shows that with four sensors, detection is possible but no isolation, with five sensors, one random failure can be tolerated and with six sensors, two random failures can be tolerated. Thus we see that from the redundancy point of view, five sensors in skewed configuration is same as to that of nine sensors in orthogonal configuration where both provides only one failure tolerance.

The navigation performance, however, degrades from the un-failed sensor geometry to reconfigured geometry after failure. This is due to increase in GDOP (Geometrical Dilution Of Precision) associated with the sensor arrangement geometry change. The standard deviation of error in the sensor output data to be used in navigation computation increases by a factor given by the GDOP of the sensor geometry.

The optimality of sensor arrangement on a cluster can be mathematically defined with GDOP, which is expressed as:

$$\text{GDOP} = \text{ Square root of the Trace of } [H^T H]^{-1} \tag{2.86}$$

Here, H matrix provides the geometrical arrangement of the sensor input axis on the cluster relative to the orthogonal body frame, so that we can write the measurement equation as:

$$[\text{m}] = [H][\text{w}] \tag{2.87}$$

where

[m] = sensor measurement matrix

[H] = sensor geometrical arrangement matrix relative to body frame

[w] = physical input vector (rate or acceleration) acting on the orthogonal body frame

Minimising GDOP, minimises the INS error, hence H matrix plays an important role in INS performance.

For skewed sensor cluster geometry starting with six un-failed sensors, the change in GDOP and its impact on navigation accuracy [**Vanderwerf et al., 1983**] with the operating sensors is shown in Table 2.4.

Table 2.4 Sensor Geometry Effect on GDOP and Navigation

Operating	GDOP	Navigation	Reconfiguration
6	1.22	1.3	No failure
5	1.41	1.6	One failure
4	1.73	2.0	Two failure

H matrix is defined from the sensor geometry and it is a $n \times 3$ matrix where n is the number of single axis sensors and using the H matrix, GDOP can be worked out using Eq. (2.86). From this H matrix, the sensor measurements can be resolved on the orthogonal frame and from these resolved components parity equations are formed axis wise. When a failure occurs and the sensor gets isolated, H is recomputed and along with it the GDOP.

EXAMPLE 2.13 (on FDI): Figure E2.3 shows an orthogonal body frame where three single axis gyros (G_X, G_Y ,G_Z) are mounted with their input axis as shown. This scheme is normally called *orthogonal sensor layout*. Will this configuration provide algorithmic redundancy?

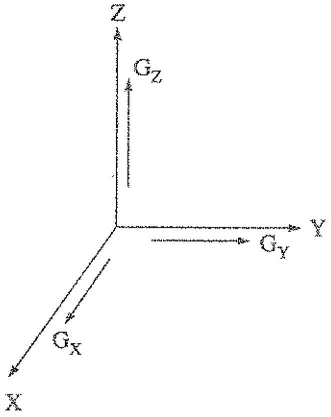


Figure E2.3 Orthogonal sensor layout.

Solution: As per the requirement of algorithmic redundancy, the number of minimum sensors needed is 3 + 2 = 5. So, the scheme cannot provide algorithmic redundancy.

How many more sensors are needed to provide one failure tolerance and in which way they should be mounted?

Now, the method of mounting should be such that there are three measurements possible in each of three axes. If the 4th gyro is mounted at 45° to X-axis and to Y-axis as well as 45° out of XY plane and towards Z-axis, this gyro will have resolved components on all the three axes. The arrangement of 4th gyro is shown in Figure E2.4. Similarly, not shown in the figure, if the fifth gyro is mounted 45° to Y and Z-axes as well 45° out of YZ plane, then this gyro will also have resolved components on all three axes. Hence, such arrangement will provide three measurements in each of three axes X, Y, Z. Forming parity equations, any one sensor failure can be uniquely identified. Figure E2.4 is normally called an ortho-skewed layout.

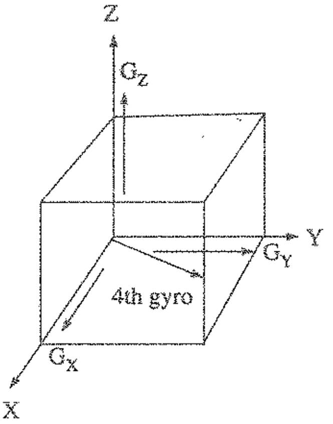


Figure E2.4 Ortho-skewed sensor layout.

EXAMPLE 2.14 (on GDOP of sensor with orthogonal mounting): Theoretically, three-dimensional navigation is possible with any combination of sensor arrangement with a stipulation that the sensor input axes are not coplanar. In the case of three single axis sensors mounted in the orthogonal manner, as shown in Figure E2.3, compute the GDOP:

Solution: H matrix, in this case, turns out to be a unit matrix and can be written as:

$$H = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Executing the matrix computation $[H^T H]^{-1}$, we get the result as H.

This means that the square root of trace of the matrix is $\sqrt{(1 + 1 + 1)}$ which is 1.73 and is the GDOP value for such type of sensor arrangement. Any other sensor arrangement with three sensors only increases the GDOP to prove that the orthogonal layout is optimum from performance point of view.

2.10.4 Use of Built In Test Equipment (BITE)

Algorithmic FDI, based on sensor outputs, has been found to give an efficient and cost effective strapdown redundant system. The scheme is normally based upon a configuration whereby each sensor has its individual power and relevant electronics up to the computer. This is normally called a sensor chain, and for the skewed configuration of Figure 2.25, there will be five such chains. However, the algorithmic scheme has a drawback, as even with four operating sensor chains, detection is possible for another failure (can be called second failure) out of these four, but the failed sensor chain cannot be isolated. With the failed sensor remaining in computation, navigation degrades and may lead to serious problem. It is also known that a three-axis measurement needs sensor geometry with only three non-coplanar sensors and that is provided by the skewed array even with any three operating sensor chains. In such a situation, a system designer looks for certain additional measurements, which can assist in identifying the failed sensor, second failure, when four sensors have been in operation prior to the failure.

Inclusion of such measurements provides the basis for BITE. The precise nature of measurement that can be additionally included depend on the *Failure Mode Effect and Critical Analysis (FMECA)*. For example, a gyro with rotating wheel configuration, where the wheel has a higher probability of failure, the wheel synchronism status [**Bose et al., 2008**] can be included as an additional measurement. It is necessary that the electronics for BITE be of high reliability, as a failure in this area will create additional decision-making problem.

2.10.5 Navigation Computation with Redundant Sensors

In redundant system, since additional sensor measurements are available, all the sensor measurements are normally used using least square estimation. Least Square Estimation (LSE) transformation, given by the matrix $[(H^T H)^{-1} H^T]$, computes the best estimates of three resolved components along the body frame from the available m number of sensor measurements. This improves performance as shown in Table 2.4 due to GDOP improvement. So, navigation, with redundant sensors, typically follows the following sequence of computations:

Sensor output → Sensor error compensation → FDI implementation → LSE → Data in body frame

However, the primary aim with additional sensors is to provide efficient redundancy management, while the additional benefit on performance is considered as secondary.

SUMMARY

The chapter has initially brought out the principle of strapdown inertial navigation explaining the terminologies like transformation matrix, ellipsoid shape of earth, gravitational acceleration and gravity. Frames of reference for navigation including inertial frames as well as non inertial earth-based frames were described as well as additional strapdown frames needed for operation of the system were explained. Strapdown equations were developed for operation in inertial as well as non inertial earth-based rotating navigation frame. Method of computing the transformation matrix using various schemes as well as self alignment scheme was explained. Error propagation in strapdown navigation and the importance of strapdown software with multi-frequency update scheme were brought out. Lastly, in-flight redundancy management scheme, with emphasis in software technique, was explained at length.

EXERCISES

2.1 A satellite is injected into space over the equator to a height of 400 km circular orbit to conduct zero gravity experiments. What is the gravitational acceleration acting on the satellite?

(i) 0 (ii) Some non zero value, if so what is the value?

[**Ans:** (ii) Some non zero value; 8.665 m/s^2]

2.2 A satellite is injected into a circular orbit of 700 km altitude with a velocity of 7.4 km/s. Which type of onboard system can be used to know the satellite position in the orbit?

(i) INS (ii) GPS [**Ans:** (ii) GPS]

2.3 In a homogeneous ellipsoidal earth model, gravitational acceleration passes through the centre of the ellipsoidal earth.

(i) True (ii) False [Ans: (ii) False]

2.4 An inertial navigation system, mechanising local level geographic frame, measures the altitude information using the vertical accelerometer data which is stable and accurate.

(i) True (ii) False [Ans: (ii) False]

2.5 In an INS with high GDOP is an indication of high navigation performance.

(i) True (ii) False [Ans: (ii) False]

2.6 Vector form of inertial navigation is expressed as:

$$\mathbf{A}^{\mathrm{I}} = \frac{d^2\mathbf{R}^{\mathrm{I}}}{dt^2} = \mathbf{f}^{\mathrm{I}} + \mathbf{g}^{\mathrm{I}}$$

where gravitation is computed using position information.

Write the above equation in component form and represent it in block schematic diagram and from it briefly explain inertial navigation operation.

2.7 A train travels from equator to North pole at uniform speed V relative to earth. Determine the Coriolis acceleration as a function of latitude L. If $V = 40$ m/s, determine the Coriolis acceleration at the equator and at the pole. Assume earth rate ω_e as 15.04°/h.

[Ans: $2\omega_e \sin LV_N$, 0, 5.83×10^{-3} m/s^2]

2.8 The error propagation characteristic of an INS in inertial frame can be expressed as:

$$\delta\mathbf{f}^{\mathrm{I}} = C_B^I \delta\mathbf{f}^{\mathrm{B}} + \delta C_B^I \mathbf{f}^{\mathrm{B}}$$

For a two-dimensional navigation and assuming that the error is only in rotation sensing, derive the navigation error equations when there is rotation of θ about the Y-axis. Initially, both the frames were coincident. If the Y-axis gyro is having a drift of 0.1°/h, calculate the position error components at the end of 200 s? Assume initial rotation of 10° and a sustained acceleration in body frame components as 10 m/s^2.

[Ans: X-axis = 10.65 m, Y-axis = 7.46 m]

2.9 Given the following information: Velocity measured in *i-frame* (ECI-frame) as $V_x = V_y = Vz = 50$ m/s and the transformation matrix *G-frame* (NED) to *i-frame* as:

$$\text{Transformation matrix } C_G^i = \begin{matrix} -\sin L\cos\phi & -\sin\phi & -\cos L\cos\phi \\ -\sin L\sin\phi & \cos\phi & -\cos L\sin\phi \\ \cos L & 0 & -\sin L \end{matrix}$$

where, $L = \phi = 30°$

Compute the velocities in *G-frame* identifying which are their geographic directions?

[Ans: $V_N = 9.0$ m/s, $V_E = 18$ m/s, $V_D = -83.48$ m/s]

2.10 What is meant by active FDI? In a two-dimensional navigation, such as on surface of earth, how many single axis sensors are minimum required for providing algorithmic FDI? Sketch a possible arrangement of these sensors and explain FDI operation. Write down the H matrix also.

2.11 Briefly explain the significance of the term GDOP with respect to sensor arrangement on the cluster?

In an INS with three sensors, what is their arrangement with the body frame that gives the best GDOP? Compute the GDOP value. When the number of sensors is increased to six, will the performance improve and if so why? **[Ans:** GDOP = 1.73]

2.12 The orthogonal transformation matrix between *G-frame* (ENU) to *E-frame* is given as:

$$\begin{bmatrix} -\sin\phi & -\sin L\cos\phi & \cos L\cos\phi \\ \cos\phi & -\sin L\sin\phi & \cos L\sin\phi \\ 0 & \cos L & \sin L \end{bmatrix}$$

where L is the latitude and ϕ is the longitude. Initial position of the vehicle is at zero latitude and longitude.

(i) Compute the initial transformation matrix C_E^G.

(ii) If the vehicle moves along north–east so that $L = \phi = 45°$, compute the new C_E^G.

$$\text{[\textbf{Ans:} (i) } C_E^G = \begin{bmatrix} -\sin\phi & \cos\phi & 0 \\ -\sin L\cos\phi & -\sin L\sin\phi & \cos L \\ \cos L\cos\phi & \cos L\sin\phi & \sin L \end{bmatrix} \text{ (ii) } \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}]$$

REFERENCES

Bose, A., Puri, S.N., and Banerjee, P., *Modern Inertial Sensors and Systems*, PHI Learning, Delhi, 2008.

Britting, K.R., *Inertial Navigation System Analysis*, Wiley-Interscience, New York, 1971.

Chatfield, Averil B., Fundamentals of High Accuracy Inertial Navigation, Progress in Astronautics and Aeronautics, AIAA, **174**, USA, 1997.

Fernandez, M. and Macomber, George R., *Inertial Guidance Engineering*, Prentice Hall, New Jersey, 1962.

International Gravity Formula, Wikipedia, 1967.

Kayton, Myron and Fried, Walter R., *Avionics Navigation system*, 2nd ed., John Wiley & Sons, 1997.

Pejas, Arthur J., Optimum Orientation and Accuracy of Redundant Sensor Arrays, AIAA, 71–59, New York, 1971.

Salychev, Oleg., *Applied Inertial Navigation: Problem and Solutions*, BMSTU Press, Moscow, Russia, 2004.

Savage, Paul G., *Introduction to Strapdown Inertial Navigation Systems*, Strapdown Associates, Maple Plain, MN55359, November 1999.

Titterton, D.H. and Weston, J.L., *Strapdown Inertial Navigation Technology*, 2nd ed., Vol. 207, Progress in Astronautics and Aeronautics, Co-published by AIAA-USA and IEE, United Kingdom, 2004.

US Defense Mapping Agency, World Geodetic System, 1984 (WGS 84), Its Definition and Relationship with Local Geodetic System, Maryfield, VA, Washington DC, 1991.

Vanderwerf, K. and Weaid, K., Fault Tolerant Inertial Navigation System, AIAA, New York, 1983.

3

Gyros

Gyro is an instrument, which senses inertial angular motion about its input axis without external reference. While all gyros sense inertial angular motion about its input axis, there are two broad usage classifications. The first one relates to providing inertial reference, while the second one relates to providing angular rate information. It is the gyro of the first usage type that had provided and continues to provide significant research and development for improved performance and reliability in association with the reduction of cost, power consumption and weight.

There are several applications of gyros where the performance is closely related to the targeted application. These applications span from aerospace to land and to ships and submarines. The gyros covered in this chapter are primarily meant for use in aerospace where providing inertial reference is primarily targeted, while the angular rate information is obtained either from the inertial reference gyros or from another set of gyros.

Over the last five decades, a host of physical laws were utilised to develop operational gyros. The typical physical laws are as follows:

- Gyros based on conservation of angular momentum of a spinning rotor
- Gyros based on Coriolis effect on a vibrating mass
- Optical gyros based on Sagnac effect

Current research is active on a new type of gyro based on cold atom interferometry. Subsequent sections will provide some detailed aspects on these gyros.

3.1 Spinning Rotor Gyro Principle of Operation

Assume a rotor with inertia J is spinning with angular rate ω to produce an angular momentum vector $\mathbf{H}$ as shown in Figure 3.1. If a torque vector $\mathbf{T}$ acts on this rotor perpendicular to $\mathbf{H}$, it causes the spin axis to precess by an angle $d\theta$ and vectorially, the change in momentum is:

$$d\mathbf{H} = \mathbf{H}d\theta$$

The precession happens as the magnitude of H is maintained constant; as a result, the applied torque causes a rotation of the spin axis. Direction of precession is such that the spin axis moves towards the torque axis.

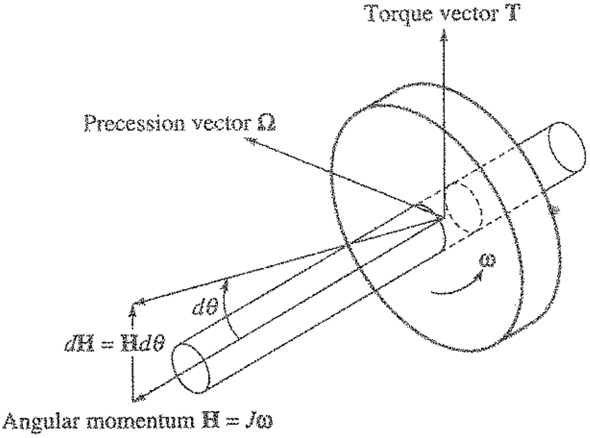


Figure 3.1 Principle of spinning rotor gyro.

Using Newton's law, the rate of change of angular momentum can be expressed as:

$$\mathbf{T} = \frac{d\mathbf{H}}{dt} \tag{3.1}$$

Since $\mathbf{H}d\theta = d\mathbf{H}$, substituting for $d\mathbf{H}$ in Eq. (3.1), we get:

$$\mathbf{T} = \frac{\mathbf{H}d\theta}{dt} \tag{3.2}$$

where $\Omega = (d\theta/dt)$ is the rate of precession of the rotor spin axis about an axis that is perpendicular to both $\mathbf{H}$ and $\mathbf{T}$. So, torque $\mathbf{T}$ is expressed as a vector cross product of inertial angular rate and angular momentum as follows:

$$\mathbf{T} = \mathbf{\Omega} \times \mathbf{H} \tag{3.3}$$

Equation (3.3) also implies that if an inertial angular rate $\mathbf{\Omega}$ acts on the rotor, a torque is produced that is perpendicular to both $\mathbf{H}$ and $\mathbf{\Omega}$ and is known as gyroscopic torque. Measurement of this torque by a suitable scheme instrumented within the gyro constitutes determination of the input inertial rate, as the magnitude of H is a known design constant.

The principle can be utilised to measure inertial angular rate of a vehicle, and can also be utilised to provide inertial reference as needed by inertial navigation. The vector cross product representation of gyroscopic Eq. (3.3) is normally converted to a scalar form using standard mathematical technique, and then Eq. (3.3) is written as:

$$T_o = \Omega_i H \tag{3.4}$$

where

T_o = torque about gyro output axis
Ω_i = inertial angular rate about gyro input axis (sensitive axis)
H = spin axis angular momentum

The input axis (IA), the spin axis (SA) and the output axis (OA) are mutually orthogonal and right handed in the sequence (IA, SA, OA), and this feature is shown in Figure 3.2.

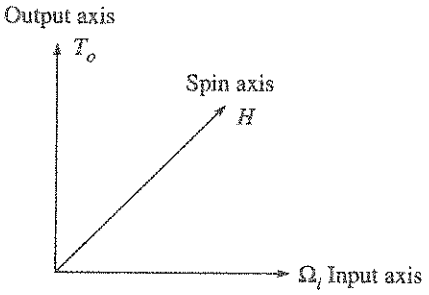


Figure 3.2 Spinning rotor gyro axes.

Precession

The unique phenomenon of precession of a spinning rotor is further explained.

Consider Figure 3.3 where a rotor is spun at a constant speed ω and supported lightly at the end of a horizontal shaft on a pivot. Question arises as to what happens to the rotor under this type of suspension.

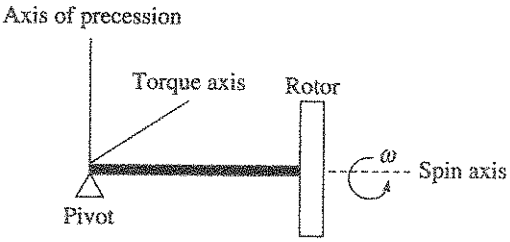


Figure 3.3 Example of precession.

Even though the rotor is pulled down by the gravitational force, it will not topple. On the contrary, the rotor will precess. It can be explained in the following manner.

The gravitational force will produce a torque, due to lever arm effect, at the pivot point whose axis will be perpendicular to the spin axis such that the torque axis will also be horizontal. This torque, being perpendicular to spin axis, will then act on the spinning rotor to cause precession of the rotor about an axis that will pass through the pivot point and the axis will be perpendicular to torque producing axis. This means the axis of precession will be vertical. So, the rotor will precess horizontally. Spin axis, torque axis and precession axis will form right handed system.

3.1.1 Torque Producing Wheel

If angular speed is not maintained constant in a spinning rotor, then a torque is produced as per Newton's law about the axis that is same as that of the momentum vector $\mathbf{H}$. This torque $\mathbf{T}_m$ is given by

$$\mathbf{T}_m = \frac{d\mathbf{H}}{dt} \tag{3.5}$$

This feature of a gyroscopic wheel, called reaction wheel, is often used in torquing a spacecraft in order to change or correct its orientation with respect to earth or sun. Normally these torque producing wheels are large in size. It is also possible to control a spacecraft using the combination of Eqs. (3.4) and (3.5). For such operation, the generalised form of vector torque equation is useful that is given by

$$\mathbf{T} = \left[\frac{d\mathbf{H}}{dt}\right] + \mathbf{\Omega} \times \mathbf{H} \tag{3.6}$$

But for use as a gyro, Eq. (3.3) is important as the speed and the rotor direction are maintained constant so that $[d\mathbf{H}/dt]$ is zero.

3.2 Types of Spinning Rotor Gyro

Spinning rotor gyros are classified into two types based on degree of rotational freedom of the spin axis:

1. Single degree of freedom gyro
2. Two degrees of freedom gyro

It is seen in Figure 3.4 that the rotor spin axis has freedom of rotation about one axis that is provided by the gimbal on which the rotor is suspended. This axis is known as the gyro output axis. Such a gyro, with suspension that provides freedom of rotation about one axis, is called single degree of freedom gyro.

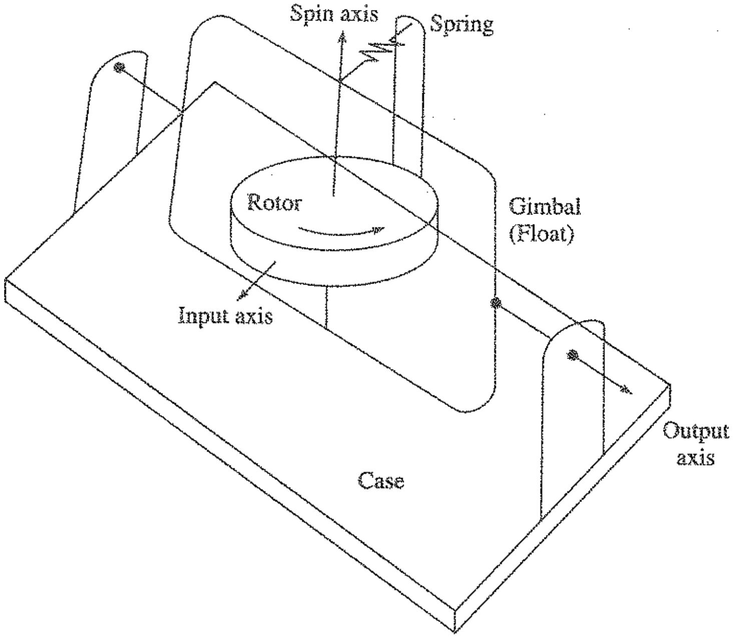


Figure 3.4 Schematic of a single degree of freedom gyro.

Figure 3.5 shows a spinning rotor gyro suspended with two gimbals which allow the spinning rotor to have rotational freedom about the remaining two perpendicular axes. In this configuration, the spin axis can remain pointing to inertial space even in the presence of rotation of the gyro about the two axes which are perpendicular to the spin vector. As a result, such gyros are also called free gyro or two degrees of freedom gyro. If we consider the spin vector as horizontal in Figure 3.5, then the two gimbals will provide rotational freedom about the vertical axis (Axis 1) and about the horizontal axis (Axis 2) that is perpendicular to the spin axis.

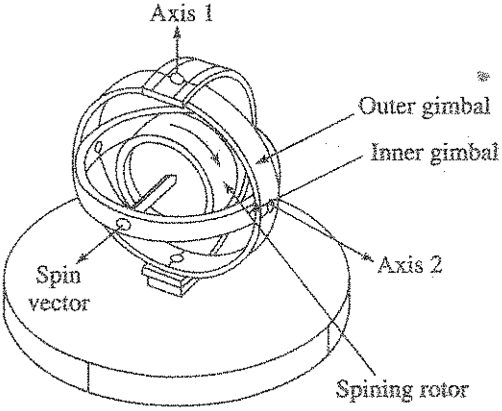


Figure 3.5 Two degrees of freedom gyro.

3.2.1 Nutation Frequency

Two degrees of freedom gyro exhibits nutation frequency in addition to the gyroscopic precession. A spinning top, with its two degrees of freedom of motion, is used to explain the phenomenon. A top is spinning with angular rate of magnitude Ω_z about the spin axis so that its angular momentum is $H_z = J_z\Omega_z$. We assume that the spinning top is symmetric about the spin axis, and this means that the transverse moment of inertias J_x and J_y are same and is J. The force of gravity mg with a lever arm distance L acts on the spinning top and produces a gyroscopic motion that can be termed as precession (Figure 3.6). The magnitude of the rate of precession will be given by T/H, where $T = mgL$. Normally, the motion is not a simple precession, but a combination of nutation along with precession, and this complex motion is shown in Figure 3.7. Here the higher frequency component moving in and out of the plane of precession constitutes the nutation frequency.

The motion of the top can be derived using the generalised form of vector torque Eq. (3.6). In the component form, the motion can be expressed as:

$$T_x = J\left[\frac{d\Omega_x}{dt}\right] + \Omega_y H_z \tag{3.7}$$

$$T_y = J\left[\frac{d\Omega_y}{dt}\right] - \Omega_x H_z \tag{3.8}$$

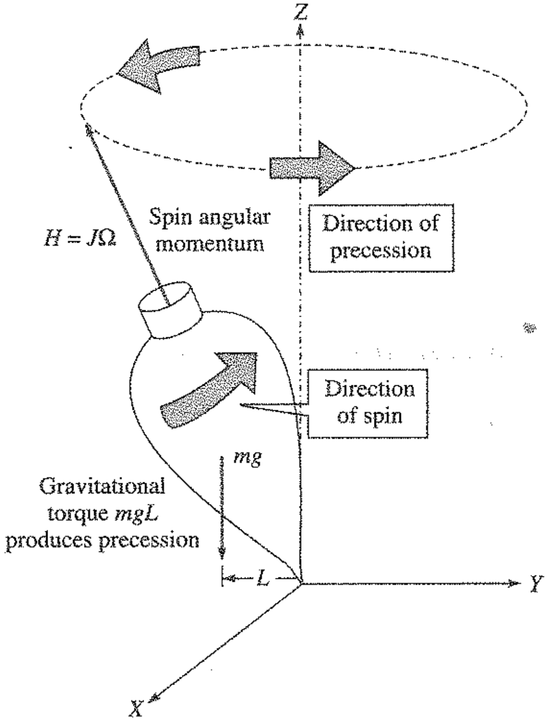


Figure 3.6 Precession of a spinning top.

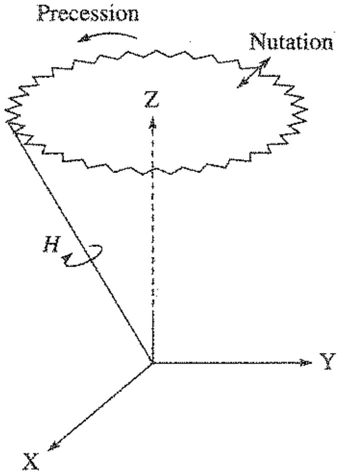


Figure 3.7 Nutation and precession of a spinning top.

The solutions of these pair of equations are given by

$$\Omega_x = A_1 \sin \omega_n t + A_2 \cos \omega_n t - \frac{T_y}{H_z} \tag{3.9}$$

$$\Omega_y = -A_1 \cos \omega_n t + A_2 \sin \omega_n t + \frac{T_x}{H_z} \tag{3.10}$$

where

$$\omega_n = \frac{H_z}{J} = \text{nutation frequency} \tag{3.11}$$

A_1 and A_2 = constants (found from the initial conditions)

The third terms in Eqs. (3.9) and (3.10) describe precessional motion. Understanding of nutation frequency is important for designing a two-axis gyro.

3.2.2 Functional Classification of Gyros

Design and application of gyros have given rise to certain functional classification as follows:

(a) Single degree of freedom rate gyro for sensing vehicle body angular rate

(b) Single degree of freedom rate integrating gyro for sensing vehicle body angular rotation and providing inertial reference as needed for inertial navigation

(c) Free gyros, vertical and directional for sensing vehicle direction with respect to local vertical and horizontal

(d) Two degrees of freedom dynamically tuned gyro for sensing angular rate and for providing inertial reference

There have been other types of spinning rotor gyros also and interested readers can refer the excellent book of **[Lawrence, 1998]**.

3.3 Single Degree of Freedom Rate Gyro

A single degree of freedom rate gyro essentially measures angular rate of an object with respect to inertial space. If the gyro is strapped to earth, then it will measure rotation rate of earth at the point with respect to inertial space. If the gyro is mounted on a table that can rotate relative to earth, then it will measure algebraic sum of the table rate and the earth rate. This characteristic of a gyro is very important. If we mount a tachometer on the table, it will only measure the table rate relative to earth. When the gyro is connected to the body of the vehicle, its output is often described with the term ω_{iB}^{B}, which is defined as angular body rate with respect to inertial space as measured in body co-ordinates.

3.3.1 Gyro Description

The rotor is driven electrically with hysteresis synchronous motor to establish the constant angular momentum. The rotor is suspended in a gimbal using low friction bearings and accurately defines the spin axis. The gimbal surrounds the rotor and the gimbal is suspended to the gyro case with torsional spring, often called torsion bar, on the right, while on the left side, the gimbal is suspended on a very low friction bearing. This suspension axis is called output axis, while the input axis is perpendicular to both spin and output axis. On the output axis, a signal generator, quite often called pick-off, is mounted for measuring the precessional motion of the gyro.

The schematic of single degree of freedom rate gyro is shown in Figure 3.8.

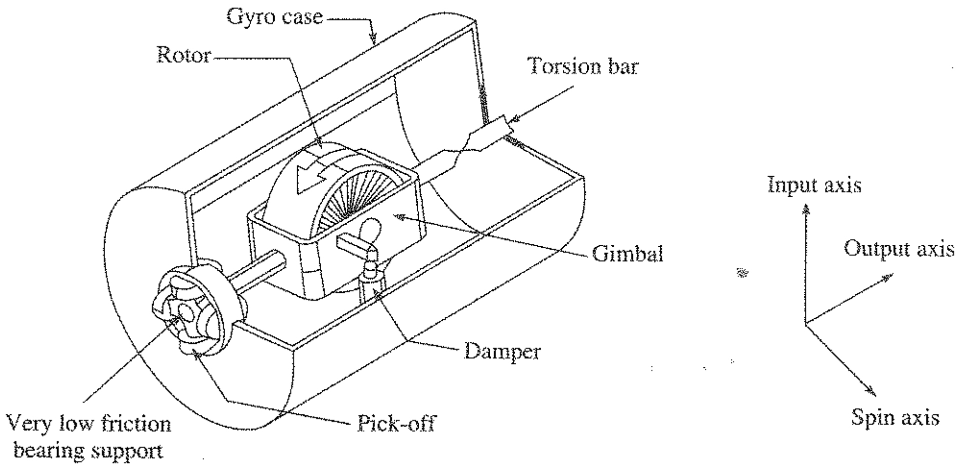


Figure 3.8 Single degree of freedom rate gyro schematic.

3.3.2 Gyro Sensitivity

The transfer function and the sensitivity of the gyro can be derived in the following manner. Consider

J = float inertia about the output axis
C = damping coefficient about the output axis
H = rotor angular momentum
K_p = pick-off scale factor
K = stiffness of the torsion bar

When an inertial angular rate Ω_i is applied to the gyro about its input axis (sensitive axis), the effect of the inertial angular rate results in the generation of gyroscopic torque T_o about the output axis. Under steady state, this precession torque is balanced by the torsional stiffness of the torsion bar about the output axis that introduces a deflection angle θ_o of the torsion bar. This is converted to an electrical output θ_e by the pick-off with gain K_p (refer Figure 3.9).

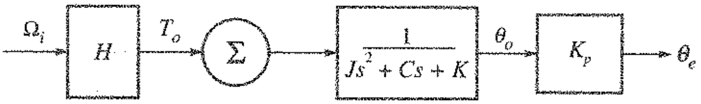


Figure 3.9 Rate gyro open loop block diagram.

The dynamic equation of motion can be represented in Laplace form as:

$$(Js^2 + Cs + K)\theta_o = H\Omega_i \tag{3.12}$$

Under steady state condition, we get:

$$\Omega_i = \frac{K}{H}\theta_o \tag{3.13}$$

Since K and H are design constants, measurement of θ_o by a pick-off with gain K_p thus provides the electrical output θ_e which is the measurement of Ω_i.

Quite often the natural frequency and damping ratio terms are important design parameters in rate gyro.

The undamped natural frequency ω_n is given by

$$\omega_n = \sqrt{\frac{K}{J}} \tag{3.14}$$

The damping ratio ξ is given by

$$\xi = \frac{C}{2\sqrt{KJ}} \tag{3.15}$$

3.3.3 Rate Gyro Features

Spin motor is typically a hysteresis synchronous motor running at a high speed of around 24,000 rpm. High speed is necessary to produce high angular momentum in a small inertia of the rotor. A synchronous motor is necessary to maintain a constant speed.

The torsion bar is typically made of a low hysteresis material such as Beryllium copper or Elgiloy. The pick-off should have high rotational sensitivity and microsyn type of detector has been successfully used. Damping to the gimbal is normally provided by some form of fluid that has low change in viscosity with temperature. Silicon oil has this property and is widely used. In some rate gyros, a low capacity torque generator is additionally provided to get a self-test feature and for response characteristic evaluation.

The critical performance parameters of a rate gyro are usually defined with the following parameters:

(a) Low bias
(b) Low bias shift with temperature
(c) Low hysteresis
(d) High linearity
(e) Suitable response in time and frequency domain

Bias is an error in a gyro as it manifests as an output even when input to the gyro is zero.

Rate gyro usage

A rate gyro is used for

- Measurement of inertial angular rate about a body fixed axis.
- Providing damping to control systems in aircrafts, launch vehicles, spacecrafts, missiles and ships.

3.4 Single Degree of Freedom Rate Integrating Gyro

Single degree of freedom Rate Integrating Gyro (RIG) was a prime mover in the development of inertial navigation system technology five decades ago. The principle of working of RIG has some similarity with rate gyro, but its technology is quite complex. This is because the RIG has to instrument the inertial reference frame that should ideally not rotate (drift) at all in an inertial navigation system.

3.4.1 Operating Principle

In a RIG, the rotor is electrically driven at a constant speed to establish a constant angular momentum. The rotor is supported on a gimbal using high precision ball bearings which define the spin axis. The gimbal, in turn, is then mounted on a pair of bearings with practically negligible friction to define what is known as the gyro output axis. On this axis, a signal generator, quite often called pick-off, is mounted for measuring the precessional motion of the gyro. These features are shown in Figure 3.10.

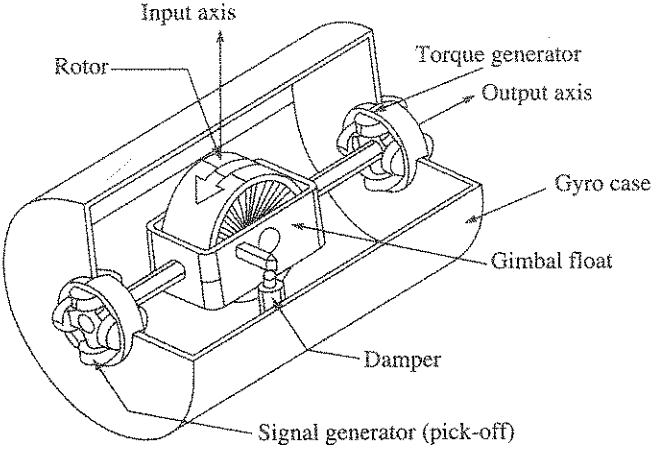


Figure 3.10 Single degree of freedom rate integrating gyro schematic.

RIG transfer function is explained with the gyro loop block diagram shown in Figure 3.11 where

J = gimbal float inertia about the output axis
C = gimbal float damping coefficient about the output axis
H = rotor angular momentum
K_p = pick-off scale factor

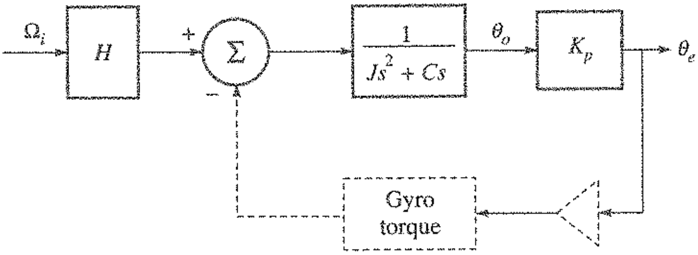


Figure 3.11 RIG block diagram.

Notice the absence of stiffness term K in the gyro model. When an inertial angular rate Ω_i is applied to the gyro about its input axis, the effect of the gyroscopic torque is to cause a rotation θ_o of the gimbal float about the output axis.

The Laplace equation describing the output motion can be written as:

$$(Js + C)\dot{\theta}_o(s) = H\Omega_i(s) \tag{3.16}$$

So

$$\dot{\theta}_o(s) = \frac{H}{(Js + C)}\Omega_i(s) \tag{3.17}$$

For a step angular rate input,

$$\Omega_i(s) = \frac{\Omega_i}{s} \tag{3.18}$$

Time domain solution of Eq. (3.17) becomes:

$$\dot{\theta}_o(t) = \frac{H}{C}\Omega_i[1 - e^{-(C/J)t}] \tag{3.19}$$

For a time approaching infinity, Eq. (3.19) reduces to

$$\theta_o(t) = \frac{H}{C}\int \Omega_i dt \tag{3.20}$$

Hence it is seen that when the gyro is mechanised with no output axis restraining torque ($K = 0$), the output is proportional to the integral of the applied angular rate and hence the name rate integrating gyro. Since gyro output grows with time due to integration of the input angular rate, it is necessary that the magnitude of output angle θ_o is restricted. From various design considerations, typically, it is kept well below ±1 degree. This is realised at the sensor level by closing the loop through the gyro torque generator, as shown by the dotted line in Figure 3.11, so that the rotation is nulled. Such a gyro with the loop when closed as shown by dotted line, is called a rate gyro, a feature that is needed in a strapdown system.

Since RIG was primarily used for providing inertial reference, the gyro drift needed to be made as low as possible and was a formidable technological challenge as there were several mechanisms to cause significant amount of drift in a gyro. For example, if the friction torque on the gyro float suspension bearings is not completely eliminated, then the residual effect can be calculated in the following manner.

EXAMPLE 3.1 (on RIG output axis friction effect: For a typical RIG, assume,

Output axis bearing friction torque value = 1×10^{-8} kg-m, $H = 1 \times 10^{-2}$ kg-m^2/s, gravitational acceleration $g = 9.80$ m/s^2. Compute the gyro drift.

Solution: The gyro drift is calculated using the gyroscopic Eq. (3.4). Putting the values and solving, we get:

$$\Omega_i = \frac{1 \times 10^{-8} \times 9.80}{1 \times 10^{-2}} \text{ rad/s}$$

Since, 1 rad/s = 57.4 × 3600°/h

In the more familiar drift unit, $\Omega_i = 2.02$ °/h

The error is interpreted by saying that only when the inertial input axis angular rate exceeds 2.02°/h, the gyro will be able to detect the rotation. So, even such an insignificant amount of friction torque is not acceptable for autonomous inertial navigation.

3.5 Two-axis Gyro

Two-axis gyro feature, shown in Figure 3.5, was not good enough to provide inertial reference due to friction in the gimbal bearings, unwanted torque produced by the flex leads feeding electrical power to the motor and gimbal unbalance. All these torques gave rise to high drift, and further they were not stable also. As a result such gyros could not be used for maintaining inertial reference in a navigation system, but were primarily used for providing vehicle attitude with respect to vertical or horizontal where larger drift could be tolerated. Alternative technologies were explored and continued research for navigation application led to the emergence of a gyro that is known as Dynamically Tuned Gyro (DTG).

3.6 Dynamically Tuned Gyro

In the year 1963, Howe of American Bosch Arma received a patent of a gyro, which used dynamical inertial effect to cancel flexure-generated torque. The flexure spring stiffness is independent of the rotor spin magnitude, but the negative dynamic inertial spring stiffness depends on the rotor spin rate, and this enables mutual cancellation of the two at a particular speed called tuned speed. The gyro, which became a reality with this concept, was named *dynamically tuned gyro*.

3.6.1 Operating Principle

The operating principle of a DTG is explained with Figure 3.12. The figure shows a rotor connected to a shaft driven by an electric synchronous motor through a single gimbal flexure system. There is a pair of flexures connecting the shaft with the gimbal and another flexure pair connecting the gimbal with the rotor but orthogonal to the first pair. Observe that the rotor is outside the gimbal which is another significant difference to this configuration.

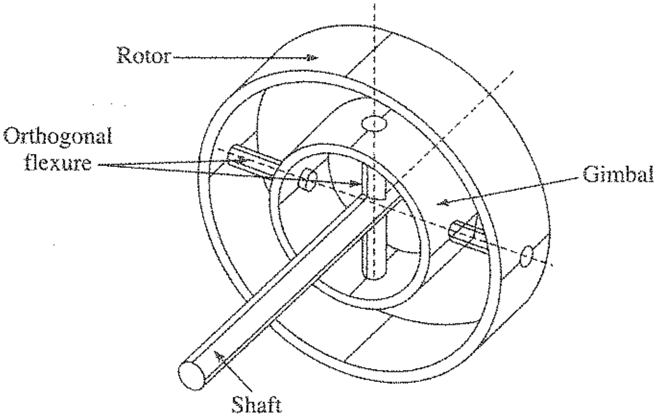


Figure 3.12 Dynamical parameters in a DTG rotor.

The open loop motions are given by the following two equations which are adapted from [**Howe and Savet, 1964**]:

$$\left(A+\frac{A_g}{2}\right)\ddot{\theta}_x+(C+A_g)N\dot{\theta}_y+[K_x-(A_g-C_g/2)N^2]\theta_x=0 \tag{3.21}$$

$$\left(B+\frac{B_g}{2}\right)\ddot{\theta}_y-(C+B_g)N\dot{\theta}_x+[K_y-(B_g-C_g/2)N^2]\theta_y=0 \tag{3.22}$$

where

A, B = transverse inertias of the rotor
C = polar inertia of the rotor
A_g, B_g = transverse inertias of the gimbal
C_g = polar inertia of the gimbal
K_x, K_y = torsional stiffnesses of the flexures
θ_x, θ_y = rotor deflections with respect to the shaft, about X and Y axes respectively
N = spin speed of the shaft

There are certain higher order terms, which consist of viscous damping terms, terms involving cos $2Nt$, sin $2Nt$ and cross axis rotor shaft torques. Initially, the effect of these torques on the rotor is neglected and not considered in Eqs. (3.21) and (3.22). The third term in both the equations, consists of a positive spring torque, K_x or K_y and a dynamic negative spring term, $-(A_g - C_g/2)\ N^2$ or $-(B_g - C_g/2)N^2$. A free gyro condition can be realised if the rotor is decoupled from the shaft. This is achieved by cancelling the positive spring torque with the dynamic negative spring torque. There is a speed N_0 when the positive flexure stiffness equals the dynamic spring stiffness. The equality conditions are given by

$$K_x=\frac{A_g-C_g}{2}N_0^2 \tag{3.23}$$

$$K_y=\frac{B_g-C_g}{2}N_0^2 \tag{3.24}$$

Solving for the tuned speed N_0, we get

$$N_0^2=\frac{K_x+K_y}{A_g+B_g-C_g} \tag{3.25}$$

Normally, by design, $K_x = K_y = K$ and $A_g = B_g$. So Eq. (3.25) reduces to:

$$N_0^2=\frac{K}{A_g-C_g/2} \tag{3.26}$$

This tuning condition in a DTG is shown graphically in Figure 3.13.

Under tuned condition, DTG equation of motion under open loop condition becomes:

$$\frac{A+A_g}{2}\ddot{\theta}_x+(C+A_g)N_0\dot{\theta}_y=0 \tag{3.27}$$

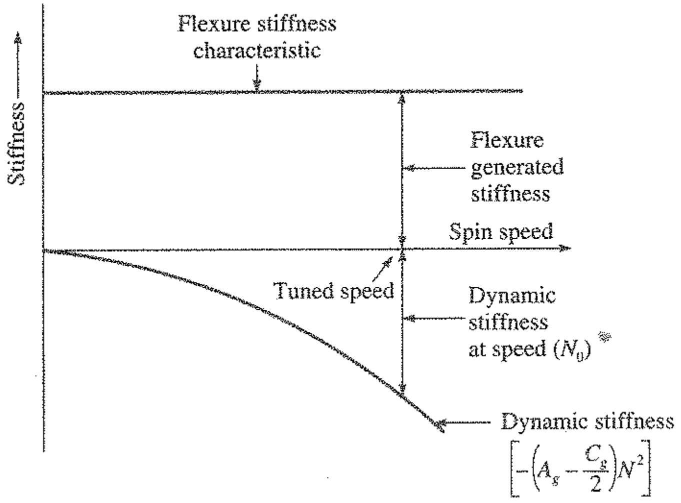

Flexure stiffness characteristic
Stiffness
Flexure generated stiffness
Spin speed
Tuned speed
Dynamic stiffness at speed (N_0)
Dynamic stiffness $\left[-\left(A_g - \frac{C_g}{2}\right)N^2\right]$

Figure 3.13 Tuning condition in a DTG.

$$\frac{B + B_g}{2}\ddot{\theta}_y - (C + A_g)N_0\dot{\theta}_x = 0 \tag{3.28}$$

when $A = B$ and $C >> A_g$, the equations are further reduced to:

$$J\ddot{\theta}_x + H\dot{\theta}_y = 0 \tag{3.29}$$

$$J\ddot{\theta}_y - H\dot{\theta}_x = 0 \tag{3.30}$$

where

J = rotor transverse inertia

H = angular momentum of the rotor = CN_0

It will be noted that Eqs. (3.21) and (3.22) have neglected damping and some other terms. Considering these factors also, three important design conditions for a DTG to make it an ideal free gyro are:

(a) Spin speed should be equal to the tuned speed N_0

(b) Damping torque on the rotor should be zero

(c) Torque at twice the spin frequency ($2N$), should be zero

The tuned speed N_0 should be precisely controlled to achieve the free gyro condition. Operation away from tuned speed leads to bias torque, which in turn produces bias drift.

Damping torque on the rotor arises from flexure hysteresis, gas drag on the rotor and the leakage of the magnetic field of the torquer magnets.

From the design point of view, DTG figure of merit F_m is given by **[Craig, 1972]**

$$F_m = \frac{C}{A_g + B_g - C} \tag{3.31}$$

Maximisation of F_m needs minimisation of the denominator. Since the value of C is decided to achieve certain angular momentum H, maximisation of F_m is normally achieved by proportioning the gimbal inertias and by making the gimbal small. Both, high time constant as well as high figure of merit, lead to high performance DTG with stable drifts and constitute important aspects of the gyro design. For a navigation grade DTG, F_m typically lies between 75 and 200, while time constant is between 50 s and 100 s.

In an idealised free gyro DTG, torque on the rotor is always zero, and the vehicle rotation angle with respect to inertial space, say ϕ_x, can be directly read by the DTG pick-off (Figure 3.14).

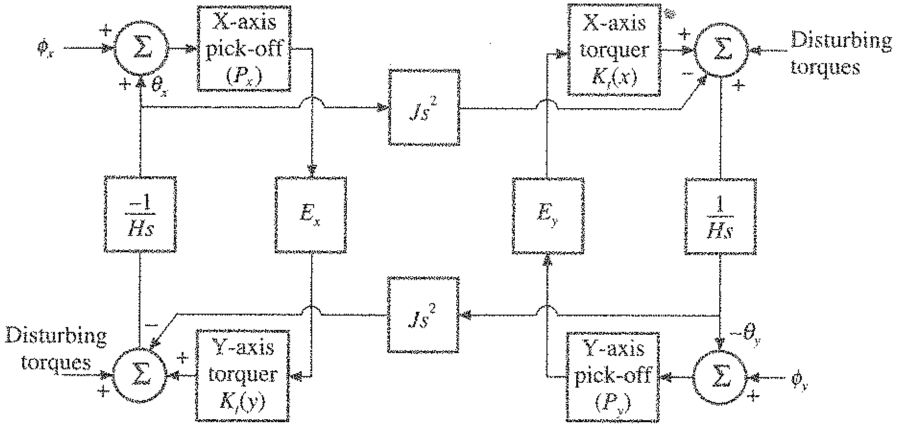


Figure 3.14 DTG closed loop block diagram.

However, Eq. (3.30) shows that the open loop gyro response is oscillatory in nature due to the presence of nutation frequency requiring appropriate closed loop stabilisation. As a result, the pick-off output is always nulled. This rebalance operation is depicted in Figure 3.14 where the rebalance torque generated by the gyro torquer counterbalances the vehicle rate induced gyroscopic torque. Thus the X-axis pickoff output feeds to Y-axis torquer to counterbalance inertial angular rate about X-axis and vice versa. A measurement of gyro torquer rebalance current can then be interpreted to know the vehicle rotation rate. Thus a DTG becomes a rate gyro in an actual operation and not a free gyro.

3.6.2 Descriptive Features

Typical features of a DTG are brought out in Figure 3.15, while its typical engineering mechanisation is shown in Figure 3.16. The spin motor is normally a hysteresis synchronous motor, which is designed to run at a speed that gives optimal drift performance as well as life. Even though, the motor is a synchronous motor, its speed is changed to approach the tuning speed by changing the frequency of the motor excitation voltage. After tuning is realised, the frequency is locked and the motor runs at a constant speed called the tuned rotor speed. Residual small amount of mistuning leads to bias drift which is calibrated.

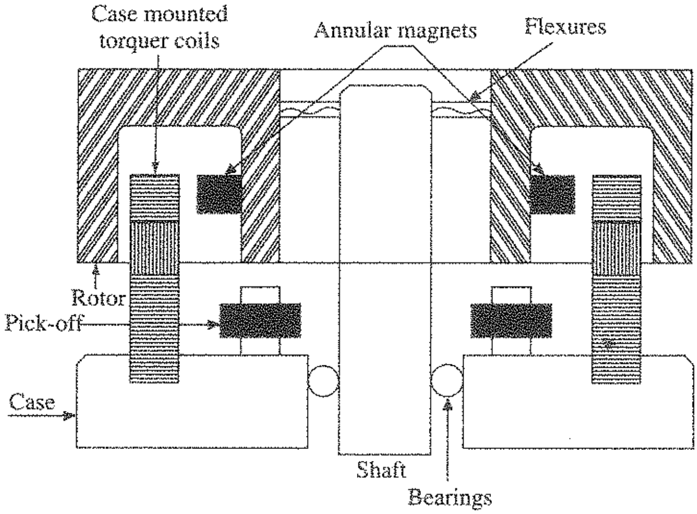


Figure 3.15 Features of a DTG.

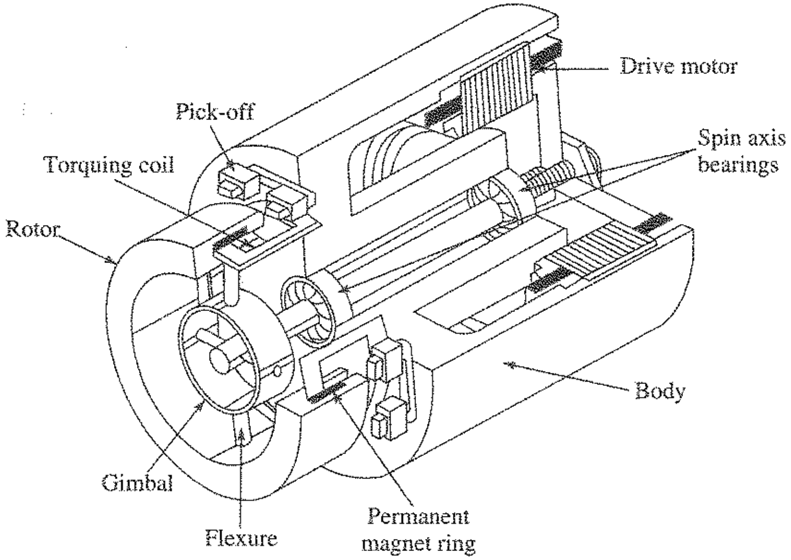


Figure 3.16 Engineering view of a cutaway DTG.

There are two pick-offs to measure case rotation angles relative to the inertially stable rotor axes. Each pick-off consists of a primary coil and a secondary coil. Two such pick-offs are positioned 180 degrees apart for X-X axis and similar pick-off arrangement for Y-Y axis (refer Figure 3.17). The outputs of these paired pick-offs, are differentially connected. In the presence of a case rotation, air-gap reduces for one pick-off, while it increases for the other pick-off, producing a net output from the paired pick-offs proportional to the rotation. The advantage of a differential connection is that it reduces certain common mode pick-off sensitivities such as temperature. Typical frequency of excitation for the pick-off is >15 kHz, while the voltage is 4–6 V.

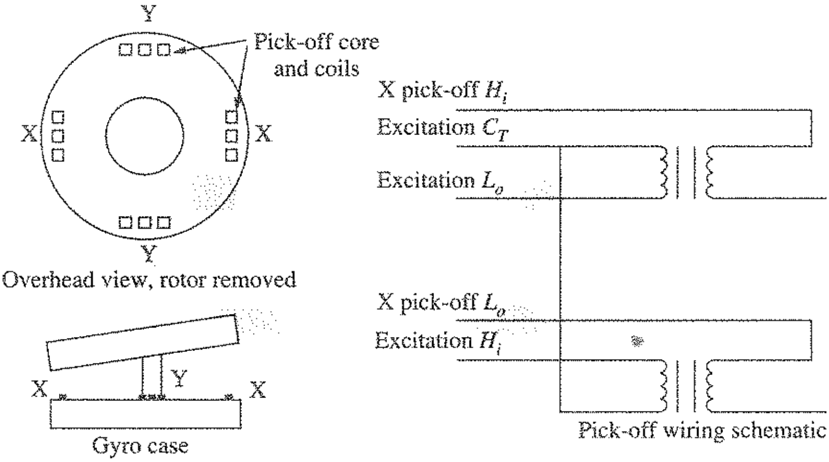


Figure 3.17 DTG pick-offs.

Torquers

In a strapdown system, torquers are needed for rebalance operation of a DTG. As the angular rate capability needed is high, powerful torquers are needed. The torquer magnets are fixed on the rotor and forcer coils on the stator case as shown in Figure 3.15. To generate torques around the defined gyro axes, X-axis and Y-axis, four forcer coils are used each spanning nearly one quadrant with the opposite pairs connected so that their torques are additive. The torque T, which is required to be produced in each axis to counterbalance the input angular rate Ω, is given by the gyroscopic Eq. (3.3).

Typical magnets currently used are the Rare Earth Samarium Cobalt magnets, which provide very high-energy product (product of flux density and magnet field intensity) and excellent linearity in the operating second quadrant of the magnet.

EXAMPLE 3.2 (on DTG rebalance torque): A DTG with angular momentum of 1×10^{-2} kg-m^2/s experiences a rate of 100°/s at an angle of 30° to X-sensitive axis in the X-Y plane. Find the rebalance torque to be provided by the torquers. If the gyro torque constant is 500°/s/A, find the rebalance currents required. Assume value of gravitational acceleration, $g = 9.80$ m/s^2

Solution: The resolved angular rates are:

About X-axis: 100 cos 30 = 86.6°/s; About Y-axis: 100 sin 30 = 50°/s

The Y-axis torquer will be activated to counterbalance the X-axis angular rate. Its magnitude is calculated from gyroscope equation as:

$$T_y = \Omega_x H$$

Putting the values and solving, we get:

$$T_y = \frac{86.6 \times 1 \times 10^{-2}}{57.4 \times 4 \times 9.80} \text{ kg-m} = 1.50 \times 10^{-3} \text{ kg-m}$$

$$\text{Y-axis rebalance current} = \frac{86.6}{500} \text{ A} = 0.173 \text{ A}$$

Now, the X-axis torquer will be activated to counterbalance the Y-axis angular rate. We write

$$T_x = \Omega_y H$$

Putting the values and solving, we get:

$$T_x = \frac{50 \times 1 \times 10^{-2}}{57.4 \times 9.80}\ \text{kg-m} = 0.86 \times 10^{-3}\ \text{kg-m}$$

X-axis rebalance current $= \frac{50}{500}\ \text{A} = 0.10\ \text{A}$

Gimbal flexures

DTG gimbal flexures play the most important role in gyro operation to provide its inertial grade performance. Materials with low hysteresis and high strength, such as maraging steel, are typical choice for gimbal flexures. Operating stress level in the flexure is to be checked against mechanical shock load and ensure adequate margin. Limit stops are used to restrict rotor angular excursions on either side, and thus protect the flexures against over stressing.

3.6.3 Errors and Their Model Representation

There are several error sources which contribute to drift in DTG. Some of these drift errors are discussed as follows:

Bias drift due to mistuning and damping

Mistuning, which is DTG operation away from the tuned speed, results in bias drift. It can happen if the flexure stiffness changes due to change in DTG temperature as most flexure materials exhibit change in elastic constants under varying temperature. This change will be more pronounced when DTG temperature is not controlled. At system level, the rotor speed of two gyros are set slightly away from each other to reduce $1N$ axial vibration effect, and this often leads to mistuned operation of at least one gyro. $1N$ axial vibration refers to an applied vibration whose frequency is same as the first nutation frequency of the gyro. In the case of damping, the rotor damping torque produces bias drift. As a result, DTG rotor is operated under partial vacuum.

Spin axis mass unbalance drift

It is caused when the rotor centre of mass is shifted along the spin axis away from the flexure-defined plane. Acceleration acting perpendicular to spin axis then causes drift due to this unbalance.

Quadrature mass unbalance drift

It occurs due to certain imperfection in a flexure or in flexures. This happens when a defective flexure, axially loaded with acceleration, twists. This twisting then results in a drift.

Thermal sensitivity

DTG thermal sensitivity manifests both in scale factor as well as in drift. In scale factor, this is due to change in magnet gap flux density, such as samarium cobalt, with temperature. For a typical samarium cobalt magnet ($SmCO_5$), the scale factor change is around 480 ppm/°C. A possible solution is to use magnetic shunt made of temperature compensating nickel–iron alloy. Alternately, very low temperature sensitivity samarium cobalt magnets are now available for use.

Drift change with temperature is addressed either by controlling DTG temperature or by a thermal model which can be of first or second order.

Thus it is clear that DTG output consists of a combination of true and erroneous information. By understanding the errors as discussed earlier, it has been possible to model these error terms along with the true output. Such a model is shown in Eq. (3.32). As DTG is a two-axis gyro, there are two equations to represent the gyro model. The model shows that the input axes (sensitive axes) are defined by X-axis and Y-axis respectively and about which inertial angular rates Ω_x and Ω_y are acting.

$$\begin{aligned} S_xK_x &= [\Omega_x + B_x + D_{xx}a_x - D_{xy}a_y] \\ S_yK_y &= [\Omega_y + B_y + D_{yy}a_y - D_{yx}a_x] \end{aligned} \tag{3.32}$$

where

S_x, S_y = output measured in milli-ampere (mA), for X-axis and Y-axis respectively
K_x, K_y = gyro torquer scale factor in °/h/mA for X-axis and Y-axis respectively
Ω_x, Ω_y = input angular rate about the reference X- and Y-axes respectively in °/h
a_x, a_y = acceleration acting along X-axis and Y-axis respectively in units of g
B_x, B_y = bias drift in °/h for X-axis and Y-axis respectively
D_{xx}, D_{yy} = spin axis unbalance in °/h/g for X-axis and Y-axis respectively
D_{xy}, D_{yx} = quadrature axis unbalance in °/h/g for X-axis and Y-axis respectively

Now, various unknown coefficients in the model are found out using series of tests and all these require a well instrumented test set up. This is called calibration. The model shown is a truncated version and depending on application, the model becomes larger with more unknown parameters. Readers can refer to the book of **[Bose et al., 2008]** to learn more about DTG model and calibration.

3.7 Coriolis Vibratory Gyros

The vibrating gyros which are based on the Coriolis force are termed as Coriolis Vibratory Gyros (CVG). These gyros are termed as solid state gyros as no rotating parts and/or bearings are involved. Such gyros have considerably reduced number of parts in comparison with the classical or the other forms of solid state gyros, and some of these vibratory gyros are amenable to fabrication by micromachining technology. All these factors put together indicate that CVGs in general exhibit considerable improvement in reliability, low power consumption and lower mass. These gyros are being developed to measure vehicle body rate as well as to provide inertial reference in a navigation system for application spanning land, air and space usages.

3.7.1 Coriolis Force

Coriolis force manifests on a moving mass in a rotating frame of reference. It can be described in the following manner. Consider a ball of mass M is moving outward with a velocity V from the centre of a table which rotates at angular rate Ω. When the motion of the ball is viewed from an inertial frame of reference, it shows a straight line path (Figure 3.18).

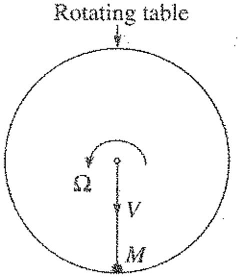


Figure 3.18 Motion of the ball viewed from inertial frame.

When the motion of the ball is observed from the rotating table, it shows a curved trajectory (Figure 3.19).

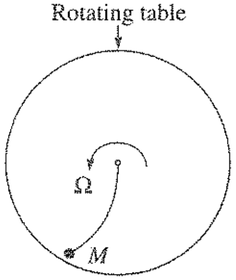


Figure 3.19 Motion of the ball viewed from the rotating table frame.

This apparent shift of the trajectory in a rotating frame of reference is caused by a force called Coriolis force $\mathbf{F}_c$ which in vector notation, is expressed as:

$$\mathbf{F}_c = 2M(\mathbf{V}_r \times \mathbf{\Omega}) \tag{3.33}$$

where

$\mathbf{V}_r$ = velocity vector relative to the rotating frame

$\mathbf{\Omega}$ = angular rate vector perpendicular to velocity

The relationship reveals that the Coriolis force is perpendicular to velocity vector $\mathbf{V}_r$ and the angular rate vector $\mathbf{\Omega}$. In scalar form, the magnitude of the force can be expressed as:

$$F_c = 2MV_r\Omega \sin \theta$$

where θ is the angle between the two vectors. This feature of Coriolis acceleration shows that if the velocity is parallel to the rotation axis, the Coriolis acceleration is zero.

The phenomena of Coriolis force is widespread in natural science to explain tidal wave behaviour, deflection of winds moving along the surface of the earth to the right of the direction of travel in the Northern hemisphere and to the left of the direction of travel in the Southern hemisphere. In the year 1852, French scientist, Leon Foucault demonstrated the rotation of earth using an instrument that is known as Foucault's pendulum. The plane of the pendulum orbit changes due to the earth rate acting perpendicularly on the pendulum velocity. This is also an example of Coriolis force.

3.7.2 Coriolis Vibratory Gyro Concept

In this section, it will be explained how the Coriolis force can be utilised to realise a conceptual gyro. Figure 3.20 shows a resonating proof mass that is supported on a frame by means of four springs. The velocity direction is +ve when the proof mass moves away from the table centre of rotation. The structure is kept on a table which is then rotated with angular rate magnitude Ω in the counter clockwise direction. The rate axis is perpendicular to velocity and the +ve axis direction is looking out of the paper.

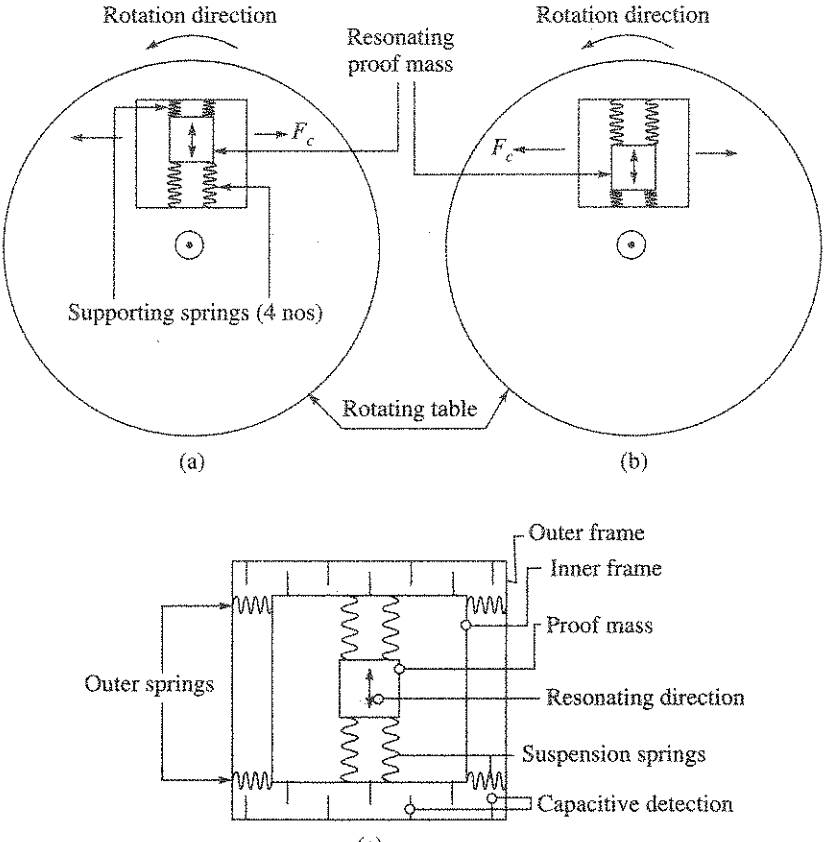


Figure 3.20 Coriolis vibratory gyro concept.

During the first half cycle of proof mass velocity V, the upper two springs are compressed as the proof mass moves away from the centre [refer Figure 3.20(a)]. As per right hand rule, the Coriolis force F_c is to the right and the corresponding reaction force on the frame is towards the left. In the next half cycle [refer Figure 3.20(b)], when the proof mass moves towards the centre, the Coriolis force F_c is to the left and the corresponding reaction force on the frame is towards the right. As per right hand rule, F_c lies on the plane of the paper. The reaction force can then be measured by symmetrically suspending this frame onto an outer frame by means of springs with equal stiffness and the resultant compression and expansion of the springs can be detected by suitable electrical sensors such as capacitors. If the outer springs have stiffness K, then the displacement resulting from the reaction force can be written as:

$$2\frac{\Omega VM}{K} \tag{3.34}$$

By measuring the displacement with capacitors, the inertial rate Ω can be measured as other parameters are known design constants. It should be appreciated that V is sinusoidaly changing its amplitude, the displacement will also be oscillating requiring suitable processing to extract the rate signal. The conceptual gyro schematic is shown in Figure 3.20(c).

The above demonstration of a Coriolis gyro can be utilised to bring out a more general working principle of a Coriolis gyro. In a Coriolis vibratory gyro, one of the resonant modes of an elastic body is excited to designed amplitude. When the device rotates about a particular body fixed axis, the resulting Coriolis force acting on the vibrating mass element of the structure excite a second resonant mode. The magnitude of vibration in this second mode is directly proportional to the applied inertial angular rate.

EXAMPLE 3.3 (on Coriolis gyro): A proof mass of 0.1×10^{-3} kg is made to resonate with a peak amplitude of 0.01×10^{-3} m at a frequency of 2 kHz. Find the Coriolis force when the mass is rotated at 57.4°/s

(i) Perpendicular to the axis of oscillation

(ii) Parallel to the axis of oscillation

Solution:

(i) Coriolis force $F_c = 2MV\ \Omega \sin\ \theta$; where $\theta = 90°$, $V_{pk} = 2 \times 3.14 \times 2.10^3 \times 0.01 \times 10^{-3}$ m/s and $\Omega = 57.4/57.4 = 1$ rad/s.

Putting the values and solving, we get:

$$F_c(pk) = 2 \times 0.1 \times 10^{-3} \times 2 \times 3.14 \times 2.10^3 \times 0.01 \times 10^{-3} \times 1 \times 1 \text{ N} = 25.1 \times 10^{-6}\text{N}$$

(ii) $\theta = 0$; So $F_c = 0$

This example highlights that the Coriolis force is quite small even when the rate is as high as 57.4°/s. So, detection of rates of the order of 0.1°/h is a serious technological issue for such type of gyros.

3.7.3 Coriolis Vibratory Gyro Structures

Currently the configurations of large number of *Coriolis vibratory gyro structures* have evolved, and these configurations can be categorised under the following:

- Beam structures, e.g. prismatic, triangular
- Tuning fork structures, e.g. single tine, dual tines
- Shell structures, e.g. hemispherical, ring, cylindrical

Some of these structures are shown in Figure 3.21.

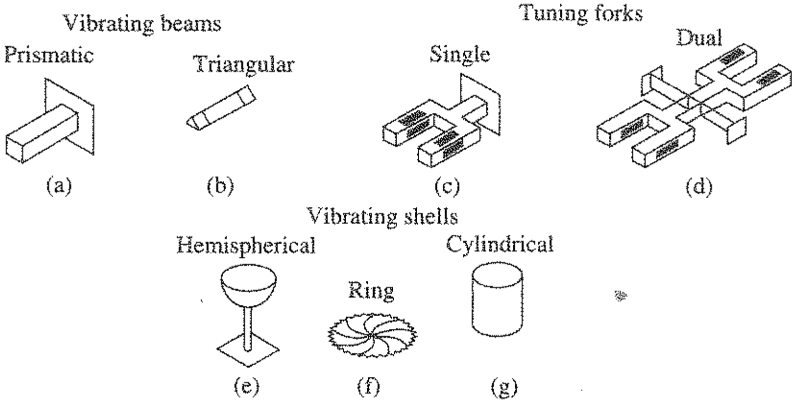


Figure 3.21 Coriolis vibrating gyro structural configurations.

These structures fall into two groups depending on the nature of the vibration modes involved. In the first group, the driven mode and the sensed mode are different. The single as well as dual tine tuning fork comes under this category. In the second group, the two modes are identical. The axis symmetric shell structures comprising hemispherical, ring and cylindrical come under this category. The second group, where the modes are identical, has demonstrated improved gyro performance, and within this group, the hemispherical configuration has demonstrated navigation grade gyro performance.

3.7.4 Tuning Fork Structure Gyro

Tuning fork structure gyro is a gyro which has two operational modes. Its operation is explained with Figure 3.22. We fix orthogonal body co-ordinates X, Y, Z on the device with centre at mass centre of the two tines. In this way, vector cross product equation can be reduced mathematically to a scalar form, and we can identify the sensor axes as follows:

Z – Input axis

Y – Drive axis

X – Sense or output axis

The tuning fork structure has two identical limbs which are called tine 1 and 2. The two tines are excited to their resonance along Y-axis lying on the plane of the paper. The method of excitation can be either piezoelectric or electrostatic. In Figure 3.22, the black colour drive and sense pattern deposited on the tines indicate piezoelectric means and the tine material is piezoelectric quartz. In the excitation scheme, tine 1 is made to vibrate in phase opposition to tine 2, which means that the directions of tine velocities are opposite to each other. This is to ensure that the structure has a vibration node at the mass centre of the tines. The first mode frequency ω_y, also called driven mode, is given by

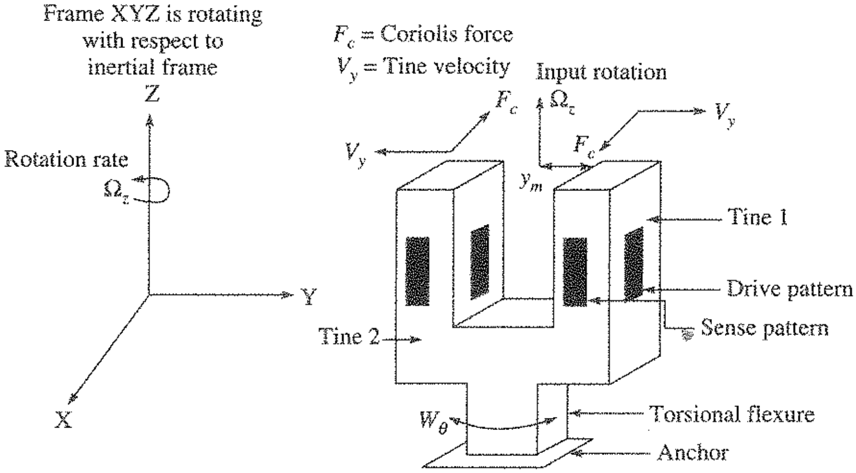


Figure 3.22 Tuning fork structure Coriolis vibratory gyro operation.

$$\omega_y = \sqrt{\frac{k_y}{M}} \tag{3.35}$$

where

k_y = tine stiffness along Y direction

M = mass of the tine

When both the tines vibrate, the scheme results in sinusoidal tine velocities V_y in phase opposition, but equal in magnitude. The sensitive axis for the gyro structure is along Z-axis lying on the plane of the paper. When an inertial angular rate of magnitude is Ω_z applied about Z-axis, anti-parallel Coriolis forces F_c, of equal magnitude and perpendicular to the plane of the paper, are generated as shown in Figure 3.22. The Coriolis force acting on each tine is given by

$$F_c = 2MV_y\Omega_x \tag{3.36}$$

Since the tine velocity is oscillating, the Coriolis force is also oscillating but in phase opposition. The forces, being away from the centre of rotation, will form a couple which in turn will result in a torsional frequency ω_θ which is the second mode or the sensed mode frequency. The torsional frequency is given by

$$\omega_\theta = \sqrt{\frac{k_\theta}{I_z}} \tag{3.37}$$

where

I_z = device moment of inertia about Z-axis

k_θ = torsional stiffness of the flexure

The magnitude of the torsional rotation is directly proportional to the applied input rate. Sensing this torsional rotation using suitable detection means enable detection of the applied input rate. Thus we see that the sensed mode frequency is different from the driven mode frequency for this type of structure, and effectiveness of the structure as a gyro depends on the sensitivity.

Tuning fork gyro sensitivity

The sensitivity of the tuning fork gyro, under certain simplified assumptions, can be worked out from the drive axis motion and the sensed mode motion.

Assume that the tuning fork tine structure is elastic about Y-axis only, which is the drive axis, while the tine is quite stiff about the remaining axes and that the mass M is concentrated at the top of the tine. Also, the elastic tine deflection is quite small. The force-deflection equation can be written as:

$$M\ddot{y} + C_y\dot{y} + k_y y = F_d = F_0 \cos(\omega_d t) \tag{3.38}$$

where

$F_d = F_0 \cos(\omega_d t)$ = sinusoidal tine drive force with drive frequency ω_d

C_y = damping constant associated with internal damping of the tine as well as surrounding air or gas

y = elastic tine deflection

Equation (3.38) represents a mechanical oscillator excited by electrical force. The equation of motion can be written as:

$$y(t) = \frac{F_0}{M\sqrt{(\omega_y^2 - \omega_d^2) + 4\omega_d^2\beta^2}} \cos(\omega_d t - \alpha) \tag{3.39}$$

where

$$\beta = \frac{C_y}{2M}$$

$$\alpha = \tan^{-1}\left[\frac{2\omega_d\beta}{(\omega_y^2 - \omega_d^2)}\right] \tag{3.40}$$

when an inertial angular rate Ω_z is applied to this structure along Z-axis, Coriolis force on one tine is given by

$$F_c(x) = 2MV_y\Omega_z \tag{3.41}$$

The Coriolis force on the other tine is anti-parallel, and these two forces form a couple with radius arm of y_m and the net torque T is given by

$$T = 4MV_y\Omega_z y_m \tag{3.42}$$

where V_y can be obtained from Eq. (3.39) by differentiation as follows:

$$V_y = \dot{y}(t) = -\frac{F_0\omega_d}{M\sqrt{(\omega_y^2 - \omega_d^2) + 4\omega_d^2\beta^2}} \sin(\omega_d t - \alpha) \tag{3.43}$$

Substituting V_y in Eq. (3.42) gives:

$$T = -\frac{4F_0\omega_d\Omega_z y_m}{\sqrt{(\omega_y^2 - \omega_d^2) + 4\omega_d^2\beta^2}} \sin(\omega_d t - \alpha) \tag{3.44}$$

The oscillating torque causes rotational motion of the flexure support at the base, and assuming no damping torque, the torque balance equation is:

$$T = I_z\ddot{\theta} + k_\theta \theta \tag{3.45}$$

where

I_z = device moment of inertia about Z-axis
θ = torsional rotation magnitude
k_θ = torsional stiffness of the flexure

The torsional frequency of this suspension is:

$$\omega_\theta = \sqrt{\frac{k_\theta}{I_z}} \tag{3.46}$$

Combining Eqs. (3.43), (3.44) and (3.45) and rearranging gives solution of θ as:

$$\theta(t) = \frac{A\Omega_z}{(\omega_\theta^2 - \omega_y^2)} \sin(\omega_d t - \phi) \tag{3.47}$$

where

$$A = -\frac{4F_0\omega_d^2 y_m}{\sqrt{(\omega_y^2 - \omega_d^2) + 4\omega_d^2\beta^2}} \tag{3.48}$$

Therefore, sensitivity of tuning fork structure is given by

$$\frac{\text{Output}}{\text{Input}} = \frac{\theta}{\Omega_z} = \frac{A}{(\omega_\theta^2 - \omega_y^2)} \sin(\omega_d t - \phi) \tag{3.49}$$

From the above derivations, we find that the sensitivity can be increased by

(a) Matching the torsional frequency with the tine drive frequency
(b) Matching the tine mode frequency with the drive frequency
(c) Minimising the damping to improve the Q factor
(d) Maximising drive amplitude y_m, which is possible by exciting the resonant mode

The above sensitivity improvement criteria are generally applicable to broad range of CVG with different types of structures where the modes are different.

3.7.5 Hemispherical Resonator Gyro

In this section, the operation of hemispherical resonator structure gyro will be explained where the driven mode and the sensed mode are identical. This gyro is of particular relevance as it has demonstrated navigation grade performance as well as providing very long life.

Bryan's discovery

Research for an inertial grade solid state gyro looked back to a very interesting discovery made by Bryan in the year 1890. Bryan wrote

> "If we select a wine glass when struck gives, under ordinary circumstances, a pure and continuous tone, we shall on twisting it around hear beats, thus showing that the nodal

meridians do not remain fixed in space. And if the observer will turn himself rapidly round, holding the vibrating glass all the time, beats will again be heard, showing that the nodal meridians do not rotate with the same angular velocity as the glass and observer".

Bryan concluded that the nodal angular velocity was about three-fifths that on the body.

Loper and Lynch, working for Delco Electronics Division of General Motors, USA, published a paper in the early eighties of the last century reporting the development of an inertial grade gyro for rotation sensing using Bryan's discovery. The gyro is called Hemispherical Resonator Gyro (HRG).

Operating principle of HRG

Consider that a thin section hemispherical shell, whose typical physical appearance is shown in Figure 3.23, is made to vibrate in its fundamental flexural mode, hereafter referred as $n = 2$ mode (also called lowest inextensional mode), by a suitable forcing scheme. This mode pattern of the resonator defines that in the first half cycle, the resonator deforms to its greatest ellipsoidal geometry and then returns to its hemispherical shape. In the next half cycle, a similar deformation takes place but spatially shifted by 90° in azimuth. This mode of vibration leads to four antinodes (A, C, E, G) and four nodes (H, B, D, F), as shown in Figure 3.24(a). The directions of the velocity (V_1, V_2, V_3, V_4, etc.) are such that they are radial at the antinodes and tangential at the nodes. Assume that this principal wave lies on the plane of the paper. We can assume that initially the node axis Y–Y′ is aligned with a case fixed axis X–X′ as shown in Figure 3.24(b).

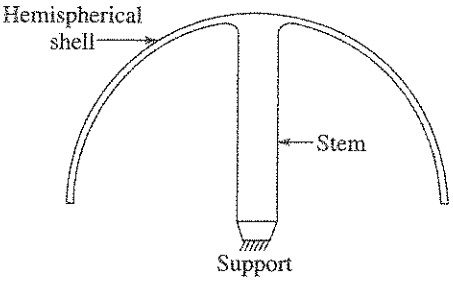


Figure 3.23 Hemispherical resonator configuration.

The shell is subjected to an angular rate $\mathbf{\Omega}$ about its sensitive axis, which is along the axis of the stem, with +ve direction pointing perpendicularly out of the paper. Assume that the stem is mounted on a turn table which provides the rotation. Coriolis forces are generated which are proportional to $\mathbf{V} \times \mathbf{\Omega}$ and the direction of these forces, being perpendicular to both $\mathbf{V}$ and $\mathbf{\Omega}$. The resultant of these forces excites another standing wave whose antinodes are now shifted to the original wave nodal points and the nodal points coincident with the original wave antinodal points. As a result, this new wave is called orthogonal wave.

The superposition of the original wave and the new orthogonal wave results in a phenomenon, where the resultant wave rotates relative to its own casing and to inertial space through an angle θ which is proportional to the inertial rotation θ_c of the case as shown in Figure 3.24(c).

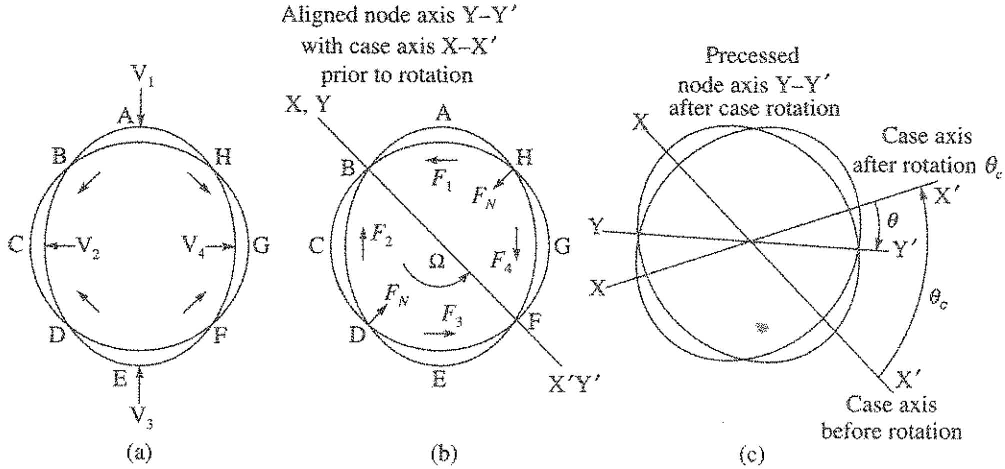


Figure 3.24 Operating principle of hemispherical resonator gyro.

The precession angle θ, which results in new pattern node axis (Y–Y′), is related to the applied inertial case rotation θ_c by an angular gain factor K, given by **[Bryan, 1890]** as:

$$\theta = -K\theta_c \tag{3.50}$$

The negative sign indicates that the pattern nodal axis rotation lags (moved CW) with respect to the CCW inertial rotation direction of the gyro case. The angular gain factor K is found to depend not on the size of the resonator, but on its shape and for the smooth hemispherical shape, it is 0.277 and for cylindrical, it is 0.4, the shape used by Bryan. So, by measuring the pattern node axis rotation angle, it is thus possible to determine the inertial rotation of the case.

EXAMPLE 3.4 (on HRG rotation detection): Assume initially, the gyro case axis and the pattern node axis are aligned as shown in Figure 3.24(b). A CCW inertial rotation is applied to the gyro case which results in the node axis to rotate in the same direction by 27°. Find the inertial rotation of the gyro case and the pattern rotation θ?

Solution: From Figure 3.24(c), it is clear that the pattern node axis has moved by

$$(\theta_c - \theta) = 27°$$

So

$$(\theta_c - K\theta_c) = 27°$$

For hemisphere $K = 0.27$, solving for θ_c gives

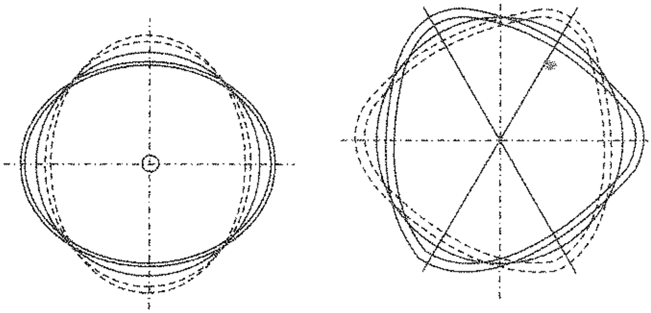
$$\theta_c = \frac{27}{1 - 0.27} = 36.98°$$

Pattern rotation angle $\theta = -0.27 \times 36.98 = -9.98°$.

Resonator modes

The inertial property of the pattern is preserved only when fundamental flexural mode of the hemisphere is excited, e.g. $n = 2$ or $n = 3$. These two flexural modes are shown in Figure 3.25(a) and Figure 3.25(b) respectively for better clarity. Even though, no gyro is currently operational in $n = 3$ mode, this mode has certain advantages relative to the already operational $n = 2$ mode.

But there are other nearby modes of the hemisphere, which, if get excited, will not work as gyro. These modes are not flexural and are called $n = 0$ and $n = 1$. The $n = 1$ is basically the bending mode and is so called because the part of it results from the compliance of the stem. At lower latitudes of the resonator, nearer to the open end, it is primarily a rigid body rotation about an axis that is perpendicular to stem axis [**Lynch, 1987**]. The $n = 0$ mode represents contraction and expansion of the resonator. It is important that both these lower modes are suppressed through appropriate resonator design.



(a) $n = 2$ mode of flexing (a) $n = 3$ mode of flexing

Figure 3.25 HRG resonator flexural operating modes.

Configuration of hemispherical resonator gyro

A typical HRG with three principal parts is shown in Figure 3.26. The top one is the outer enclosure containing the forcer electrodes, the middle one is the hemispherical resonator with tines and support stem and the bottom enclosure contains the pick-off electrodes. The assembly is enclosed in a metal housing with getter under an evacuated environment.

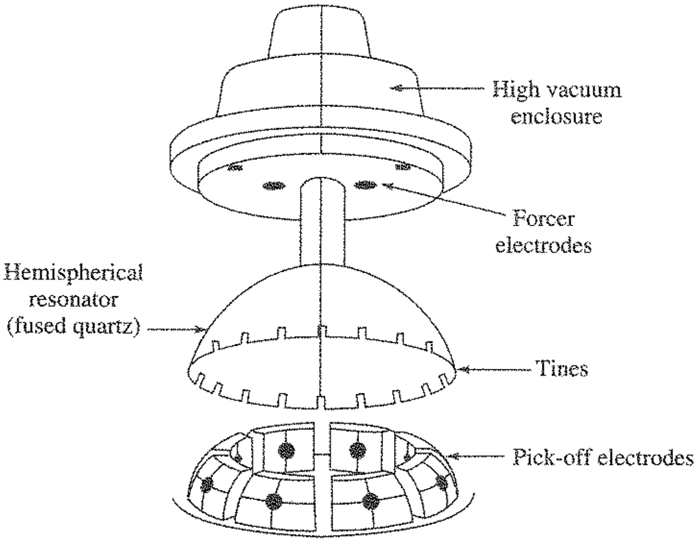


Figure 3.26 Features of a hemispherical resonator gyro.

The electrostatic forcing is provided by two sets of electrodes, in which one is a ring shaped electrode located between typical latitudes of 25° and 45° and facing the resonator. The ring electrode and the resonator form a part of a spherical design. The other forcer is a set of discrete electrodes, typically 16 in number, symmetrically placed between latitudes of 0° and 25°. The material used for the forcer housing is also fused quartz to provide a material compatibility with the resonator.

Discrete forcer electrodes are utilised for initiating resonator vibration. Using a ring forcer, there will be no net force on the resonator if gap remains uniform all around. This is the reason why the ring electrode is not suitable for initiating the vibration when the resonator is undeformed at the start. However, once the vibration is initiated, the function of the ring electrode is to accurately control the vibration amplitude and sustain the vibration by compensating the energy losses due to damping as the rate at which the amplitude a of the standing wave diminishes due to damping is proportional to a/τ.

The hemispherical resonator, the most critical part of the gyro, is made from fused quartz, which has extremely low internal damping and provides very high Q factor (~5×10^7). This enables resonator Q exceeding five million in an inertial grade gyro using appropriate manufacturing technology. The Q factor is further boosted electronically with parametric resonator drive. The figure of merit for the resonator is the time constant t, which is a measure of its ability to sustain the free vibration when forcing is withdrawn. The time constant t (for derivation refer Appendix C) is given by the relation:

$$\tau = \frac{2Q}{\omega} \tag{3.51}$$

where ω is resonator vibration frequency. Typical t value reaches around 500 s for navigation grade gyro.

Pick-offs provide the measurement of rotation θ of the precessed node axis using a set of capacitive electrodes which are typically 8 in numbers as shown in Figure 3.27.

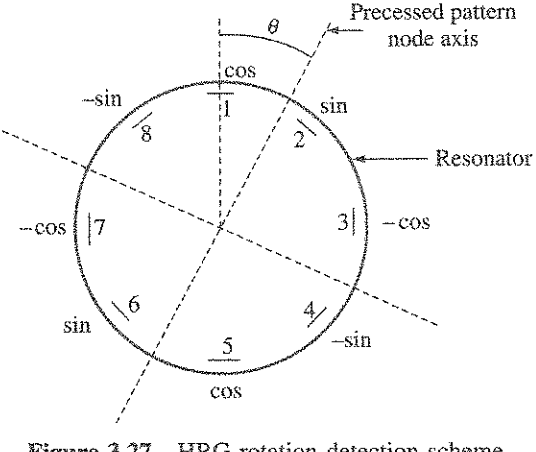


Figure 3.27 HRG rotation detection scheme.

Considering the direction of movement of the precessed angle in $n = 2$ mode, the group of four electrodes (1, 3, 5, 7) constitute one channel of information. If this channel is designated as cosine channel, then electrodes group (2, 4, 6, 8) can be designated as sine channel by noting that electrode group (2, 4, 6, 8) is 90° electrical away from electrode group (1, 3, 5, 7). The amplitude-modulated outputs of the two channels, with resonator frequency ω as the carrier, can be written as:

$$E_c = \text{output of cosine channel} = 1 - 3 + 5 - 7$$

$$E_s = \text{output of sine channel} = 2 - 4 + 6 - 8$$

After phase sensitive demodulation of E_c and E_s, angle θ can be obtained by

$$\theta = \frac{1}{2}\tan^{-1}\frac{{}^*E_s}{{}^*E_c} \tag{3.52}$$

${}^*E_c$ and ${}^*E_s$ are demodulated outputs of cosine and sine channel respectively. However, it may be noted that Eq. (3.52) is computationally unstable when θ approaches 90°, so additional processing is involved to tackle this problem. The computed angles are called electrical angles and to be divided by 2 to obtain the mechanical precession of the nodal axis. The most successful pattern precession angle detection method is called whole angle tracking mode which can provide large angular rate range and low angular resolution. Suitable high input impedance electronics, closely located, are necessary for processing the capacitive outputs to improve upon signal-to-noise ratio.

HRG can also be operated in rebalance mode when the precession angle θ is nulled, and the force necessary to make θ zero becomes a measure of the input inertial rate. However, the angular rate range is limited to around 10°/s due to capacitive nature of torquing.

HRG errors

There are several errors in HRG to contribute to drift, and it becomes necessary to take care of them in design as well as in assembly process. A dominant error source has been found on non uniform mass distribution along the circumference of the resonator shell geometry arising out of non uniform shell thickness. This aspect is shown in Figure 3.28.

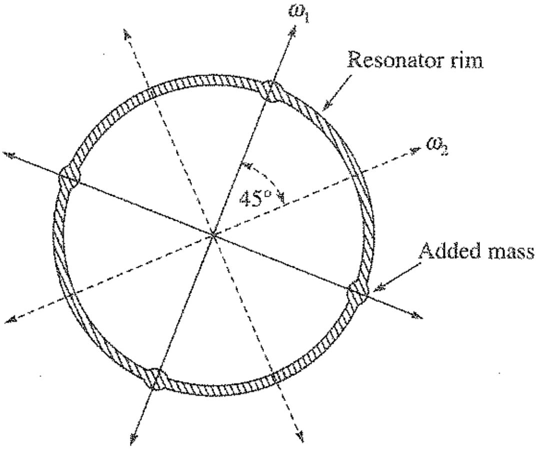


Figure 3.28 Resonator with non uniform mass distribution.

A resonator is shown to have a non uniformity so that the shell thickness is slightly more at four points, which is equivalent to added mass at those four equispaced points. This leads to two distinct flexing mode natural frequencies ω_1 and ω_2 where ω_1 is less than ω_2 due to heavier mass at these points. The resultant of these two flexing mode natural frequencies, each corresponding to a definite standing wave location, gives rise to travelling wave components which cannot provide inertial reference. Another valid description is that a secondary wave is created due to the non uniform mass distribution, having its antinodes coinciding with the nodes of the initial wave, and oscillating in phase quadrature with the initial wave.

The quadrature wave has a time growing property and as a result, a time is reached when the gyro cannot detect inertial rotation. Methods have been devised to control the quadrature drift well within an acceptable limit using mechanical means as well as electrical method. In the mechanical method, a series of tines, refer Figure 3.26, are made at the hemisphere lip and material is physically removed by polishing technique so as to appreciably reduce the non uniform mass distribution. This is followed by closed loop electrical forcing using discrete electrodes which never allow the quadrature drift to grow. **Lynch and Loper, 1990** have made pioneering contributions in solving the quadrature problem and readers can refer to their publication listed in the references at the end of this chapter.

HRG with newer technology

Considerable research during the last decade has enabled evolution of new configuration with the aim to reduce size and manufacturing cost. Currently, the resonators are dispensing with the tines as well as the ring electrode which means, currently, all are discrete electrodes of adequate number. In another development [**Rosellini et al., 2008**], the longitudinal vibration component along the stem is detected with flat capacitive plates in lieu of curved plates described earlier. The resonator size has been brought down to 20-mm diameter yet providing inertial grade performance.

3.8 Basic Principle and Theory of Operation of Optical Gyros

3.8.1 Sagnac Effect Gyro Principle

All modern optical gyros use the principle of Sagnac effect in detecting inertial rotation (Figure 3.29).

If two counter-propagating light beams travel in a closed circular path after entering at point P in inertial space, then, in the absence of an inertial rotation perpendicular to the plane of the circular path, the travel time for both the beams are same as the beams traverse the same optical path length and exiting at the same point P in inertial space. The travel time t_0 is given by

$$t_0 = \frac{2\pi R}{c} \tag{3.53}$$

When the closed optical path is rotated at angular rate Ω, then the travel time for both the beams are not same as the beams have entered at P but leaving from another point P_1 in inertial space. This means that the rotation has created a path length difference ΔL between the counter-propagating beams in the closed optical path. This phenomenon is called *Sagnac effect*. The difference in the beam travel time Δt, due to path length difference and to a first order of approximation, is given by (**Rudloff, 1999**) as:

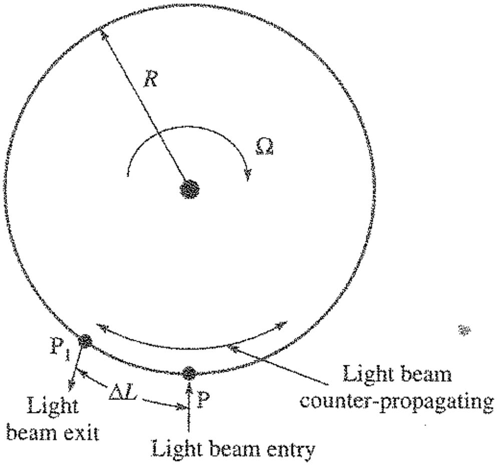


Figure 3.29 Sagnac effect gyro principle.

$$\Delta t = \frac{4\pi R^2}{c^2}\Omega = \frac{4A}{c^2}\Omega \tag{3.54}$$

where

A = area enclosed by the closed optical path πR^2
c = velocity of light (3×10^8 m/s)
Ω = input angular rate about inertial space

Detection of the time difference Δt by a measurement scheme, which can be instrumented, provides information on the input rate as area A is a design constant. Equation (3.54) can also be formulated as path length difference ΔL where ΔL is given by $c\Delta t$.

In order to measure the rotation rate, the direct measurement of the time delay or the path length change, provides an impracticable solution to an engineered optical gyro. This can be understood when we notice that the time difference Δt is only around 2×10^{-26} second for a gyro of area 100 cm^2 and which is to detect an inertial angular rate of 0.01°/h (a typical sensitivity needed for autonomous navigation). This time difference is much smaller even compared to the resolution of an atomic clock. The corresponding optical path length difference ΔL is also extremely small, around 2×10^{-18} metre, which is much smaller than even hydrogen atom whose diameter is around 10^{-13} metre.

Alternative to time measurement considered was phase measurement. For a beam of frequency f, the change in phase $\Delta\phi$ between the clockwise and counterclockwise beams is given by

$$\Delta\phi = \Delta t 2\pi f$$

Substituting for Δt, for a ring shaped Sagnac setup, the phase change becomes:

$$\Delta\phi = \frac{8\pi A f}{c^2}\Omega \tag{3.55}$$

This is a scheme that is implemented in a passive fibre-optic gyro with an ingenuity that increases area A manifold to increase the sensitivity, and the gyro is described later in Section 3.8.3.

Working further for another scheme and using interferometer principle, the phase shift change of the recombined waves is seen as a relative shift (ΔZ) of the interference fringe pattern given by

$$\Delta Z = \frac{\Delta\phi}{2\pi} \quad = \frac{4Af}{c^2}\Omega$$

Further, since $f = c/\lambda$ where λ is the free space light wavelength. On rearranging we get:

$$\Delta Z = \frac{4}{c\lambda}(A\Omega) \tag{3.56}$$

This type of passive Sagnac ring interferometer for detection of angular rate is shown in Figure 3.30. It is to be noted that the axis defining the plane of the optical area is parallel to the rotation rate for maximum sensitivity. Relative fringe shift is defined as the ratio of the absolute fringe shift to the distance between two neighbouring fringes.

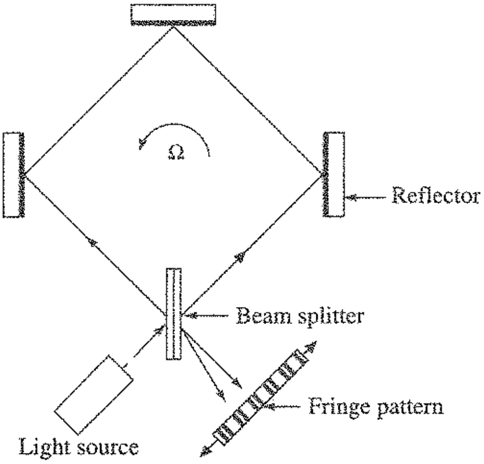


Figure 3.30 Passive Sagnac ring interferometer.

If we consider $\lambda = 0.63$ μm and compute the relative fringe shift, ΔZ becomes 1×10^{-11}, which is also extremely low. While Figure 3.29 depicts a scheme where the light source is passive, such a scheme has been actually and successfully implemented using active laser resonator, and the gyro is called ring laser gyro.

Further, it has been shown [**Armenia et al., 2010**] that the Sagnac effect is same if the homogeneous light medium is having a refractive index in lieu of vacuum.

3.8.2 Ring Laser Gyro

The principle of laser is covered in various publications and for a clearer appreciation on the topic of laser gyro resonator where various phenomena and terminologies are involved, Appendix A attempts to provide some descriptive details. The important part of ring laser gyro operation is to create an *active ring type laser resonator* which can detect the rotation induced Sagnac path length difference.

Principle of operation

Two operating conditions to be fulfilled to achieve a resonator, either a linear or ring, are similar. They are given as under:

(i) In a resonator, the gain provided by the amplifying medium must exceed the losses in the cavity due to scattering, transmission and absorption.

(ii) The resonator oscillation is determined by the condition that the optical path length L for the beam to return to itself, is an integral number m of vacuum wavelength λ, so that the relation $L = m\lambda$ is ensured.

In a linear resonator, the oppositely directed travelling waves with equal amplitude and frequency constitute the standing wave. In the case of ring laser interferometer, the CCW beam and the CW beam can be independent so that each can oscillate with different amplitude and frequency. Noting this aspect as the ring resonator characteristic, the second condition, $L = m\lambda$ can then be written by substituting λ as:

$$f_+ = \frac{mc}{L_+} \tag{3.57a}$$

$$f_- = \frac{mc}{L_-} \tag{3.57b}$$

where f_+ represents CCW beam frequency and L_+ represents corresponding optical path length. Similarly, f_- and L_- represent frequency and optical path length respectively for the CW beam.

The relative frequency difference can then be written as:

$$\frac{\Delta f}{f} = \frac{\Delta L}{L} \tag{3.57c}$$

In the ring laser gyro, rotation induced Sagnac effect creates the optical path length difference ΔL, which then can be measured by measuring the frequency shift. This can be deduced from Eq. (3.54) as follows:

$$\Delta L = c \cdot \Delta t = \frac{4A}{c} \cdot \Omega \tag{3.57d}$$

Substituting ΔL from Eq. (3.57c) in Eq. (3.57d) and after further simplification, Sagnac effect can be finally expressed as shift in frequency given by

$$\Delta f = \frac{4A}{\lambda L} \Omega \tag{3.57e}$$

Under resonating conditions inside the ring laser cavity, the clockwise and the counterclockwise laser beams are created (refer Figure 3.31) that travel with identical frequency in the resonator cavity.

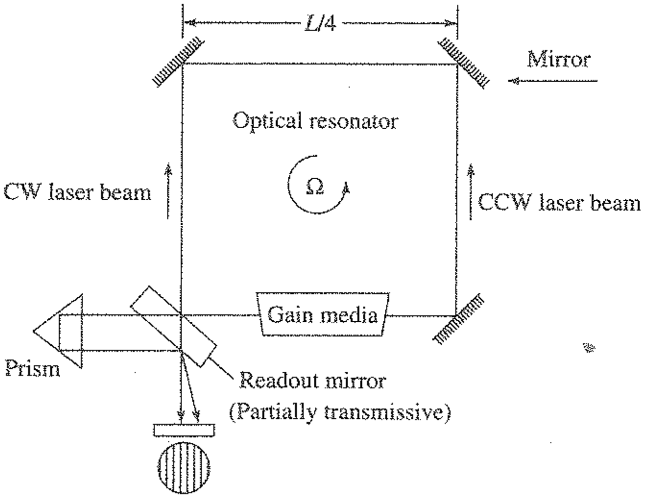


Figure 3.31 A basic ring laser gyro operation.

When the ring laser cavity is rotated, Sagnac effect causes a change in the optical path length of one beam relative to the other, which in turn results in the two beams to have different oscillation frequency. In this way, the optical path length difference is transformed to a much better instrumentable optical frequency difference measurement, and thereby provides an elegant scheme that forms the basis of RLG rate detection. The laser resonator being an integral part of the gyro makes RLG to be defined as an active gyro.

The output power of the resonator is available through a partially transmissive mirror known as read out mirror. This power output is optimised through a single longitudinal mode that is tuned so as to align with the centre of the gain line, also called peak power point, which is shown in Figure 3.32.

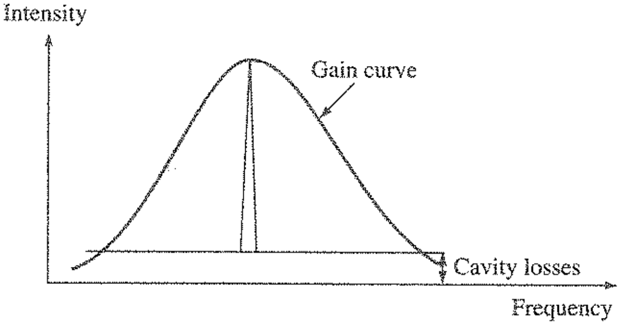


Figure 3.32 Resonator operations with single mode.

Readout of rotation

The principle of sensing the rotation in RLG is through the observation of the fringe pattern movement (Figure 3.33). The coherent combination of CW and CCW beams is carried out through a combination of dielectric mirror, which is a part of the resonator, and a combining prism of refractive index n outside the resonator. This combination results in an interference pattern whose amplitude can be described as a co-sinusoidal function with high and low intensity points known as fringes. The time and space dependent interference pattern can be expressed by the equation:

$$I(x, t) = I_0\left[1 + \cos\left(2\pi\Delta f t + 2\pi\frac{\gamma}{\lambda}x + \phi_0\right)\right] \tag{3.58}$$

where

I_0 = average beam intensity

Δf = frequency difference due to Sagnac effect

γ = angle between the output beams

x = spatial co-ordinate measured along detector array

ϕ_0 = constant phase difference between the beams

When the input angular rate is zero, Δf will be zero. Eq. (3.58) shows that the fringe is stationary in space with fringe spacing d which is a design constant. The photodiodes measure the intensity of the interference fringes. As the fringes pass by the diodes, sinusoidal electrical output signals are generated, with each cycle of the sine wave corresponding to the movement of one fringe over the diodes. Two photodiode detectors are used, which are separated by $\lambda/4$. This is equivalent to 90° in fringe space, and they are used to detect the direction of rotation by observing which diode output is leading the other.

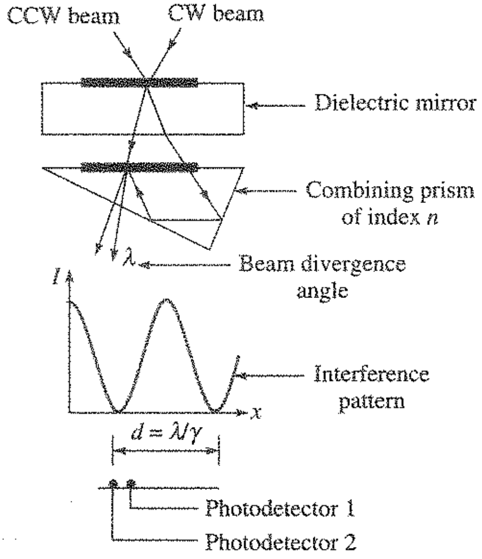


Figure 3.33 RLG rotation readout scheme.

Rate sensitivity derivation from geometry

Ring laser gyro configuration may have square block geometry or equilateral triangular geometry. The rate sensitivity can also be derived from the geometry aspect as shown below.

It is assumed in this derivation that the RLG is a square block having four equal sides of length $L/4$, and the rotation rate Ω is acting about point O which is the centre of square block RLG as shown in Figure 3.34.

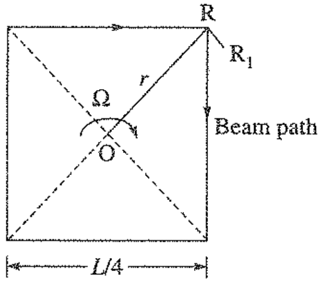


Figure 3.34 RLG geometry.

In the absence of rotation, the transit time of light beam, starting at point R and returning to the same point has been stated earlier, and is given as:

$$t = \frac{L}{c}$$

When rate Ω is imparted to RLG, point R moves to R_1 through a distance d in the light transit time, where d is given by

$$d = \Omega r t = \frac{\Omega r L}{c} \tag{3.59}$$

For the square block $r = \dfrac{L}{4\sqrt{2}}$

Projecting d on the beam path, we get shift in path length as:

$$\Delta L = d \cos 45 = \frac{d}{\sqrt{2}}$$

On substitution of d and r, we have:

$$\Delta L = \frac{2(L/4)^2}{c}\Omega$$

Area A of the square block is $(L/4)^2$, so path length shift can be expressed as:

$$\Delta L = \frac{2A}{c}\Omega \tag{3.60}$$

Since $L = m\lambda$, to satisfy the resonator path length change to wavelength change, we can write:

$$\Delta\lambda = \frac{\Delta L}{m} = \frac{\lambda \Delta L}{L} \tag{3.61}$$

We can now express the clockwise beam frequency change as:

$$\frac{\Delta f_{\mathrm{CW}}}{f_{\mathrm{CW}}} = \frac{\Delta\lambda}{\lambda} = \frac{\Delta L}{L} \tag{3.62}$$

The shift in the CCW beam frequency magnitude is same as that of CW frequency. Hence the total shift in frequency Δf between the two beams is:

$$\Delta f = 2\Delta f_{\text{CW}} = \left(\frac{2\Delta L}{L}\right) f \tag{3.63}$$

Substituting ΔL from Eq. (3.60) in Eq. (3.63), and simplifying by noting that $f = c/\lambda$, we get:

$$\Delta f = \frac{4A}{\lambda L}\,\Omega \tag{3.64}$$

where

Δf = frequency difference between CCW and CW laser beams
A = closed optical path area
λ = laser beam wavelength in vacuum
L = closed path optical length.

Sometimes, this frequency difference is also called *Beat frequency*.

The operational feasibility of the RLG can be easily visualised from the fact that for the earlier example of 0.01°/h rotation rate detection, the frequency shift will be around 0.0087 Hertz with typical laser wavelength of 0.63 μm. This is relatively easier to measure electronically. However, it should also be noted that at this wavelength, the frequency of the laser beam is around 5×10^{15} Hertz and that the resonator frequency as well as the optical path length must be extremely stable to ensure accurate measurement of this small frequency difference relative to the absolute value of the beam frequency. This makes RLG technology highly challenging.

In reality, RLG rate detection scheme is mechanised to measure rotation angle by counting the fringes. Equation (3.64), when integrated over a specified time, can be expressed as:

$$N = \frac{4A}{\lambda L}\,\Delta\theta \tag{3.65}$$

where N represents the number of beat frequency that is observed during the integration time when the gyro is rotated through an angle $\Delta\theta$. With this typical readout scheme, RLG behaves like a rate integrating gyro. The scale factor of RLG can be deduced by putting $N = 1$ in Eq. (3.65), and this gives the gyro scale factor S_k as:

$$S_k = \frac{\lambda L}{4A} \tag{3.66}$$

Defined in this way, we see that the unit of scale factor S_k is in radian/pulse or radian per count or in a more convenient scale used by inertial community to read as arc-sec/pulse. However, it is quite possible to define the scale factor as the inverse of S_k, which will mean the unit of scale factor as counts/radian. By measuring the number of fringes moving across the field of view for a fixed time, the angular rate can be determined by using the scale factor S_k. A simple counting of the fringes provides for the inertial rotation angle.

Number of mirrors or sides in ring resonators

In ring resonator to create CCW beam and CW beam, minimum of three or more mirrors are necessary and arranged in a defined planar configuration. In the development history of RLG, three or four or five-mirror design were invoked. Later on only three-mirror version and the four-mirror version got stabilised. It may be prudent to use the term 'reflector' as resonator designs are in operation with total reflecting prisms instead of mirrors. This aspect is discussed later.

Explanation of basic RLG operation, for a three-sided equilateral triangular geometry, remains same, and similarly the sensitivity is also given by Eq. (3.64).

EXAMPLE 3.5 (on RLG rate sensitivity): A triangular shaped RLG is having a side length of 10 cm and operating with He–Ne gas laser having wavelength of 0.63 μm. Calculate the frequency shift to detect 0.01°/h rotation rate and also the scale factor?

Solution: The frequency shift is given by

$$\delta f = \frac{4A}{\lambda L}\Omega$$

Here $L = 30 \times 10^{-2}$ m, $\Omega = \frac{0.01}{57.4} \times 3600$ rad/s $= 4.85 \times 10^{-8}$ rad/s, $\lambda = 0.63 \times 10^{-6}$ m

$$\text{Area } A = \sqrt{\frac{3(L/3)^2}{4}} = 1.73 \times \left(\frac{30 \times 10^{-2}/3}{4}\right)^2 = 43.25 \times 10^{-4} \text{ m}^2$$

Putting the values and solving gives:

$$\delta f = 4.43 \times 10^{-3} \text{ Hz}$$

Scale factor

$$S_k = \frac{0.63 \times 10^{-6} \times 0.3}{4 \times 43.25 \times 10^{-4}} = 1 \times 10^{-5} \text{ rad/pulse} = 2.06 \text{ arc-s/pulse.}$$

RLG features

RLG design and development were focussed for a high performance autonomous navigation grade strapdown gyro with high reliability and producibility. As expected, the technology evolved through number of configurations involving various designs on all the key areas. Typical functional elements of RLG are:

(i) Ring resonator (primarily consists of gas medium, gas discharge and mirrors or reflectors)

(ii) Path length control

(iii) Block material

(iv) Lock-in solution.

These are further described with the help of Figure 3.35 in the following subsections.

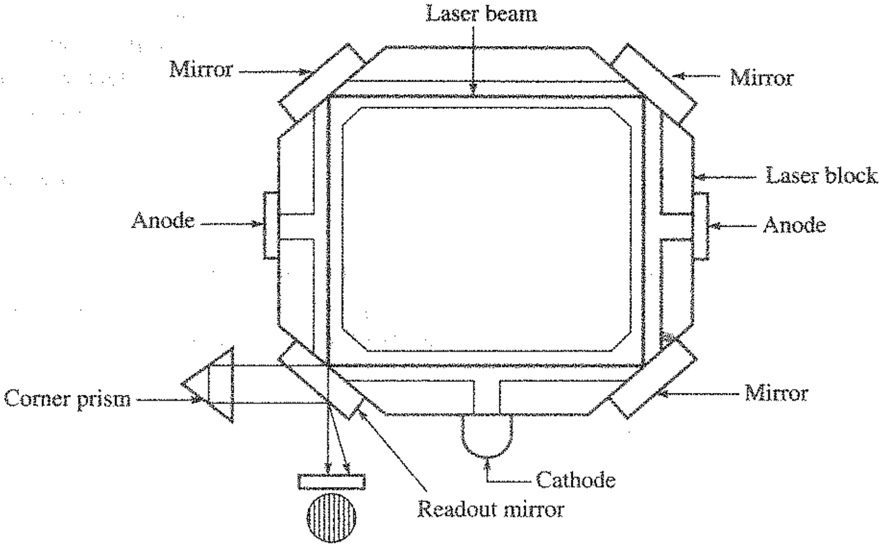


Figure 3.35 Ring laser gyro simplified configuration.

Ring resonator: The most widely used lasing medium inside the cavity is a gas that is a mixture of helium (He) and neon (Ne). This mixing is necessary as neon alone is not excitable to lead to population inversion. While population inversion is necessary for laser operation, it is not sufficient for gyro use as the resonator is also to be designed for stable operation and not optimised for high power. Besides these requirements, low diffraction losses in the cavity, avoiding gain competition between the beams, single longitudinal mode operation, avoiding mode locking at low rates, gyro performance and long operating life, all put difficult requirements for He–Ne resonator.

He–Ne laser has been designed to emit light at wavelengths 0.5435 μm (green), 0.612 μm (orange), 0.6328 μm (red) and 1.523 μm (infrared). The current standard generally followed, after other earlier experimentations on different wavelengths, is to use the transition which is 0.6328 μm. However, to avoid gain induced mode competition between the two beams, neon is a mixture of isotopes Ne^{22} and Ne^{20} in equal proportion. Gas fill pressure and the He–Ne mixture ratio are important controlling parameters for resonator operation, and these are design specific. Control on these parameters establishes important gyro parameters such as laser gain, gyro thermal sensitivity and performance.

Gas mixture has to be suitably excited to ensure the lasing operation and to ensure proper resonator operating conditions. The gas is excited by generating plasma in the gas which is ionised through the application of high voltage (typically a few hundreds of volts) between the cathode and the two anodes. When the discharge occurs, ions flow. The flow of ions generates a drag, called Fresnel drag, a source of bias in RLG. The discharge is split, with the two anodes, to generate opposite flow and this cancels the Fresnel related bias. Thus two-anode configuration of RLG resonator becomes a design standard.

An alternative approach to high voltage discharge is realised through the use of high frequency discharge. To ignite the high frequency discharge, high frequency pulse transformer is used. It forms high voltage pulse of amplitude between 1.5 kV to 2 kV and 1.5 μs to 3 μs duration. Currently, both the schemes are operational.

Mirrors play crucial role in resonator performance. Three or more mirrors are necessary to realise the ring laser resonator and currently, two versions of gyro are mostly operational which are using either three mirrors or four mirrors. There are four dielectric mirrors shown in Figure 3.35. The number depends on the number of sides in RLG. The reflecting surfaces of the dielectric mirrors are designed to selectively reflect the frequency associated with the particular helium–neon laser transition being used. For stable resonator operation, one or two mirrors are curved with concave surface. Concave mirrors, in general, have lower diffraction losses compared to the plane mirror. Besides such crucial role in resonator operation, the quality of the mirrors and their alignment during gyro assembly are both critical. Optical quality mirror surfaces are needed to reduce differential phase shifts between the two beams and to reduce lock-in effect.

The mirrors get rigidly attached to laser block by optical contact process that additionally provides contamination free assembly required for long gyro life. Avoidance of any form of contamination during resonator assembly is also important. Over a time, the mirror quality tends to degrade due to the interaction of the discharge on the mirror surface. Mirror technology, for stable resonator operation, gyro accuracy and long life, constitutes a key area in RLG development and production.

An alternative to mirror technology has been through the use of *Total Reflecting Prisms* (TRP). The TRP scheme is shown in Figure 3.36 where light is totally and internally reflected within the prism. Currently, both the schemes are operational.

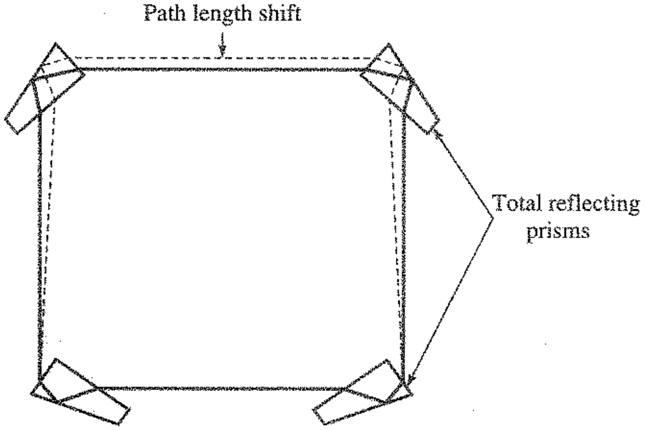


Figure 3.36 Four sided RLG with total reflecting prisms and temperature dependent beam path length shift.

Path length control: In a gyro with reflecting mirrors, the change of path length is due to thermal expansion or contraction of the block material even though the block material chosen is characterised by very low coefficient of thermal expansion. Resonator self heating due to discharge as well as environmental temperature variation contribute to change in path length. The dotted lines in Figure 3.37 show the shift in gyro path length from the original path length shown by solid line. In a typical form of path length control, the scheme ensures that the path length fits to the centre of the gain line, the line of peak average beam power. In Figure 3.37, the path length control mirror is shown at top right corner that is moved by piezo driven actuator, to correct the change in path length. The process is also called cavity tuning. The scheme thus allows constancy of optical power as well as the scale factor of RLG.

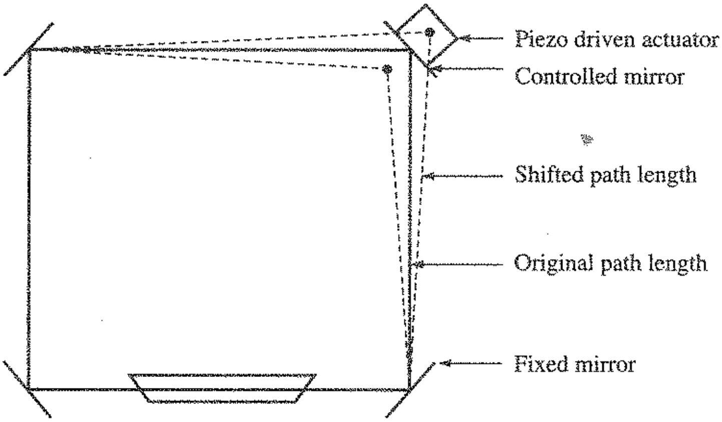


Figure 3.37 Cavity path length control in a RLG with mirrors.

When total reflecting prisms are used, the change in path length is attributed due to the change of refractive index of the TRP and variation of optical path length in the TRP due to temperature. This causes tens of wavelength shifts in path length across the operating temperature with a variation of around $0.8\lambda/°C$ due to gyro self heating. As a result, the cavity tuning with piezo driven actuator is not suitable. The dotted line in Figure 3.36 shows the shifted beam path with temperature. The path length shift from the centre of gain line is observed through synchronous variation of the gyro output amplitude. The path length control for this type of RLG configuration is achieved by air density variation in one arm of the laser block by using controlled heater in a closed loop mode.

Block material: RLG block material needs some special characteristics to provide performance and life. The material should have ultra low thermal coefficient of expansion. It should allow very low diffusion of helium and lastly amenable to precision machining and optical quality polishing. Materials with trade names Cervit, Zerodur and Sital are normally used. Cervit is not currently in production. Boroslicate is another candidate material, but it has relatively high thermal expansion coefficient. This could be suitable for lower performance RLG.

Lock-in problem and solutions: The lock-in phenomenon in a RLG is its unique feature with severe detrimental effect on performance. The phenomenon of lock-in originates, when two oppositely directed travelling waves couples to provide frequency synchronisation that results in the two beams oscillating at the same frequency. The frequency synchronisation, commonly known as lock-in, results in a scenario where RLG is able to sense input rotation above a threshold value only. The phenomenon is shown in Figure 3.38. This problem manifests when the input rate to the gyro approaches towards zero from either direction. If Ω_L is the lock-in rate, then the effective lock-in band is $2\Omega_L$. Since the behaviour manifests before entering the lock-in rate, this extended range of improper behaviour is shown as lock-in region.

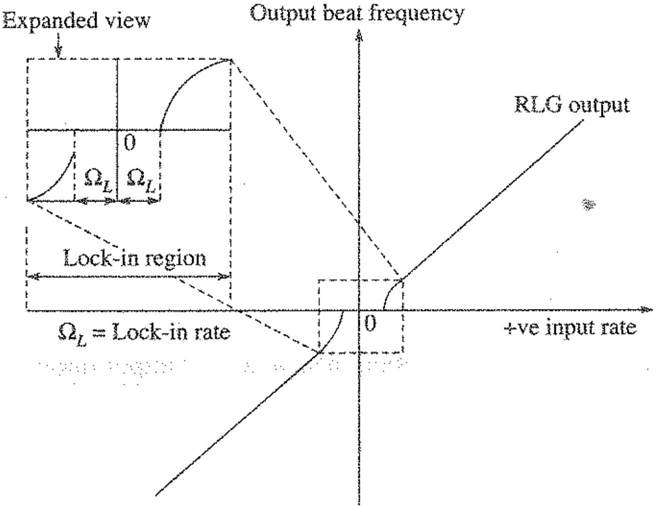


Figure 3.38 Lock-in phenomenon in a RLG.

The typical lock-in rate, in production version of RLG, can be around or more than 300°/h. This means that within this value, RLG will not be able to detect angular rate. Beyond this lock-in rate and up to the lock-in region, the scale factor becomes highly non-linear. Possible lock-in reason is due to back scattering of laser beam from the mirrors. This would mean in a square block gyro with four mirrors, there will be four back scattered wave of CW beam which will couple with CCW beam. A similar phenomena for the back scattered CCW beam coupling with CW beam where, with the understanding of the mechanism of lock-in, designs [**Faucheux et al., 1988**] have been evolved to reduce its absolute magnitude.

Various lock-in elimination schemes have evolved over time. Out of these, two operational schemes are described which are known as:

(i) Mechanical dither
(ii) Multi-frequency gyro

The first scheme involves mechanical motion to the gyro while the next scheme is based on magneto-optical effect.

Mechanical dither: Mechanical dither scheme utilises a concept that by rapidly moving the gyro through the lock-in region, the time constant of the system would keep the gyro from getting locked. The basic scheme uses sinusoidal motion about the gyro input axis with amplitude much more than the lock-in band. This enables the gyro to stay out of lock-in condition most of the time except a small period when the amplitude is less than the lock-in magnitude. However, this approach did not totally solve the problem, and a final solution was achieved with adding amplitude noise to the sine dither. The detrimental effect of random dither noise on gyro is in making the gyro output itself noisy, whose value R_w is given by [**Aronowitz, 1999**] as:

$$R_w = \Omega_L \sqrt{\frac{S_k}{2\pi \Omega_D}} \tag{3.66}$$

where

Ω_D = dither rate amplitude

Ω_L = lock-in magnitude

The noise R_w in the output is called angular random walk drift whose unit is $°/\sqrt{\text{h}}$. It is to be noted that the random walk originates due to integration of white noise process. The foregoing discussion on lock-in rate and its elimination is summarised in Table 3.1, illustrating the effect of dither on a RLG with certain assumed values such as scale factor of 2.5 arc-s/pulse and lock-in magnitude of 300°/h.

Table 3.1 Effect of Dither on RLG Lock-in Threshold

Type of dither	Rate threshold	Output noise
None	Ω_L = 300°/h	0
Sinusoidal (Ω_D = 175°/s)	$\Omega_L(d)$ = 8.8°/h	0
Randomised sine	0	$R_w = 0.003°/\sqrt{\text{h}}$

EXAMPLE 3.6 (on mechanical dither): Refer Example 3.5 of a triangular shaped RLG, having a side length of 10 cm and operating with He–Ne gas laser having wavelength of 0.63 μm. Assume the lock-in magnitude to be 300°/h, and the dither amplitude is 200 arc-s peak at a dither frequency of 500 Hz. Calculate the gyro random walk.

Solution: Scale factor S_k = 2.06 arc-s/pulse (obtained earlier) = 2.06/3600°/pulse

$$\Omega_L = 300°/\text{h}$$

$$\text{Dither rate } \Omega_D = \frac{2\pi \times 500 \times 200}{3600} = 174.4°/\text{s} = 6.28 \times 105°/\text{h}$$

On substituting the values, we get:

$$R_w = 300\sqrt{\frac{2.06}{3600 \times 2\pi \times 6.28 \times 10^5}}\ °/\sqrt{\text{h}} = 0.0035°/\sqrt{\text{h}}$$

It should be noted that the normally used unit of random walk is in $°/\sqrt{\text{h}}$ rather than rad/$\sqrt{\text{s}}$. In SI unit, the computed value will be 0.1 × 10^{-5} rad/$\sqrt{\text{s}}$. It is necessary to take care in the conversion of appropriate units.

Multi-frequency gyro: Multi-oscillator ring laser gyro is an innovative scheme where the lock-in problem is eliminated by a change in the resonator design along with magneto-optical biasing. The scheme aims to create large optical bias, also called Faraday bias, in the resonator so that the gyro would always work outside the lock-in threshold range.

This gyro is also called multi-frequency gyro or a four-frequency gyro as against two-frequency gyro discussed earlier. The operating principle of this type of gyro is shown in Figure 3.39.

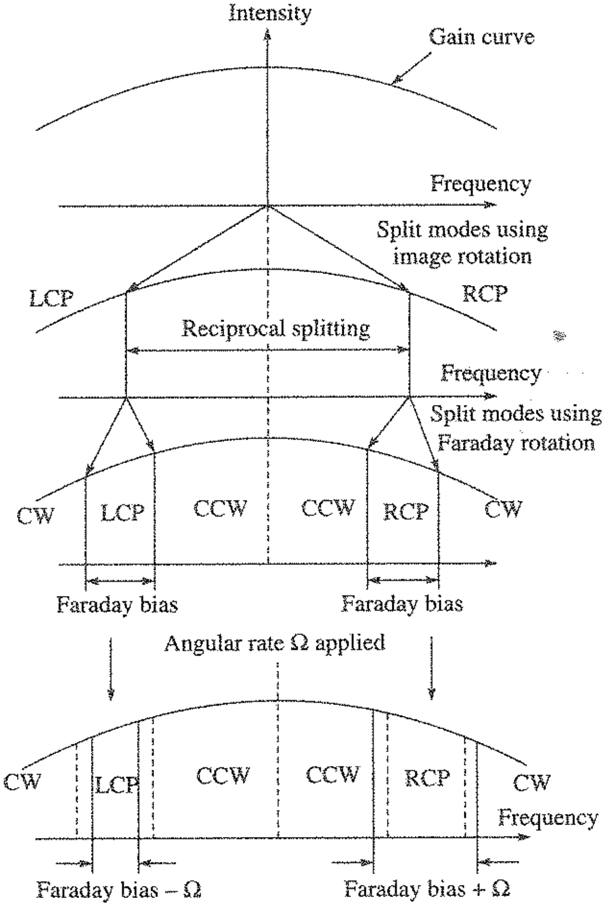


Figure 3.39 Principle of multi-frequency RLG.

In this gyro, the CW and CCW laser frequencies are initially split into left circularly polarised (LCP) beam and right circularly polarised (RCP) beam. The process being termed as *reciprocal splitting*. These polarised beams are then split into four frequencies in such a manner that a pair of CW and CCW beam is left circularly polarised (LCP), whereas the other pair of CW and CCW beam is right circularly polarised (RCP). In effect it creates two gyros in one resonator with large but equal Faraday bias in both. When angular rate Ω is applied, Sagnac effect changes the magnitude of split frequencies, increasing in one gyro while decreasing in the other gyro. Effect of these operations in eliminating lock-in and at the same time providing workability as a gyro, can be explained by the following equations:

$$\Delta f(\text{RCP}) = B_f + S_k\Omega \tag{3.67}$$

$$\Delta f(\text{LCP}) = B_f - S_k\Omega \tag{3.68}$$

Taking the difference, we get:

$$\Delta f(\text{RCP}) - \Delta f\,(\text{LCP}) = 2S_k\Omega \tag{3.69}$$

The difference operation gets rid of the Faraday bias B_f and at the same time doubles the RLG scale factor. If the highest operating rate of the gyro is $\Omega_{\max}$ and lock-in rate is Ω_L, then the required Faraday bias is given by

$$|B_f - S_k\,\Omega_{\max}| > \Omega_L \tag{3.70}$$

In effect, the four-frequency gyro is also called a differential gyro.

Thus the scheme avoids the mechanical dithering with their attendant complexity and cost. Currently, such multi-oscillator gyro has been evolved with considerable amount of ingenuity **[Volk et al., 1999]** and is referred as zero lock gyro. There are other innovative methods on lock-in **[Aronowitz, 1999]** which have produced navigation grade operational gyros.

EXAMPLE 3.7 (on Faraday bias): In a square block multi-frequency RLG, the rate scale factor is 5 × 10^3 Hz/°/s and has a operating range of 100°/s. If the lock-in band is 300°/h, calculate the Faraday bias required.

Solution: The applicable equation is given by

$$|B_f - S_k\,\Omega_{\max}| = \text{or} > \Omega_L$$

where

$S_k = 5 \times 10^3$ Hz/°/s

$\Omega_{\max} = 100°/\text{s}$ and $\Omega_L = 300°/\text{h}$

Now, Ω_L can be expressed as: $\Omega_L = 300°/\text{h} = \dfrac{300 \times 5 \times 10^3}{3600}\ \text{Hz} = 416.6\ \text{Hz}$

So, Faraday bias B_f = or > $5 \times 10^3 \times 100 + 416.6$ Hz = 500416.6 Hz

RLG errors

Typical RLG errors, which remain after elimination of lock-in, are as follows:

(i) *Bias:* The flow of the ions, after the discharge, has been characterised as Langmuir flow which in turn results in Fresnal drag which is a type of bias. This type of drag is neutralised with two-anode configuration as shown in Figure 3.35. The residual effect in this drag correction gives rise to bias. Sensitivity of the gas flow to thermal effect leads to bias instability.

(ii) *Noise:* Random walk type of noise is caused due to randomised mechanical dither and has been discussed earlier. Even in non-dithered four-frequency gyro, noise is present due to cavity losses and also due to residual spontaneous emission in lieu of stimulated emission. The magnitude and characteristics of these errors in a gyro are calibrated using suitable model and procedure as described in **[IEEE STD 647, 1995]**.

3.8.3 Fibre-Optic Gyro

The next engineering approach to successfully realise optical gyro based on Sagnac effect is called Fibre-Optic Gyro (FOG).

In FOG, phase measurement was implemented as an alternative to time measurement. For a beam of frequency f, the change in phase $\Delta\phi_s$ between the clockwise and counterclockwise beams is given by

$$\Delta\phi_s = \Delta t 2\pi f \tag{3.71}$$

Substituting for Δt, for a ring shaped Sagnac setup, the phase change becomes:

$$\Delta\phi_s = \frac{8\pi A f}{c_0^2}\Omega \tag{3.72}$$

For a circular optical path, the sensitivity is enhanced by the number of fibre turns N that has the effect of increasing area A by N times. The Sagnac phase shift is then given by

$$\Delta\phi_s = \frac{2\pi LD}{\lambda c_0}\Omega \tag{3.73}$$

where

D = mean coil diameter

λ = free space light source wavelength

L = total coil length that is equal to πND

c_0 = vacuum velocity of light

Ω = inertial rate component parallel to or passing through the centre of the circular coil axis

The gyro schematic is shown in Figure 3.40.

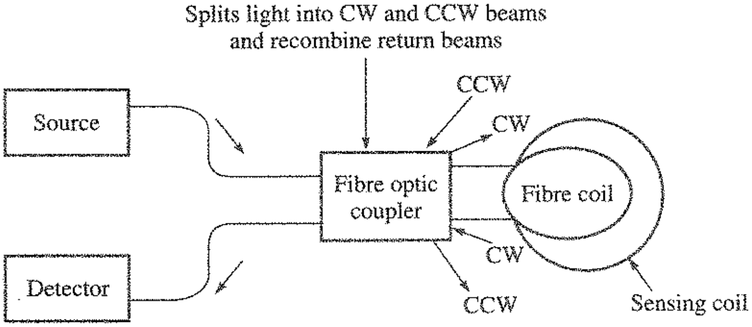


Figure 3.40 Operating principle of an interferometric fibre-optic gyro.

In the operation of FOG, the laser light from a source is split into two beams, CW and CCW, in a fibre-optic coupler that travel simultaneously in opposite directions through a reciprocal path in the fibre coil which encloses area NA, and then recombined again in the coupler. Recombination results in the superposition of the two waves, and as a result this type of gyro is widely known as Interferometric Fibre-Optic Gyro (IFOG).

The intensity response of the recombined waves with interferometer phase difference, is a cosine function (refer Figure 3.41), given by

$$I = I_m\,[1 + \cos\Delta\phi_s] \tag{3.74}$$

where I_m is the mean value of the light intensity.

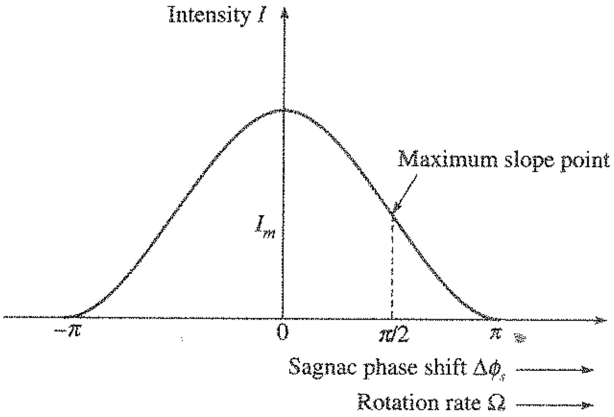


Figure 3.41 Basic intensity response of the interferometer.

It is thus seen that at zero rate, the sensitivity of intensity is zero, and the maximum sensitivity is at rates corresponding to ±$\pi/2$ phase shift operating points. Also, the intensity response is not sensitive to the change of direction of rotation around zero rate. It becomes a necessity that the intensity response is shifted (biased) by $\pi/2$ so that maximum sensitivity is obtained at zero rate.

When this is done, it leads to a sinusoidal response pattern with ability to detect the sign of rotation around zero rate, and this is shown in Figure 3.42. The photodetector measures the intensity to provide an electrical output, which consists of a DC component proportional to the average optical power incident on the photodetector, and an AC component proportional to the phase difference between the counter propagating light waves.

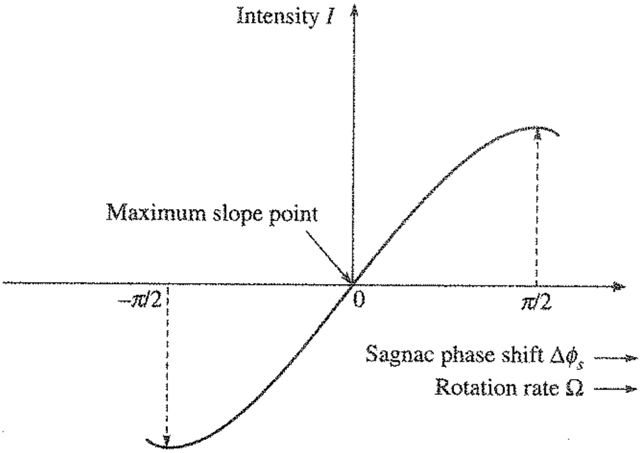


Figure 3.42 Biased IFOG intensity response with modulator.

This phase shift is the sum of two components:

(a) Rotation induced Sagnac phase shift

(b) Phase difference induced due to the phase modulator

Electronic processing takes care in removing the DC component and extraction of the rate signal. Also, it can be noted that with the sinusoidal response pattern, the open loop output range of IFOG is limited by the rate magnitude which produces a Sagnac phase difference of $\pm\pi/2$. Further, the actual operating range will be further limited by the non-linearity of the rate output.

The gyro open loop scale factor K_O is given by:

$$\Delta\phi_s = K_O\Omega \tag{3.75}$$

where

$$K_O = \frac{2\pi LD}{\lambda c_0} \tag{3.76}$$

A dimensional analysis of K_O shows that it has the unit of seconds. Further, it is seen that a better scale factor is obtained by increasing the number of turns and/or by changing the laser wavelength. This aspect is illustrated in Table 3.2 where $\Omega_{\pm\pi/2}$ is the operating rate range that is limited by the Sagnac phase shift of $\pm\pi/2$. The limitation of angular rate range in a high sensitivity gyro is overcome by closed loop technique that is described later.

Table 3.2 Open Loop IFOG Features and Performance

	L	D	λ	$\Omega_{\pm\pi/2}$	K_O
High sensitivity	1000 m	10 cm	1.55 μm	±66°/s	1.35 s
Lower sensitivity	200 m	3 cm	0.85 μm	±660°/s	0.15 s

EXAMPLE 3.8 (on IFOG rate sensitivity): An IFOG has the following features:

Coil diameter = 10 cm; No. of turns = 1200; $\lambda = 1.3$ μm

(i) Calculate its scale factor and the rate that gives phase shift of $\pi/2$.

(ii) Calculate Sagnac phase shift to detect inertial reference sensitivity of 0.05°/h.

Solution:

(i) Scale factor, $K_O = \dfrac{2\pi LD}{\lambda c_0}$

where $L = \pi \times 0.1 \times 1200$ m $= 376.8$ m

Substituting the values,

$$K_O = \frac{2\pi \times 376.8 \times 0.1}{1.3 \times 10^{-6} \times 3 \times 10^{8}} = 0.606 \text{ s}$$

$$\Omega_{\pm\pi/2} = \frac{\pi/2}{0.606} \text{ rad/s} = 136.5°/\text{s}$$

(ii) Sagnac phase shift to provide inertial reference sensitivity of 0.05°/h can be computed as

$$\Delta\phi_s = \frac{0.606 \times 0.05}{57.4 \times 3600} \text{ rad} = 1.46 \times 10^{-7} \text{ rad}$$

Detection of such a small phase shift requires that phase shifts, which are not due to angular rate, are eliminated by design. First requirement is to ensure 'reciprocity' in the gyro. Reciprocity means that for an optical wave in a linear medium, there is an opposite wave which propagates with exactly the accumulated phase and the same attenuation.

The initial experimentation showed that a fibre gyro was not intrinsically reciprocal which meant that even in the absence of rate, CW and CCW beams did not traverse identical path prior to their recombination. Eliminati‹ n of such errors forms the basis of designing and realising an IFOG which are described in the subsequent paragraphs.

IFOG features

The design feature of an all fibre open loop IFOG is described (refer Figure 3.43), which shows typical functional elements of the gyro.

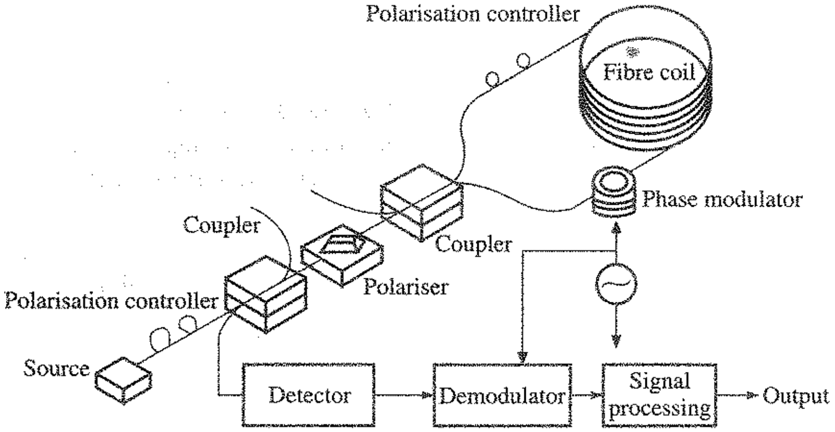


Figure 3.43 Features of an all fibre open loop interferometric fibre optic gyro.

This feature is commonly known as the minimum reciprocal all fibre configuration. A reciprocal optical path length is one whose properties are independent of the direction of propagation. Ensuring that light is fed in the fibre coil through a coupler and the returning wave is filtered through an identical coupler at the output solved the problem. Thus there are two *couplers* in IFOG to provide the reciprocity.

Source

The light in an IFOG is broadband (low coherence) in nature. During the initial phase of research in IFOG, high coherent laser beam, as used in RLG, was used. But it was not suitable as the light reflected at couplers, joints and other scatterings, swamped the rotation induced Sagnac signal. When a low coherence source, also called broadband source, is used, the backscattered light is incoherent and will not form fringes with the direct light.

Since early 1980s, Super Luminescent Laser Diode (SLD) has been in use as a source for IFOG. SLD light is broadband in nature. This is a p-n junction semiconductor emitter based on stimulated emission with amplification that is made insufficient for feedback oscillation to build up. Its typical wavelength is 0.85 μm with GaAlAs diode, but the wavelength has a high temperature sensitivity of around 400 ppm/°C that directly affects the gyro scalefactor. SLD sources with other wavelengths are also in use. The output power in SLD has increased over the years with current versions producing typically 10 mW to 20 mW. SLDs are costly as its characteristics are designed to suite gyros only. Semiconductor lasers are produced in bulk and hence it is cheaper. But for its use as a source in gyro, its coherency has to be sufficiently reduced by a method called 'frequency chirping'. It has similar temperature sensitivity problem and for its use in a navigation grade gyro, its temperature needs accurate control. In some design, thermo electric cooler controller is used that provides control with cooling as well as heating.

A more recent development is the use of Erbium-doped fibre, and such a source is called Super Fluorescent Fibre Source (SFS). Its high output power, typically between 30 mW and 50 mW, improves gyro signal-to-noise ratio and its wavelength temperature sensitivity is below 5 ppm/°C. Both these factors are useful in navigation grade performance. Its high output power is beneficial in designing for cost reduction through source sharing scheme, where three gyros, required in a system, are connected to a single SFS source.

Fibre coil

The characteristics of optical fibre in IFOG play significant role in gyro design and performance. These characteristic features have made IFOG fibres as something special and costly as compared to the optical fibres used in communication industry.

A standard communication optical fibre guides light by total internal reflection. It follows Snell's law, which defines the reflection and refraction of light in terms of angle of incidence and the media refractive indices. The fibre is a fine thread of glass to form the core with diameter comparable to the light wavelength. Protective cladding made of silica or pure quartz surrounds the core. Sharp bending and kinks cause loss of light. These features in optical communication fibre are common for IFOG also.

The special characteristic of IFOG fibre is in the use of *Single Mode Polarisation Maintaining type* (SMPM) as against considerable use of multimode fibres seen in communication. Multimode fibres support many spatial mode of light propagation, each with its own modal velocity. Coupling between these modes through light scattering creates non-reciprocity. Single mode is a special characteristic built in a gyro fibre, which allows only one mode at a particular light wavelength to propagate.

The next characteristic is the introduction of polarisation maintaining feature throughout the length of the fibre to reduce the polarisation-induced error. In a depolarised fibre coil, the polarisation related error in gyro is caused by light that is cross-coupled to the wrong polarisation state prior to the polariser, travels in the wrong state in part of the interferometer loop and then cross-couples back to interfere with gyro signal light. It erroneously appears as a gyro bias. It gives rise to bias instability due to both temperature and time dependency of this factor in the fibre. A polarisation maintaining fibre is in essence a single mode fibre that preserves the plane of polarisation of light coupled into it as the beam propagates through its length. These special features needed for IFOG considerably increases the fibre cost.

Use of polarisation maintaining fibre, is necessary in addition to configuring the gyro with *polarisers* which should suppress the second (orthogonal) polarisation mode. The optimum suppression is achieved by using the same polariser, also called single mode filter, as input and output ports of the sensor coil, thus giving a reciprocal configuration.

Couplers, which are used to initially split the light and then recombine the counter-rotating beams, also use SMPM fibres in all fibre gyro. Two such couplers are used for reciprocity requirement.

A ***phase modulator*** is used to maximise the rate detection sensitivity at the zero rate cross over point and which needs biasing the phase by $\pi/2$. Optically, this biasing is achieved by a scheme that makes path length of one beam longer with respect to other circulating beam. This is equivalent to creating a designed non-reciprocity. In the earlier versions of gyros, the light beams were mechanically phase modulated using a piezoelectric mandrel on which some fibre coil was wound and placed near the input coupler. An alternating electric field applied to the mandrel caused stretching and relaxing of this fibre coil. In the stretched condition, the path length increased and in the relaxing condition the path length became normal. By an appropriate choice of the modulating frequency of the mandrel, it was possible to shift the phase of the recombined beams by $\pi/2$. The result was the modulation of the beam phase.

The common ***detectors*** in use are PIN (acronym for p-doped, intrinsic and n-doped layers) photo detector diode that is either in Si or in GaAs or in InGaAs. A factor which influences the choice of semiconductor detector is the source wavelength as the detector material exhibits detection cutoff wavelength. Cutoff wavelength for Si detector is 1.13 μm and that for InGaAs is 1.65 μm.

The detector output, which is a modulated electrical signal, is demodulated and after necessary signal processing, provides output proportional to input rate.

For navigation grade performance, *closed loop* IFOG has become necessary to overcome some of the inherent limitations in an open loop IFOG and provide higher performance. A closed loop operation is implemented by nulling the rate induced Sagnac phase shift shown in Figure 3.44. As a result of this nulling operation, the gyro remains at the maximum slope point and the rate range increases substantially. Additionally, it offers reduction of non-linearity as well as noise in the output. Earlier, Table 3.2 has shown that high sensitivity needs long length of fibre. But the source power received at the photodetector decreases exponentially with the length ($e^{-\alpha L}$). This contradictory requirement sets limit on increasing the length.

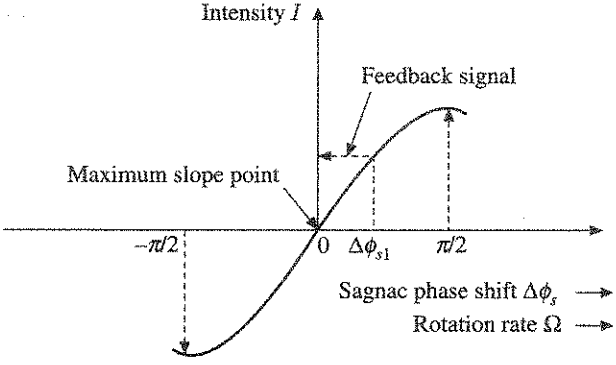


Figure 3.44 IFOG closed loop operation principle.

Errors

Even though, IFOG appears easier to produce or change the design in comparison with RLG, it has some complex set of errors which require close attention in order to improve performance. A few such errors are introduced to the readers.

A time dependent temperature gradient within the fibre coil sensor leads to gyro bias drift and the phenomenon is called Shupe effect, which is named after the inventor of this effect. Dipolar winding and quadrupolar winding have been developed to reduce the effect.

Further, when the gyro is subjected to magnetic field, it creates bias drift through a phenomenon known as magneto-optic Faraday effect. The direction of the magnetic field affects the velocity of the counter propagating beams in different manner causing the bias drift

Noise in IFOG is observed from photo detection process and the important noise mechanism is termed as *shot noise*. Photon *shot noise* appears when light energy, quantised in photons, falls on the detector to create electrical noise. This parameter limits the performance even in a high performance IFOG and is described further.

Shot noise in IFOG: Assume that the detector amplifier electronic noise is small, and the shot noise depends on detector current and the measurement bandwidth. It is also dependent on average optical power as the detector current is also dependent on this parameter. Shot noise limits the minimum detectable change in rate, which is dependent on the noise in the detector current. The noise in the detector current and the corresponding effect on the rate sensitivity can be expressed as [**Ghatak and Thyagaragan, 1999**]:

$$i_{\text{shot noise}} = \sqrt{2ei_d\Delta f} \tag{3.77}$$

Since

$$i_d = \frac{1}{2}\rho I_0$$

and

$$i_s = \frac{1}{2}\rho I_0 \Delta\phi_s$$

Expressing the SNR as the ratio of $i_s/i_{\text{short noise}}$, combining and simplifying, the shot noise limited phase shift can be expressed with unity SNR as:

$$(\Delta\phi_s)_{\min} = 2\sqrt{\frac{e\Delta f}{I_0\rho}} \tag{3.78}$$

Using Eq. (3.73), the minimum detectable rate, limited by shot noise, can be expressed as:

$$\Omega_{\min} = \frac{c_0\lambda}{\pi LD}\sqrt{\frac{e\Delta f}{I_0\rho}} \tag{3.79}$$

where

e = electron charge (1.6×10^{-19} coulomb)
i_d = average current generated by the detector
$I_0/2$ = mean optical intensity falling on the detector
i_s = signal current in the detector
Δf = bandwidth over which noise is considered
ρ = responsivity = ($\eta e/hf$); f is the frequency of the incident light
η = detector quantum efficiency

EXAMPLE 3.9 (on effect of shot noise on sensitivity): Compute

(i) shot noise limited $(\Delta\phi_s)_{\min}$ for a typical detector whose parameters are as follows:

$$I_0 = 0.1 \text{ mW}, \Delta f = 1 \text{ Hz}, \rho = 0.5 \text{ A/W}$$

(ii) Shot noise limited inertial rate of the high sensitivity gyro of Table 3.2.

Solution:

(i) The detectable Sagnac phase shift $(\Delta\phi_s)_{\min}$ is computed using Eq. (3.78) as follows:

$$(\Delta\phi_s)_{\min} = 2\sqrt{\frac{1.6\times10^{-19}\times1}{0.5\times0.1\times10^{-3}}} = 11.3\times10^{-8}\ \text{rad}$$

(ii) Taking the high sensitivity gyro parameters from Table 3.2 having

$$L = 1000\ \text{m},\ D = 10\ \text{cm},\ \lambda = 1.55\ \mu\text{m}$$

Putting the values in Eq. (3.79) and solving, we get:

$$\Omega_{\min} = \frac{3\times10^{8}\times1.55\times10^{-6}}{2\times3.14\times1000\times0.1}\times11.3\times10^{-8} = 8.37\times10^{-8}\ \text{rad/s} = 0.017°/\text{h}$$

There are other noise effects which tend to degrade the sensitivity further. The above example illustrates the necessity of shot noise reduction to get the benefit of long fibre coil length that is necessary for a high sensitivity IFOG. Use of high optical power source decreases the shot noise component of the random noise.

IFOG advantages

A major advantage in an IFOG lies in the feasibility that the sensor offers to change the performance by changing the length of fibre coil. Such a design change can offer various grades of performance needed in aerospace. For example, typical fibre coil length is >1.5 km for navigation grade sensors and is only 100 m for a gyro with considerably reduced performance or for use as a rate sensing sensor. Some of the desirable advantages in IFOG are listed as:

(i) High rate capability >1000°/s

(ii) No lock in problem

(iii) Immune to both radio frequency and electromagnetic (EMI) interferences

(iv) Negligible g and g^2 sensitivity

(v) No moving parts and gas, leading to high operational reliability and long shelf life

(vi) Fast activation time

3.9 Gyro based on Cold Atom Interferometry

One of the most fundamental and revolutionary ideas of quantum mechanics is that matter and light have a wave and particle nature. This postulate can be verified by passing a beam of electrons through a diffraction grating. The electrons, often considered particles, will produce a wave-like interference pattern. The equation linking the wave and particle nature of matter was suggested by de Broglie who formulated:

$$\lambda = \frac{h}{p} \tag{3.80}$$

where

λ = de Broglie wavelength
h = Planck's constant
p = linear momentum

Thus, matter having momentum p will have a specific de Broglie wavelength. Conversely, an electromagnetic wave having wavelength λ should have a specific linear momentum of h/λ.

Cold atom waves propagate freely in an atom interferometer, forming fringes whose location is fixed in inertial space. The atom interferometer will have similar functions like an optical interferometer such as, splitting, reflecting and mixing. To execute these functions, a sequence of laser pulses is required [**Fox, 2006**], and the scheme is shown in Figure 3.45.

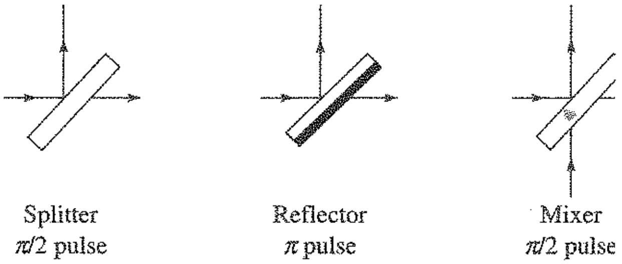


Figure 3.45 Atom interferometry scheme.

To understand this approach, which is based on quantum physics, it is reminded that photons carry momentum even though it has no rest mass. When an atom absorbs or emits a photon, its momentum changes accordingly. Therefore, one starts with a $\pi/2$ laser pulse that puts the atom in an equal superposition of the ground and excited states. While the excited state of the atom changes its momentum due to photon absorption, the ground state remains unchanged thus accomplishing the cold atom wave beam splitting. Similarly, a π-laser pulse exchanges the states, functioning as a mirror in redirecting the atom wave. Therefore, a sequence of $\pi/2 - \pi - \pi/2$ pulses makes up an interferometer.

If such interferometer is rotating (refer Figure 3.46), Path A and Path B become unequal due to Sagnac effect, and consequently the fringes appear shifted in proportion to the rotation rate.

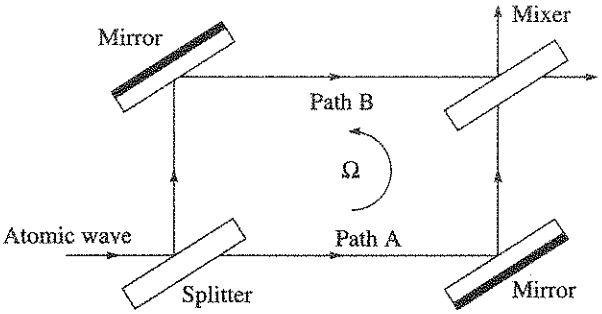


Figure 3.46 Atom interferometry showing Sagnac effect.

For atom particle velocity v and wavelength λ, the Sagnac phase shift for an enclosed loop area A is given as:

$$\Delta\phi_s = \frac{4\pi}{\lambda v}(A\Omega) \tag{3.81}$$

For atom, the wavelength, $\lambda = h/mv$, while for light, $\lambda = 2\pi c/\omega$. The phase shift expression in Eq. (3.81) is for particles interfering after traversing half of the loop. It may be noted that in the original Sagnac optical interferometry, the interference takes place after traversing the complete loop resulting in an additional factor of 2. Here also, it is more appropriate to express the bracketed terms as vector dot product $(\mathbf{A}\cdot\mathbf{\Omega})$ to convey the information that the perpendicular axis defining the plane of the interferometer area is parallel to the rotation rate vector for maximum sensitivity.

For laser light like interferometry, the atoms need to be cooled to micro kelvin region. The probable atom velocity v_{atom} at an absolute temperature T is given by the expression:

$$v_{\text{atom}} = \sqrt{\frac{3k_B T}{m}} \tag{3.82}$$

where

k_B = Boltzmann constant (1.33×10^{-23} J/K)

m = mass of the atom

For example, if cesium atom beam is used ($m = 133 \times 1.6 \times 10^{-27}$ kg) and the temperature is reduced by laser cooling to 100×10^{-6}°K, then the probable atom velocity v_{atom} will be 0.013 m/s.

This reduction results in a significant increase in the theoretical rotational sensitivity of this atom interferometer, which is about ten orders of magnitude larger compared to the sensitivity of a corresponding interferometer that uses light. Theoretically, the increase in sensitivity over the photon based optical gyro has been defined by a factor given by

$$\frac{\phi_{\text{atom}}}{\phi_{\text{photon}}} = \frac{mc^2}{\hbar\omega} \approx 10^{10} \tag{3.83}$$

where

m = atom mass

ω = photon frequency

$\hbar$ = reduced Planck's constant

This gain factor assumes that both types of gyros have same Sagnac area. This large gain assumption is not currently valid as the Sagnac area that can be realised now in cold atom gyro is quite low as compared to an optical gyro. The science and technology involves cooling of atoms, generation of coherent atom beams, atom interferometry and detection of fringe shift. A recent review paper **[Fang and Qin, 2012]** have described the current status on atom interferometric gyro and the direction of research to improve upon the performance and the packaging technology including chip scale gyro realisation. The paper also points out renewed R&D on nuclear magnetic resonance gyro.

Some current approaches **[Titterton et al., 2004]** towards a laboratory prototype of such cold atom gyro is described, while the gyro performance realised has been described by **[Gustavson et al., 1997]** so that such gyros can be produced to provide standalone high precision performance in navigation application where satellite navigation will not be suitable.

SUMMARY

This chapter on gyros has brought out the principle of operation, their features, configurations and error characteristics for a range of gyros which are currently under use primarily in aerospace and to some extent in ships as well as for land based application. This range covered the spinning rotor gyros consisting of single degree of freedom rate and rate integrating gyros as well as the two degrees of freedom gyros where specifically dynamically tuned gyro still finds considerable application. The other end of these gyros are normally called solid state gyros and the gyros covered in this range consisted of Coriolis vibratory gyro and the optical gyros. The Coriolis vibratory gyro has special significance due to the emergence of micro-fabrication methods and tuning fork gyro has been described at length. In this category, the most accurate gyro, called hemispherical resonator gyro, is addressed due its excellent performance and long reliability even though the gyro is not fabricated by the micro fabrication method. The optical gyros are currently dominating in navigation application. So, the ring laser gyro as well as the fibre optic gyro are described at length. Lastly, the great stride on the research and development on cold atom gyro has made it necessary to introduce the basics of this futuristic gyro.

EXERCISES

3.1 A perfectly balanced spinning rotor of a gyro is rotating in the direction shown in Figure E3.1.

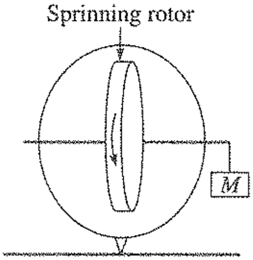


Figure E3.1

What happens when a mass *M* is added at the end as shown? Explain your choice.

(i) The gyro topples
(ii) The gyro precesses about the vertical axis
(iii) The gyro precesses about the horizontal axis
(iv) Not enough information to answer [**Ans.** (ii)]

3.2 Sketch of a spinning rotor gyro is shown in Figure E3.2.

(i) Assume normal gyro parameters, draw its block schematic and derive its input–output relation. What type of gyro it is called?

(ii) Suppose the torsion bar is removed and the gimbal is mounted to the gyro case on a friction less bearing, what will be the gyro called? Assuming normal gyro parameters, write down its input–output functional relation.

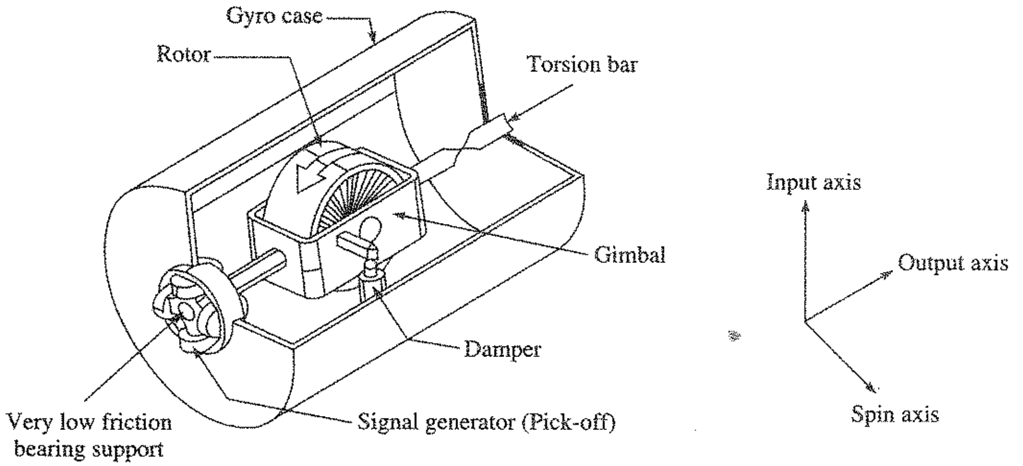


Figure E3.2

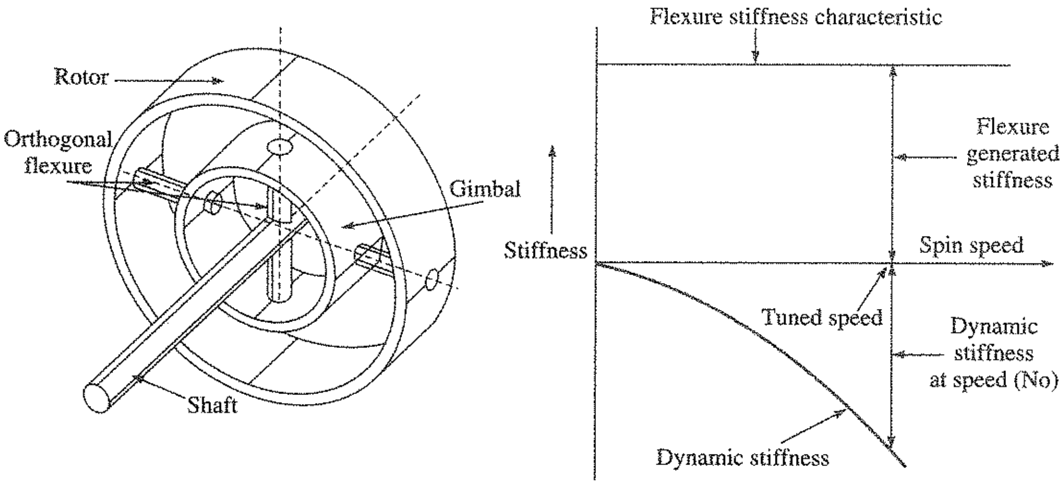
3.3 There are two figures shown [Figure E3.3(a)] below, one showing the feature of a gyro and other [Figure E3.3(b)], its important characteristic. Answer the following:

(i) What is the gyro called and why?

(ii) What is the tuned speed equation for this gyro?

(iii) How many degrees of freedom the rotor has?

(iv) Will the gyro have nutation frequency? If so, how this can be solved?


Rotor
Orthogonal flexure
Gimbal
Shaft
Flexure stiffness characteristic
Flexure generated stiffness
Stiffness
Spin speed
Tuned speed
Dynamic stiffness at speed (No)
Dynamic stiffness

(a) (b)

Figure E3.3

3.4 A DTG requires its bias drift alone to be calibrated. A set-up, shown below (Figure E3.4), is used to calibrate the gyro X-axis bias with known value of gyro scale factor.

(i) Write down the model equations and show how the bias is obtained?

(ii) Do you think the set-up is ok for gyro Y-axis bias calibration?

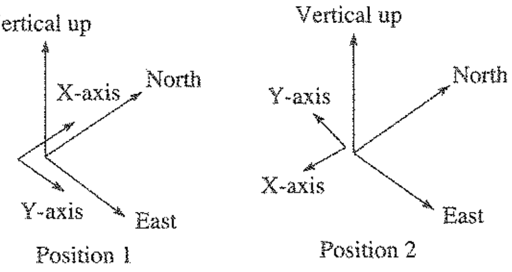


Figure E3.4

3.5 Figure E3.5 shows a pendulum bob of mass M suspended from a frictionless support with two degrees of freedom at latitude of 30° and assume free from air resistance.

(i) If you initiate the pendulum movement at time t_0 along North–South direction, what do you notice on the direction of pendulum movement after t_0? Explain.

(ii) If the above set-up is made at equator, what do you observe?

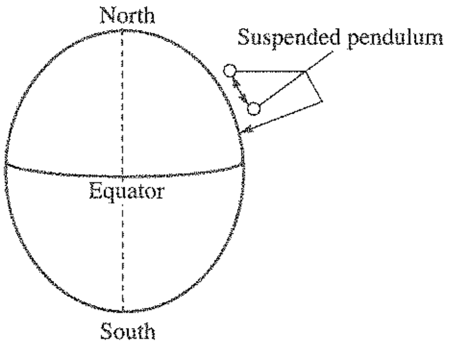


Figure E3.5

3.6 Sketch of a CVG is shown in Figure E3.6.

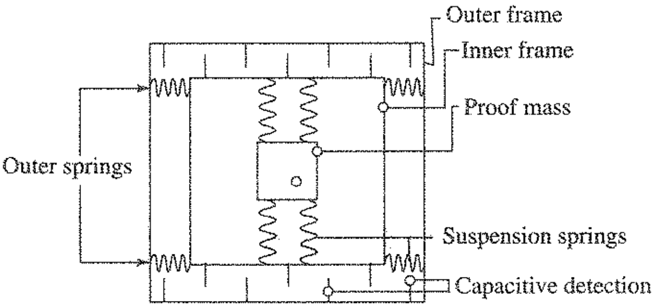


Figure E3.6

Indicate:

(i) Resonator vibration direction

(ii) Direction of input angular rate

(iii) Direction of Coriolis force

(iv) Also explain how the gyro is detecting angular rate

3.7 In an HRG, the resonator has following vibration modes:

(a) $n = 0$ (b) $n = 1$ (c) $n = 2$

(i) Which is the mode suitable for operation as a gyro?

(ii) Sketch a gyro operating mode and show on it the nodes, antinodes and the velocity direction at these nodes and antinodes.

(iii) A counterclockwise rate is applied to this resonator. Show the direction of Coriolis force at any one of the antinodal points. [**Ans.** (c)]

3.8 Write the rate sensitivity equation for a square block RLG and define the parameters. If the square block has a side of length 7 cm and operating on laser wavelength of 0.63 μm,

(i) Calculate frequency shift to detect angular rate of 0.01°/h

(ii) Calculate its scale factor in arc-s/pulse.

(iii) What will be the path length for a triangular gyro to achieve the same scale factor operated with same wavelength?

[**Ans.** (i) 5.33×10^{-3} Hz, (ii) 1.85 arc-s/pulse, (iii) 36.5 cm]

3.9 Rate sensitivity equation of a basic IFOG is given as:

$$\Delta\phi_s = \frac{8\pi A f}{c_0^2}\Omega$$

The gyro has following design parameters:

Coil diameter: 10 cm, No. of turns = 1200, $\lambda = 1.3$ μm

Find out (i) gyro scale factor (ii) Sagnac phase shift for a rate of 100°/s

[**Ans.** (i) 0.606 s, (ii) 1.055 rad]

3.10 In a basic IFOG, the intensity response of the recombined waves with interferometer phase difference is given by

$$I = I_m\,[1 + \cos \Delta\phi_s]$$

where $\Delta\phi_s$ is the Sagnac phase shift.

(i) Plot the function and explain the problem with this type of gyro.

(ii) (a) What is done to circumvent the problem? Explain with a sketch.

(b) What is the maximum angular rate range achievable without ambiguity?

REFERENCES

Armenia, Mario N., Caterina Ciminelli, Francesco Dello'lio, and Passaro, Vittoria M., *Advances in Gyroscope Technologies*, Springer-Verlag, Berlin Heidelberg, 2010.

Aronowitz, Fredrick., *Fundamentals of Ring Laser Gyro, Optical Gyros and Their Applications, AGARDOGRAPH*, **339**, 1999.

4

Accelerometer

Accelerometer is used for wide range of applications. Such areas of application cover aircrafts, missiles, launch vehicles and spacecrafts which are collectively categorised as aerospace. Besides these, there exist large non-aerospace applications. As a result, the features, design and performance of accelerometer varies widely. This chapter has concentrated primarily on accelerometers for use in aerospace.

Accelerometers are designed to actually measure inertial force by converting the applied inertial acceleration with the help of a proof mass that is built into the instrument. This inertial force is then further processed using different schemes to derive the imparted acceleration. Various types of accelerometers, which are either currently in operation for aerospace or under development, have adopted their names from these schemes. Some such schemes are named as:

1. Open loop spring–mass accelerometer
2. Closed loop spring–mass accelerometer
3. Vibrating beam accelerometer
4. Interferometric fiber optic accelerometer
5. Electron tunneling effect accelerometer

A new accelerometer, that is emerging, uses totally different principle of operation and is called cold atom interferometer based accelerometer.

Current operational accelerometers for aerospace are mostly using first three schemes, while the remaining two accelerometers are under development. Cold atom interferometer based accelerometer has targeted for very high accuracy stand alone inertial navigation application where satellite navigation aiding will not be possible.

4.1 Operating Principle

Accelerometer operates by measuring the inertia force generated when a mass accelerates. The simplest form of a mass–spring accelerometer, which measures the inertia force, is shown in

Bose, A., Puri, S.N., and Banerjee, P., *Modern Inertial Sensors and Systems*, PHI Learning, Delhi, 2008.

Bryan, G.H., On the Beats in the Vibration of a Revolving Cylinder or Bell, *Proceedings of the Cambridge Philosophical Society*, Vol. VII, pp. 101–111, November 24, 1890.

Craig, R.J.G., Theory of Operation of an Elastically Suspended Tuned Gyroscope, and Theory of Errors of a Multigimbal Elastically Supported Tuned Gyroscope, *IEEE Transaction on Aerospace and Electronic Systems*, AES-8, pp. 280–297, May 1972.

Faucheux, M., Fayoux, D., and Roland, J.J., The Ring Laser Gyro, *Journal Optics* (Paris), **19**(3), pp. 101–115, 1988.

Fox, Mark, *Quantum Optics: An Introduction*, Oxford University Press, USA, 2006.

Ghatak, A., and Thyagaragan, K., *Introduction to Fiber Optics*, Cambridge University Press, First South Asian Edition, New Delhi, 1999.

Gustavson, T.L., Bouyer, P., Kasevich, M.A., Precision Rotation Measurements with an Atom Interferometer Gyroscope, *Physical Review Letters*, **78**(11), pp. 2046–2049, 1997.

Howe, E.W. and Savet, P.H., *The Dynamically Tuned Free Rotor Gyro, Control Engineering*, pp. 67–72, June 1964.

IEEE STD 647 on IEEE Standard Specification Format Guide and Test Procedure for Single-Axis Laser Gyros, 1995.

Jian Cheng Fang and Jie Qin, Advances in Atomic Gyroscopes: A View from Inertial Navigation Applications, *Sensors*, **12**, 6331–6346, 2012.

Lawrence, Anthony, *Modern Inertial Technology*, 2nd ed., Springer–Verlag, New York, 1998.

Loper, E.J. and Lynch, D.D., Vibratory Rotation Sensor, U.S. Patent 4, 951, 508, 28 August, 1990.

Lynch, D.D., Vibration-Induced Drift in the Hemispherical Resonator Gyro, *Proc. 43rd Annual Meeting of The Institute of Navigation*, Dayton, OH, pp. 34–37, 23–25 June, 1987.

Lynch, David D., "Coriolis Vibrating Gyros", Symposium Gyro Technology 1998, Stuttgart, Germany.

Rosellini, L. and Caron, J.M., REGYS-20; A Promising HRG-based IMU for Space Application, GNC-2008, *Seventh International ESA Conference on Guidance, Navigation and Control Systems*, 2–5, Tralee, Ireland, June 2008.

Rudloff, R., Physical Background and Technical Realisation, *RTO AGARDOGRAPH* **339**, May 1999.

Titterton, D.H. and Weston, J.L., Strapdown Inertial Navigation Technology, 2nd ed., **207**, Progress in Astronautics and Aeronautics, Co published by *AIAA-USA and IEE*, United Kingdom, 2004.

Volk, C.H., et al., Multioscillator Ring Laser Gyroscopes and Their Applications, *RTO AGARDOGRAPH*, (339), May 1999.

Figure 4.1. The operating principle is described for such an open loop spring–mass accelerometer that is historically one of the earliest.

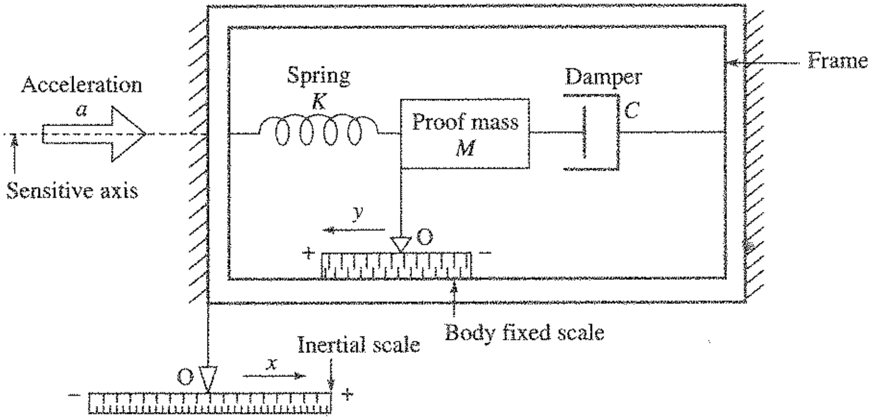


Figure 4.1 A spring–mass accelerometer schematic.

When a moving vehicle puts the frame of the accelerometer under acceleration of magnitude a, the spring deflects until it produces enough force as per Hooke's law to accelerate the proof mass M at the same rate as the frame. Physically, this means that for the acceleration direction as shown in the diagram, the spring will be under compression. If the direction of acceleration is reversed, the spring will be under tension. Such an accelerometer can be described mathematically in the following manner. Let

x = displacement of the accelerometer frame with respect to inertial reference scale
y = displacement of the proof mass M with respect to the accelerometer body fixed scale
C = damper damping coefficient
K = axial spring stiffness

The force balance equation can be written as:

$$M\frac{d^2}{dt^2}(y-x)+C\frac{dy}{dt}+Ky=0 \tag{4.1}$$

Rearranging the equation, we get:

$$M\frac{d^2y}{dt^2}+C\frac{dy}{dt}+Ky=M\frac{d^2x}{dt^2}=Ma \tag{4.2}$$

Under steady state condition, when the derivatives of y go to zero, we get:

$$Ma = Ky$$

$$a=\frac{K}{M}y \tag{4.3}$$

Since K and M are accelerometer design constants, a measurement of y provides the value for the input acceleration. Equation (4.3) is important to underline the essential characteristic of this class of accelerometer where accurate measurement of linear acceleration is of fundamental importance and not vibration or shock. The static sensitivity and the time response to reach a steady state value are governed by the following parameters:

Static sensitivity $\frac{y}{a} = \frac{M}{K}$

Undamped natural frequency $\omega_n = \sqrt{\frac{K}{M}}$

Damping ratio $\zeta = \frac{C}{2\sqrt{KM}}$ (4.4)

Equation (4.2) is often reformulated using ω_n and ζ to give:

$$\frac{d^2y}{dt^2} + 2\zeta\omega_n \frac{dy}{dt} + \omega_n^2 y = a \tag{4.5}$$

Figure 4.2 shows the transient response of the accelerometer where $\zeta < 1$. It shows that damping is essential to provide a steady state response within a short time which is dictated by navigation sampling time of accelerometer data.

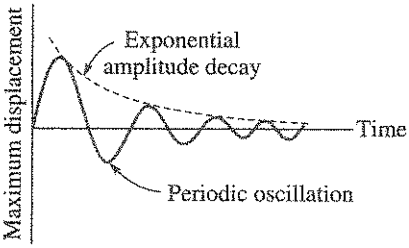


Figure 4.2 Accelerometer transient response.

EXAMPLE 4.1 (on open loop accelerometer): An accelerometer with proof mass of 2 × 10^{-3} kg is having natural frequency of 150 Hz and damping ratio of 0.7. Find out its damping coefficient and also the displacement of the proof mass relative to case to measure acceleration of 10g. (Assume g = 9.78 m/s^2).

Solution: Stiffness K can be obtained from Eq. (4.4). On substitution, we get:

$$2\pi \times 150 = \sqrt{\frac{K}{2 \times 10^{-3}}}$$

Solving, K = 1774.7 N/m.

Damping coefficient C can be obtained from Eq. (4.4). On substitution, we get:

$$0.7 = \frac{C}{2\sqrt{1774.7 \times 2 \times 10^{-3}}}$$

Solving, C = 2.63 N/m/s.

Proof mass displacement is obtained from Eq. (4.3). On substitution, we get:

$$y = \frac{2 \times 10^{-3} \times 10 \times 9.78}{1774.7} \text{ m}$$

Solving, $y = 110\ \mu\text{m}$.

4.2 Accelerometer Classification

Accelerometers are classified depending on its design and built. Quite often it is classified based on the method of sensing the proof mass displacement. The most widely used sensing methods are:

(i) Inductive (ii) Capacitive (iii) Optical (iv) Interferometric (v) Electron tunneling (vi) Piezoelectric and (vii) Piezoresistive.

Another method of accelerometer classification is based on open loop sensing or closed loop sensing. If the detector or the pick-off output is a direct measure of the acceleration, it is called *open loop accelerometer.* Whereas, if the detector output is nulled by a force feedback system, and the feedback current or voltage is a direct measure of the acceleration, then it is termed as *closed loop accelerometer.*

Another classification is based on the suspension of the proof mass. If the proof mass is having a lever arm, it is called pendulous accelerometer, whereas, if there is no lever arm, it is called *non-pendulous accelerometer* or *translatory mass accelerometer.*

In aerospace applications, where accelerometers need to be quite accurate, IEEE has defined the classification by combining the last two and in such classification, Figure 4.1 shows *open loop non-pendulous accelerometer*, Figure 4.3 shows open *loop pendulous accelerometer* and Figure 4.5 shows *closed loop pendulous accelerometer.*

4.3 Open Loop Pendulous Accelerometer

In open loop pendulous accelerometer, the proof mass of mass M is placed at a distance l, known as lever arm, from the centre of suspension, as shown in Figure 4.3. Pendulosity of the

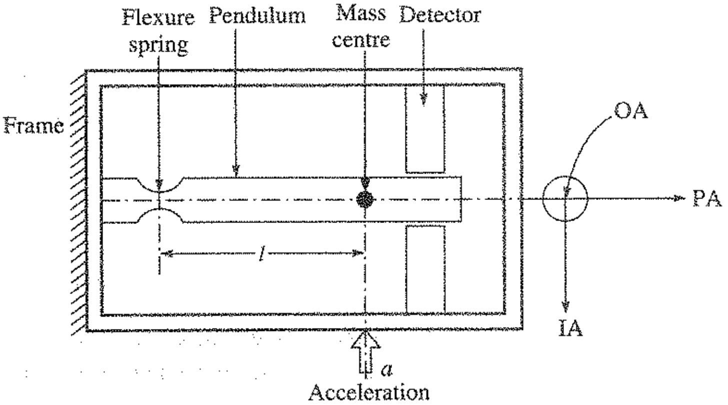


Figure 4.3 Open loop pendulous accelerometer schematic.

suspended mass is then given by Ml. Inertial acceleration acting on the proof mass produces a reaction torque about the suspension point. The suspension is normally provided by a flexure spring. A flexure spring avoids friction and is normally used against other form of suspension, e.g. jeweled bearing, used in the early days. Accelerometer damping is not explicitly shown here, although some form of damping will be there to shape the transient response.

Under steady state, acceleration a acting on the accelerometer leads to generation of inertial torque Mla that is balanced by the rotational spring stiffness K. This produces a relative deflection angle θ between the accelerometer frame and the proof mass. The torque balance equation is given by

$$(Ml)a = K\theta \tag{4.6}$$

Comparing Eq. (4.6) with Eq. (4.3), it is seen that the sensitivity of the pendulous accelerometer increases by a factor l which is the lever arm length, to highlight one advantage of a pendulous configuration over the non-pendulous version. The detector measures the deflection to provide an electrical output. The damping is either provided with a fluid or with some inert gas. The orthogonal sensor axes are defined as:

IA = input axis (sensitive axis), along which acceleration is acting
PA = pendulum axis
OA = output axis (axis about which rotational spring deflection takes place)

Axes, defined by IA, PA, OA, are normally conforming to a right handed system.

4.3.1 Cross Coupling Effect

Open loop configuration, although simple, has not been able to provide navigation grade performance due to deficiency inherent in this design in areas such as cross axis coupling and linearity.

Cross axis coupling is an indication that the accelerometer is sensitive to acceleration along the axis perpendicular to the input axis. In an open loop pendulous accelerometer, the cross axis sensitivity is high and manifestation of this error is described with the help of Figure 4.4.

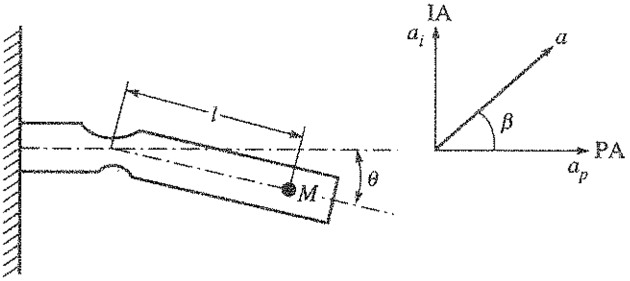


Figure 4.4 Cross coupling error in pendulous open loop accelerometer.

Assume that an acceleration a is acting on the accelerometer so that it has a component a_i along IA and a_p along PA. The torque T acting on the accelerometer proof mass is given by

$$T = Mla_i \tag{4.7}$$

Deflection θ is then given by

$$\theta = \frac{Mla_i}{K} \tag{4.8}$$

Due to this finite deflection, the acceleration component a_p sin θ acts on the input axis of the accelerometer to produce cross coupling error torque T_p, given by

$$T_p = a_p \sin \theta Ml \tag{4.9}$$

For small deflection, sin $\theta = \theta$, and combining Eqs. (4.8) and (4.9) we get:

$$T_p = \frac{M^2 l^2}{K} a_i a_p \tag{4.10}$$

$$= K_{ip} a_i a_p \tag{4.11}$$

where $K_{ip} = M^2l^2/K$ is cross coupling coefficient with unit as N-m per (m/s^2)^2. However, unit of K_{ip} is normally expressed in terms of acceleration such as g/g^2 or m/s^2/(m/s^2)^2 and this is possible by expressing the cross coupling coefficient as (Ml/K).

From Eq. (4.10), it is thus seen that the spring stiffness K needs to be high to reduce cross coupling coefficient magnitude, but high K leads to reduction of sensitivity. There are several other design issues which are difficult to sort out for improved performance in open loop configuration.

EXAMPLE 4.2 (on cross coupling error of open loop accelerometer): Calculate the spring stiffness necessary to limit the pendulum deflection to 15 arc-min when the input acceleration is 10g for a proof mass of 2 × 10^{-3} kg and pendulum length of 2 × 10^{-2} m. Find out the cross coupling coefficient. Assume g = 9.78 m/s^2.

Solution: Spring stiffness K can be obtained using Eq. (4.8). On substitution, we get:

$$K = \frac{10 \times 9.78 \times 2 \times 10^{-2} \times 2 \times 10^{-3}}{15} \text{ N-m per arc-min}$$

Solving, K = 2.6 × 10^{-4} N-m per arc-min or 0.89 N-m per radian.

Cross coupling coefficient K_{ip} can be obtained using Eq. (4.10). On substitution, we get:

$$K_{ip} = \frac{(2 \times 10^{-3} \times 2 \times 10^{-2})^2}{0.89}$$

Solving, K_{ip} = 17.97 × 10^{-10} N-m/(m/s^2)^2.

Expressing cross coupling coefficient as Ml/K, we get:

$$K_{ip} = \frac{(2 \times 10^{-3} \times 2 \times 10^{-2})}{0.89} = 4.49 \times 10^{-5} \text{ m/s}^2/(\text{m/s}^2)^2$$

$$\text{with } g = 9.78 \text{ m/s}^2, K_{ip} = \frac{4.49 \times 10^{-5}}{9.78} = 0.45 \times 10^{-5} \text{ g/g}^2$$

In a non-pendulous accelerometer, there is no defined pendulous axis and the input axis is parallel to output axis. So, cross coupling is theoretically zero and can be considered as an advantage from performance point of view.

4.4 Closed Loop Pendulous Accelerometer

In a closed loop design of pendulous accelerometer, the proof mass movement is constrained by an electrical force and by design; the inherent flexure spring force is considered as negligible. In conventional accelerometer design, this electrical force is generated by an electromagnetic torquer.

4.4.1 Sensitivity

Balancing the acceleration induced inertial force by an electrical force, we get under steady state:

$$Mla = K_t i \tag{4.12}$$

where

i = feedback current

K_t = torquer constant

So, sensitivity can be defined as:

$$\frac{\text{Output}}{\text{Input}} = \frac{i}{a} = \frac{Ml}{K_t} \tag{4.13}$$

This type of accelerometer is often referred as *torque to balance pendulous accelerometer*. The closed loop design keeps the pendulum deflection near to its null (zero value) under full operating acceleration range thereby substantially reducing the cross coupling error as well as the linearity error associated with pick-off output.

EXAMPLE 4.3 (on torque constant): For the accelerometer of nominal design parameters and the operation range as shown in Example 4.1, find out the torque constant if the current required is 50 mA for rebalance operation.

Solution: The torque constant can be found out using Eq. (4.13). On substitution, we get:

$$K_t = \frac{2\times10^{-3}\times2\times10^{-2}\times10\times9.8}{50} = 7.8\times10^{-5}\ \text{Nm/mA}$$

4.4.2 Design Features

Features of a typical closed loop accelerometer consist of the following functional elements which are shown in Figure 4.5.

(i) Detector (ii) Forcer (iii) Flexure and pivot suspension (iv) Damping (v) Rebalance servo

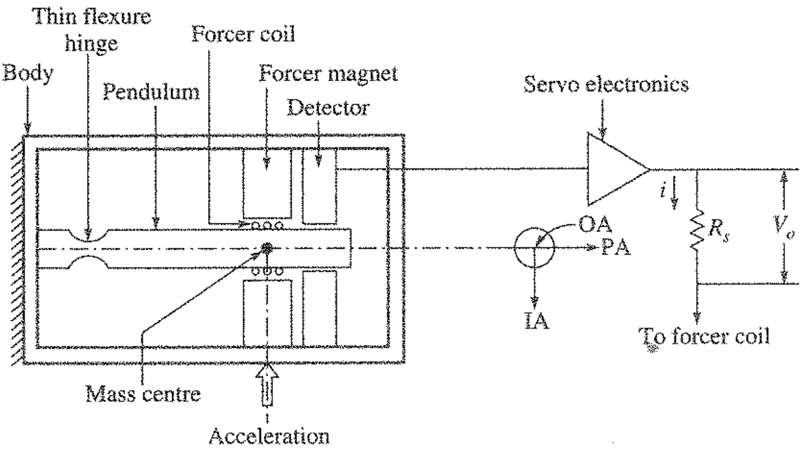


Figure 4.5 Torque to balance pendulous accelerometer schematic.

Detector

A detector senses the rotation or the linear translation of the proof mass, relative to the accelerometer case when acceleration acts along the input axis.

Common detectors, quite often termed as pick-offs, are either inductive or capacitive or optical in nature. Accelerometers based on piezoelectric or piezoresistive detection have not been successful in realising precision grade.

The inductive pick-offs are normally air core type to indicate that there is absence of soft iron material in the magnetic path. In a typical design, there are four coils which are connected in Wheatstone bridge configuration. When the pendulum is centred, i.e. under no input acceleration, the bridge output is zero. Change of air gap, due to an input acceleration, causes a change in the bridge balance, which in turn produces an output. The bridge is typically excited with sine wave with voltage amplitude between 2 V and 6 V and frequency >15 kHz. Sensitivity is moderate requiring suitable amplification, while output voltage phase shift is high; both are due to air core design. Output phase shift is reduced to an acceptable level using tuning capacitor.

Capacitive pick-off, a later induction in accelerometer design, uses a pair of capacitors C_1 and C_2 so that they are on either side of the pendulum (refer Figure 4.6).

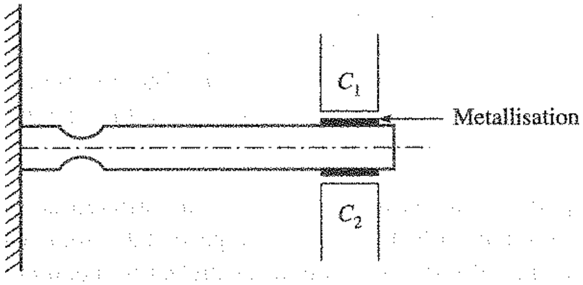


Figure 4.6 Capacitive pick-off in accelerometer.

Outputs of these two capacitors are then used in differential mode so that when the pendulum is centred under no input acceleration, the net capacitor output is zero. In the presence of input acceleration, the capacitance of one increases, while that of the other decreases, thus providing a useful output for the amplifier. In some design, the capacitors are used in Wheatstone bridge configuration. The capacitors are primarily of parallel plate type with area A and with nominal design gap d. Nominal capacitance C_0 under zero acceleration is given by

$$C_0 = \varepsilon_g \frac{A}{d} \tag{4.14}$$

where ε_g is gap permittivity ($= \varepsilon_0 \varepsilon_r$); ε_0 is free space permittivity and ε_r is relative gap permittivity.

Under an input acceleration, there is a gap change Δd, which results in a change of nominal capacitance C_0 to assume the values of C_1 and C_2 as follows:

$$C_1 = \varepsilon_g \frac{A}{d} \frac{1}{(1 + \Delta d/d)} = \frac{C_0}{1+k} = C_0(1+k)^{-1} \tag{4.15}$$

$$C_2 = \varepsilon_g \frac{A}{d} \frac{1}{(1 - \Delta d/d)} = \frac{C_0}{1-k} = C_0(1-k)^{-1} \tag{4.16}$$

Expanding by series expansion Eqs. (4.15) and (4.16) and taking the difference, we get:

$$C_1 - C_2 = 2C_0[k + k^3 + k^5 + \cdots] \tag{4.17}$$

Equation (4.17) shows that while even order nonlinearities are absent by difference process employed, the output contains odd order nonlinearities. The linear pick-off scale factor S_f is thus given by

$$S_f = 2C_0 k \tag{4.18}$$

where $k = \dfrac{\Delta d}{d}$.

Sensitivity of capacitive detection is low compared to its inductive counterpart but is electronically taken care in the current accelerometer designs.

In some accelerometer design, optical pick-offs have been used with success. In a typical design, a projection of the pendulum comes between a Light Emitting Diode (LED) and a divided silicon photo detector. The shadow of the pendulum falls on the detector such that at zero acceleration equal voltages are outputted from the two halves of the detector. Under an input acceleration the pendulum moves, the shadow also moves such that for one half of the detector the voltage rises while for the other half, it falls. The difference output is then used by the servo to restore the pendulum position. It is important in this design to ensure that the LED intensity does not change on its own.

EXAMPLE 4.4 (on capacitive detection): Find out the differential capacitance in a parallel plate capacitor system of Figure 4.5, for a gap change of 0.01 mm where the capacitor area is 25 mm^2 and a nominal gap of 0.05 mm with air as dielectric medium. Find out the detection sensitivity.

Solution: Nominal capacitance C_0 is given by Eq. (4.14). Now,

$$\varepsilon_g = \varepsilon_0\varepsilon_r = 8.854 \times 10^{-12} \times 1.000589 = 8.859 \times 10^{-12} \text{ F/m}$$

On substitution, we get:

$$C_0 = 8.859 \times 10^{-12} \times \frac{25 \times 10^{-6}}{0.05 \times 10^{-3}} \text{ F}$$

solving, $C_0 = 4.429$ pF

Therefore,
$$C_0 = 4.429 \text{ pF} \quad \text{and} \quad k = \frac{0.01}{0.05} = 0.2$$

Differential capacitance $C_1 - C_2$, can be obtained from Eq. (4.17). On substitution, we get:

$$C_1 - C_2 = 2 \times 4.429\,[0.2 + 0.2^3 + 0.2^5 + \cdots]$$

or

$$C_1 - C_2 \sim 1.76 \text{ pF}$$

$$\text{Detection sensitivity} = \frac{1.76}{0.01 \times 10^{-3}} \text{ pF/m} \quad \text{or } 0.176 \text{ pF/}\mu\text{m}$$

Quite often detection sensitivity is an important design parameter.

Forcer

The most common type of forcer, which is in use, works on the principle of a D'arsonval principle where magnetic field set up by permanent magnets interacts with coils carrying current. The scheme is shown in Figure 4.7.

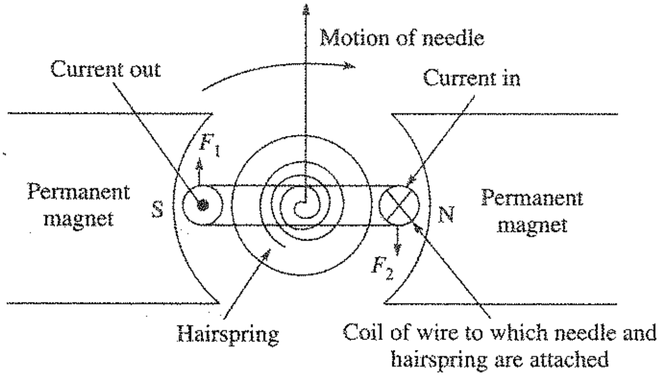


Figure 4.7 Illustration of D'arsonval principle as forcer for accelerometer.

In accelerometer, the hairspring is not necessary as the restoring force is provided by the acceleration acting on the proof mass. The two electromagnetic forces F_1 and F_2 form a couple to provide a rotation whose direction will be opposite to that produced by the inertial force.

Force F_m in such a design with current i in the coil is given by

$$F_m = \pi d N_t B_i \tag{4.19}$$

where

d = coil diameter
N_t = number of turns in the coil
B = flux density in the gap

The direction of electromagnetic force is perpendicular to both B and i. The position of the magnets and the coils should be so arranged that the electrical force F_m passes through the proof mass centre. The design is optimised so that maximum force is generated with minimum of current, and this is realised by operating around the magnet BH-max point and using magnets with high-energy product such as Samarium Cobalt ($SmCO_5$) or its next higher version Sm_2CO_{17}. The latter has a magnetic energy product that lies in the range of 25–28 millions gauss-oersted. Working around the BH-max point reduces the volume of magnet needed and hence the magnet mass. However, samarium cobalt magnet shows high sensitivity to temperature and typically, the flux density changes by 480 ppm/°C change in its temperature. This directly changes accelerometer rebalance scale factor and needs to be addressed in the design. Three methods, which are available to tackle this problem, are:

(i) Operate the accelerometer at a fixed temperature.
(ii) Use of temperature compensating shunt for the magnets.
(iii) Use of special type of low temperature coefficient samarium cobalt magnet.

The magnets are normally located in the accelerometer case and some form of shielding is introduced to reduce the effect of external magnetic field. The forcer coils in two halves, are placed on the pendulum. The current is conducted to the forcer coil by taking gold wire over the hinges. The coils thus contribute to the proof mass and the location of the mass centre relative to the suspension point. Thus the design of forcer coil must address to minimisation of power dissipation and provide suitable path for heat transfer out of the pendulum. If digital capture loop is considered, the coil time constant must be kept low.

The typical range of values of accelerometer scale factor, with analog capture, lies between 1.0 and 5.0 mA/g. This has an interpretation that this is the magnitude of feedback current needed to balance an acceleration of magnitude $1g$ (≈ 9.8 m/s^2).

EXAMPLE 4.5 (on forcer): If the operating flux density of the $SmCO_5$ magnet is 0.7 T and coil diameter is 10 mm, find out the number of turns needed to close the loop with 50 mA current for the accelerometer of Example 4.1.

Solution: For the forcer coil of N_t turns, the developed force is given by Eq. (4.19). On substitution, we get:

$$F_m = (3.14 \times 10 \times 10^{-3} \times 0.7 \times 50) N_t$$

From Example 4.3, the torque required to close the loop is:

$$7.8 \times 10^{-5} \times 50 = 39.12 \times 10^{-4} \text{ N-m}$$

Therefore, force F_m (to be generated by the forcer) $= \dfrac{39.12 \times 10^{-4}}{2 \times 10^{-2}} = 19.56 \times 10^{-2}$ N

Equating, $(3.14 \times 10 \times 10^{-3} \times 0.7 \times 50) N_t = 19.56 \times 10^{-2}$. So, $N_t = 178$

Capacitive torquing is an alternative to electromagnetic torquing with an advantage of negligible power dissipation on the pendulum under varying acceleration. It is also advantageous from digital rebalance point of view. But the biggest drawback is in the limited torquing capability when compared with the electromagnetic version. As a consequence to this torquing limitation, which reduces sensing range, macro sized accelerometer, with relatively large proof mass, normally uses electromagnetic torquing.

Flexure and pendulum suspension

Flexure hinge is considered as ideal for very low static friction and practically demonstrating infinite resolution. In a closed loop accelerometer, the entire restraint torque is expected to be provided by the servo operated forcer system. As a result, such flexure suspension matches ideally with the closed loop requirement. This leads to use of very thin section flexure hinge whose thickness is designed that is purely based on the survivability to mechanical shock and vibration along with sufficient strength to micro deformation. The choice of hinge material is based on strength, low hysteresis and geometrical stability across temperature. Typical such materials are elgiloy, beryllium copper and maraging steel. Fused quartz is also a good hinge material, especially with the development of ultrasonic machining and micromachining. Normally, two coplanar flexure suspensions are used as shown in Figure 4.8, while Figure 4.6 shows another view of the thin flexure pendulum suspension. During the normal closed loop operation, the electrical forcing system opposes the acceleration load and provides a near stress free condition for the hinges that supports the pendulum.

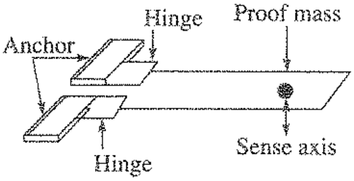


Figure 4.8 Co-planar pendulum suspension with thin flexure hinges.

However, during non-operating phase, there is no electrical capture of pendulum deflection, so that the stiffness of the flexure hinges alone supports the pendulum under the gravitational pull. This gravitational force results in the thin flexure hinges getting sufficiently deflected. With flexure thickness typically designed between 12 μm and 20 μm, the pendulum deflection under this gravitational pull can be sufficient to stress the hinges severely. This stress is controlled within the acceptable limits by limiting the pendulum deflection with limit stops on either side of the pendulum. However, spacing of the limit stops should be such that under maximum operating linear acceleration and vibration, the limit stops are not hit. Fluid damping is provided to protect the flexure from non-operating shock.

Damping

Damping is needed in an accelerometer to provide an acceptable output response in time domain as well as in frequency domain. From Eq. (4.2), we find that the damping force $F_d(t)$ in time domain is given by

$$F_d(t) = CV_y(t) \tag{4.20}$$

where

C = damping coefficient

$V_y(t)$ = velocity of the proof mass

Common damping methods followed are:

(i) Fluid damping, this again can be divided into:
- Using compressible fluid
- Using non-compressible fluid

(ii) Feedback loop damping (possible for accelerometers with feedback loop)

Fluid damping in an accelerometer can occur in the following manner:

- Through shear resistance
- Through squeeze film

Shear resistance damping: When the proof mass moves in a space, which is surrounded by fluid, damping force is induced by the shear resistance between the contacting surfaces of the proof mass and the surrounding fluid. This type of shear stress was originally postulated by Isaac Newton and the fluids, satisfying this postulation, are normally called Newtonian fluids.

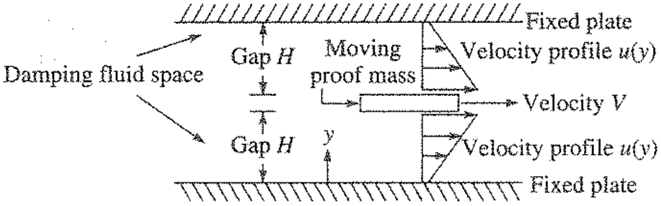


Figure 4.9 Shear resistance damping in accelerometer.

The principle of shear resistance damping can be explained with the help of Figure 4.9. The proof mass is moving with a velocity V and surrounded by the damping fluid between two fixed plates. For a non-slip flow condition, the velocity profile $u(y)$, on both faces of the proof mass, is considered linear. The shear stress τ, at either the top or the bottom face, can be expressed as:

$$\tau = \mu \frac{du(y)}{dy} \tag{4.21}$$

where μ is dynamic fluid viscosity.

Since, the velocity profile is considered linear, the velocity profile can be written as:

$$u(y) = \frac{Vy}{H} \tag{4.22}$$

where H is the width of the gap between the top and the bottom surfaces of the proof mass and the surrounding plates. Hence, the shear stress at the contacting surfaces τ_0 can be expressed as:

$$\tau_0 = \mu \frac{V}{H} \tag{4.23}$$

So, the shear force F_D acting on both the top and the bottom faces of the proof mass can be written as:

$$F_D = \tau_0(2Lb) = \frac{2\mu Lb}{H}V \tag{4.24}$$

where L and b are the length and width respectively of the proof mass. So, the damping coefficient C can be written as:

$$C = \frac{F_D}{V} = 2\mu\frac{Lb}{H} \tag{4.25}$$

Since L, b and H are decided based on several other design considerations, choice of μ becomes an important consideration for arriving at the suitable value of C. Typical non-compressible damping fluid can be a liquid such as silicone oil. Compressible damping fluid is a gas such as air or nitrogen or helium.

The typical ratio of dynamic viscosity at 20°C for helium and silicone oil is 19.4 : 740. This gives a choice whether to use gas or oil. Further, the sensitivity of dynamic viscosity of silicone oil with temperature is quite low; hence, it is widely used where liquid damping medium is necessary. In the case of gas damping, helium may be preferred over air as it has higher thermal conductivity that is useful in dissipating the heat generated in the forcer coil.

Squeeze film damping

Fluid provides damping in a narrow gap by squeezing the fluid and hence the name squeeze film damping. The mechanism of this type of damping is little complex and depends on the choice of a fluid which can be either compressible or non-compressible.

Squeeze film damping occurs when a plate moves in close proximity to another solid surface, in effect alternately stretching or squeezing any fluid that may be present between the moving plate and the solid surface. This fluid can act as a spring and or a dashpot having significant effect on the dynamics of the system. In working out a mathematical formulation, certain assumptions are made like quasi steady fluid characteristic, flow dominated by viscosity and that the gap is small compared to overall plate width. Precise nature of damping involved in a modern macro sized accelerometer whose size has been reduced over the years, needs careful analysis due to combination of the above factors that are involved. Readers interested to know more on this topic may refer the book of [**Tai–Rau Hsu, 2002**].

Where silicone oil is used in accelerometer, its volume expansion with temperature needs to be controlled using metallic bellows. Considerable amount of perfection is required in the oil filling process ensuring cleanliness, total avoidance of air entrapment and leak proof joints for the entire life of the accelerometer. Air entrapment leads to bias instability problem.

Rebalance servo

Accelerometer rebalance servo plays a critical role in accelerometer performance and reliable operation, which in turn define the following typical requirements:

- To provide the operating linear acceleration range, which is usually expressed in units of g.
- To provide high dc (low frequency) loop stiffness, so that the steady state error is minimised. The stiffness is usually expressed in arc-s/g.

- A required loop bandwidth, which is consistent with the environmental acceleration characteristic of vibration and shock.
- A high operating dynamic range, typically of the order of 1 million, for navigation application.
- A suitable scale factor either in volts/g for an analog rebalance loop or in pulses per s/g for a digital rebalance loop.
- An acceptable step or frequency response feature with adequate gain and phase margin.

Considering the above requirements, accelerometer servo loop design has evolved which is either analog or digital. A digital loop provides interface with computer without the need for analog to digital conversion which is needed in an analog loop. But characteristics of all accelerometers are not suitable for digital rebalance loop and as a result analog rebalance loop continues.

Analog rebalance

Basic features of analog rebalance loop are shown in Figure 4.10, which is known as a proportional control loop. In this figure, a_i is acceleration along input axis; P is accelerometer pendulosity; J is inertia of pendulous mass; C is damping coefficient; K_f is flexure stiffness; K_p is pick-off scale factor; K_e is servo gain consisting of all the electronic elements in the loop; R_c is forcer coil resistance; R_s is sense resistor; $L(s)$ is forcer coil inductance; i is feedback current; K_t is accelerometer torque constant; V_o is accelerometer output in volts.

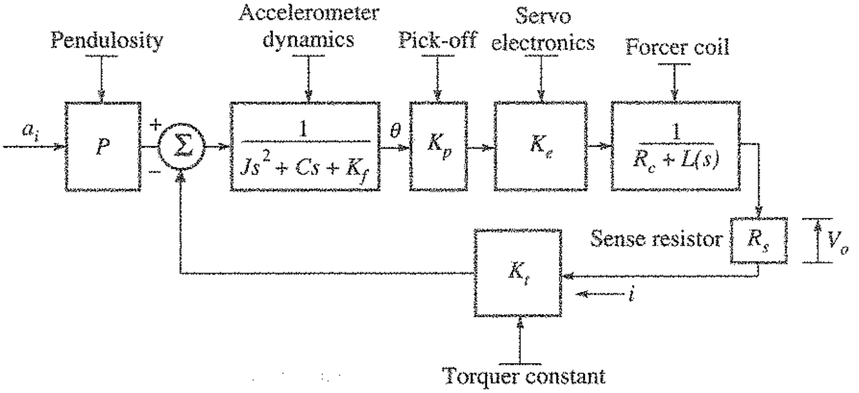


Figure 4.10 Analog rebalance accelerometer servo loop with proportional control.

Overall loop gain is given by

$$K = \frac{K_p K_e K_t}{R_c + R_s} \tag{4.26}$$

Stiffness of the flexure K_f is neglected being quite low compared with the servo stiffness. Since K_p and K_t are the accelerometer design constants, electronic gain K_e can be determined, which will provide loop response, stiffness and the desired bandwidth. Higher the overall loop gain K, higher will be the loop stiffness which in turn will lower the deflection of the pendulous mass under acceleration. This lower deflection in turn will lower the cross axis coupling. Closed loop dc stiffness value of around 1–2 arc-s/g and bandwidth >150 Hz are typical in accelerometer for aerospace applications.

In this class of accelerometer, where accurate measurement of linear acceleration is of fundamental importance, the requirement of reasonably high bandwidth can be explained in the following manner:

(i) The transient response of the accelerometer loop must settle well within a time that is consistent with the accelerometer data update frequency discussed in Chapter 2.

(ii) The loop must not saturate under environmental shock and vibration spectrum. Quite often, vibration isolators are introduced to relieve the environmental load on accelerometer.

Digital pulse rebalance is an alternative to VFC where the inertial sensor, having a rebalance loop feature, is digitally servoed that enable interfacing with the computer without the of an analogue to digital converter. This scheme is explained Section 8.11 in Chapter 8.

4.4.3 Accelerometer Errors and Model

All accelerometers will have errors and these are categorised under different modelleble terminologies. As discussed in Chapter 2, the systematic part of these modelleble errors is compensated in computer which signifies the importance of modelling. Typical terminologies describing the errors and their corresponding model terms are:

(a) Bias
(b) Scale factor error
(c) Scale factor non linearity error
(d) Cross axis coupling error
(e) Axis misalignment error

Bias K_0 is an output in the accelerometer in the absence of acceleration. It changes with time, temperature or even under accelerometer switch ON/OFF condition.

Scale factor K_1 changes with time and temperature resulting in error.

Scale factor nonlinearities (K_2 and K_3) indicate that under different linear acceleration magnitudes, the scale factor is not the same and the dispersion from the linearity is the error. K_2 and K_3 are the model parameters which represent the nonlinearities.

Cross axis coupling error terms (K_{ip} and K_{io}) have been described earlier in Section 4.3.1.

Axis misalignment error terms (θ_o and θ_p) are due to misalignment of the actual input axis from reference input axis.

Typical model of a closed loop pendulous accelerometer with analogue rebalance is given as:

$$A(\text{ind}) = \frac{E}{K_1} = [K_0 + a_i + K_2 a_i^2 + K_3 a_i^3 + K_{ip} a_i a_p + K_{io} a_i a_o + \theta_o a_p - \theta_p a_o] \qquad (4.27)$$

Here

a_i, a_p and a_o = linear acceleration acting along input axis, pendulous axis and output axis respectively.

A(ind) = acceleration measured by the accelerometer normally in units of g.

E = accelerometer output measured electrically.

Accelerometers with digital rebalance may have terms like scale factor asymmetry which means that the positive half scale factor is slightly different from its negative half.

These model terms are calibrated using specified test methodology and with appropriate instrumentation. Depending on application and performance requirement, the model can be truncated. Readers can refer the book [**Bose et al., 2008**] for more on accelerometer model calibration.

4.5 Vibrating Beam Accelerometer

Vibrating Beam Accelerometer (VBA) works on the principle of force to frequency conversion where the inertial force on the proof mass causes a change in the frequency of a resonating beam. In a resonating beam, the beam is excited by an external force to make it vibrate at its natural frequency. The beam shape is likely to have different vibration mode frequencies. In the operation of a resonant accelerometer, the beam is made to vibrate at its transverse mode natural frequency. When a force is applied longitudinally, which is perpendicular to the direction of vibration of the beam, the beam resonant frequency changes and the change in frequency is a direct measure of the applied force. From this information, the input acceleration is determined as a change in no load pulse stream frequency output. Thus VBA is a solid state accelerometer with inherently digital output and is treated as an open loop sensor. In the early eighties of the last century, development of low cost tuning fork oscillator with piezoelectric crystalline quartz for watch industry provided sufficient motivation for research on VBA as a low cost alternative to classical spring mass accelerometer.

4.5.1 Force to Frequency Conversion Principle

Principle of the force to frequency conversion is explained with the help of Figure 4.11 where a beam of length l and thickness t is made to resonate at frequency f_0 along the beam width direction, defined as transverse mode, with one end attached to a proof mass M and the other end is rigidly fixed.

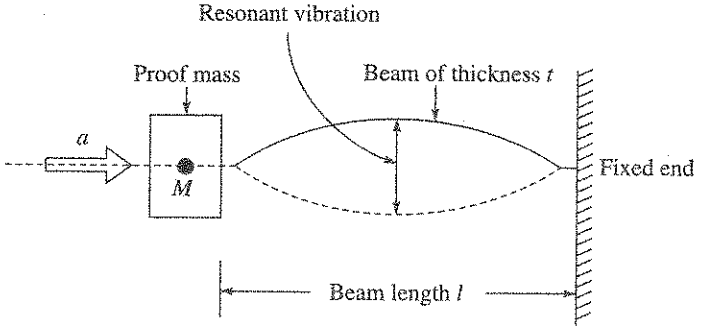


Figure 4.11 Force to frequency conversion.

In the absence of input acceleration, the transverse resonant frequency f_0 of the suspended beam is given by

$$f_0 = \frac{1}{2\pi}\sqrt{\frac{K}{m}} \tag{4.28}$$

When acceleration of magnitude a is applied to this beam along the beam length direction, the resultant inertial force Ma causes a shift in the beam transverse resonant frequency and the magnitude f_a of the shifted frequency is given by **[Norling, 1991]**

$$f_a = \frac{1}{2\pi}\sqrt{\left(\frac{K}{m} + \frac{\beta}{m}Ma\right)} \tag{4.29}$$

$$= f_0\sqrt{\left(1 + \frac{\beta}{K}Ma\right)} \tag{4.30}$$

where

K = transverse beam stiffness
m = beam equivalent mass
β = design constant

Later on in Section 4.5.4, these equations have been derived using the beam parameters.

Equations (4.29) and (4.30) can be analysed to derive the following features of such accelerometer:

(a) The shift in frequency is such that when the beam is in tension, frequency f_a increases and it goes down when the beam is under compression, which happens due to reversal of acceleration direction.

(b) For a given acceleration, frequency shift is more with increase in proof mass.

(c) It is also seen that the force to frequency change is nonlinear in nature.

(d) When the beam is under compression, the resonance stops when the factor $\beta Ma/K$ equals –1. This condition then puts a design limit on the permissible acceleration range for the beam under compression. This condition is important as it will be shown later that in VBA there are two beams. When one beam is under tension, the other beam is under compression.

4.5.2 Resonator Features

The most important element in a VBA is the resonator as the accelerometer performance, to a large extent, is governed by the resonator quality factor Q. We will elaborate this aspect in the subsequent section.

Resonator Q factor

VBA resonator is a mechanical structure designed to vibrate at its resonant frequency with a high Q factor. The widely known definition of a high Q system is 2π times the ratio of the total energy stored under resonance condition divided by the energy lost in a single cycle under resonance condition and is a dimensionless parameter. However, there have been derivatives from this definition and for very high Q structure, the resonator Q-factor can be defined as:

$$Q = \pi f_0 \tau \tag{4.31}$$

where

τ = resonator time constant

f_0 = resonator unloaded frequency

Resonator time constant τ is a number in the unit of seconds which gives the duration the resonator vibration amplitude will take to fall to $1/e$ of the initial amplitude when the electrical excitation is withdrawn. This equation for Q is same as that is shown in Chapter 3, for HRG. Readers can refer to Appendix C on Q-factor for the derivation of Eq. (4.31).

A high Q-factor means a sharp response peak as damping in such structure will be quite low. Such a characteristic, where low frequency as well as high frequency is attenuated, is defined as band-pass filter.

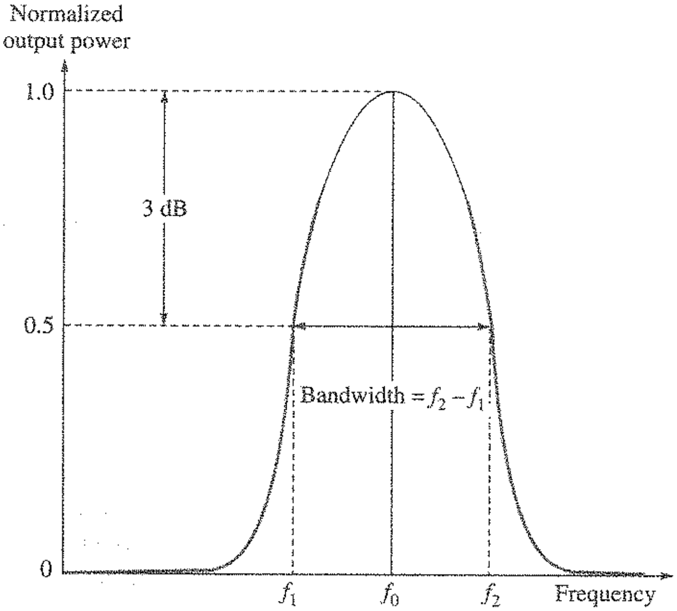


Figure 4.12 Band-pass feature of a high Q resonator.

In a band-pass filter (Figure 4.12) the characteristic can be related to circuit Q as:

$$Q = \frac{f_0}{f_2 - f_1} \tag{4.32}$$

where

f_0 = centre frequency

f_1 and f_2 = half power point frequencies (0.707 of centre frequency amplitude) as shown in Figure 4.12.

For a very high Q resonator, the magnitude difference of $f_2 - f_1$ reduces sharply. Such a characteristic can then be interpreted to say that the detectable resolution in frequency shift is given by

$$\Delta f = \frac{f_0}{Q} \tag{4.33}$$

Since VBA measures acceleration by detecting a change in its resonant frequency, the implication of high Q is that the detectable acceleration resolution improves.

EXAMPLE 4.6 (on VBA resolution): Assume a typical VBA has f_0 = 30,000 Hz, Q = 60,000, scale factor = 100 Hz/g. Find out the acceleration resolution.

Solution: Using Eq. (4.33),

$$\Delta f = \frac{3 \times 10^4}{6 \times 10^4} \quad \text{or} \quad \Delta f = 0.5 \text{ Hz}$$

Using the scale factor, the acceleration resolution is

$$\frac{0.5 \times 1}{100} = 0.005 \text{ g}$$

Further, a high Q means good frequency stability and the power demand on the electronic circuitry, used to sustain the oscillation, is minimised. A high Q-factor shows that the resonant structure is well isolated from its surroundings and that the influence of external factors will therefore be minimised. In essence, a high Q is a figure of merit in a VBA. In navigation grade VBA, such Q-factor reaches >50,000. Achieving a high Q is a combination of design and technology, as there are several factors which contribute to the losses. Some more elaborations on this aspect are presented as follows:

Beam material: The ideal beam material for the resonator is the one that has very high internal Q-factor, which implies very low material damping. Piezoelectric crystalline quartz, possessing internally very high Q ($>5 \times 10^6$) and excellent mechanical properties, became the first choice in the eighties. Quartz is non magnetic, hard, strong but not extremely brittle. It has very low internal losses and an infinite fatigue life. The advent of ultra pure silicon crystal has shown that it has also excellent properties for becoming a high Q resonator. Rapid growth in silicon micromachining is an advantage over quartz and the primary technical difference between the two materials is that silicon is not piezoelectric.

Resonator anchoring: Minimisation of energy losses to the surrounding structure due to resonator anchoring becomes a key aspect of resonator design and technology. Currently, there are two operational schemes:

(i) Single beam resonator
(ii) Dual beam resonator

The first approach utilises a single beam with integral end isolation to counterbalance the turn around acceleration forces of the single beam motion. While it has evolved a good design with very good performance, it is complex and is not conducive towards low cost fabrication.

In the dual beam resonator, the end conditions are optimised where it tries to cancel rather than isolate the vibration energy leaving the resonator. The dual beam, also called dual tine, resonator is conventionally known as Double Ended Tuning Fork (DETF). A conventional tuning fork has one end closed and the other end open. When both ends of a tuning fork are closed, it gives rise to a DETF structure.

Figure 4.13 shows a DETF resonator schematic. The scheme requires two identical tines in the plane of vibration where one tine is driven 180° out of phase with respect to the other tine. In this dual tine design, two tines, T_1 and T_2, when vibrated in phase out mode, the bending and shear stresses combine at the tine end and cancel each other within a few tine widths. This way the need for a complex isolator is eliminated and the anchoring can be done directly without the loss of vibration energy.

The requirement that the two tines should be geometrically identical imposes considerable manufacturing constraint. The development of quartz micromachining technology has made the realisation of matched tines a reality and in large numbers so that from a one-inch wafer, dozens of identical resonators are produced. As a result, low cost resonators can be produced in bulk which in turn has given considerable edge to DETF scheme over the single beam resonator scheme.

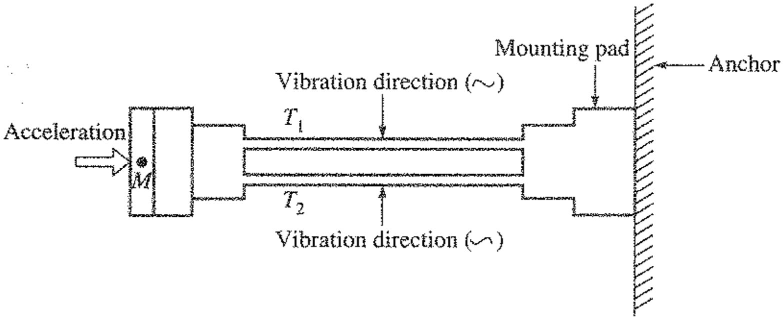


Figure 4.13 Dual tine resonator with double ended tuning fork shape.

4.5.3 Quartz Resonator Electrical Analog

Piezoelectric crystalline quartz beam resonates through the inherent property of piezoelectric material where an appropriately applied sinusoidal voltage causes the material to vibrate. Such a behaviour, where the electrical energy has been transformed to mechanical motion, has been electrically modelled and the electrical analog of the quartz beam in resonance is shown in Figure 4.14.

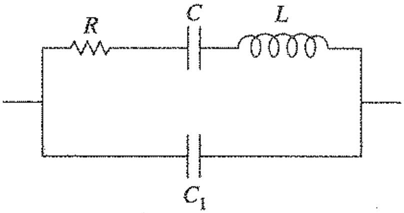


Figure 4.14 Electrical analogue of quartz beam resonator.

In Figure 4.14, L, C, R are defined as motional parameters and C_1 is equivalent shunt capacitance, sometimes referred as static arm capacitance.

There are two important dimensionless parameters which define the quartz crystal and they are:

(a) Figure of merit M of crystal stability
(b) Resonator Q-factor

Electrically, M is defined as the ratio of shunt arm impedance to motional arm impedance at series resonance. The relationships defining these parameters are expressed as:

$$M = Q\frac{C}{C_1} \tag{4.34}$$

$$Q = \frac{1}{2\pi f_0 RC} \tag{4.35}$$

where f_0 is the series resonance frequency, the frequency at which the motional arm reactance is zero and it is given by

$$f_0 = \frac{1}{2\pi\sqrt{LC}} \tag{4.36}$$

At resonance, the circuit effectively becomes a parallel combination of R and C_1. It will be noted that the maximisation of M requires maximisation of Q.

In the event, silicon is chosen as beam material, the beam is normally made to resonate using capacitive forcing scheme and requires different electrical analogue model.

4.5.4 Vibrating Beam Acceleration Sensor and Push-Pull Operation

In this section, the currently evolved configuration of VBA will be discussed. Development of an accelerometer configuration with resonator reveals certain interesting aspect of design, which is a considerable departure from the conventional accelerometer described earlier. All modern vibrating beam acceleration sensors are designed to operate in push–pull mode, which is conceptually similar to operating two such separate accelerometers in back to back configuration. This means for a given acceleration direction, when one VBA beam is under tension, the other VBA beam is under compression. In reality, this mode of operation is realised in a single VBA through an integrated design and fabrication. Such a configuration is common to both the types of accelerometer involving either the quartz beam or the silicon beam.

In *push–pull operation of a VBA*, two resonators are arranged in such a way that when one resonator is in tension, the other resonator is in compression (Figure 4.15). When acceleration a_i acts along the axial direction as shown, the top resonator beam 1 is in tension, and the bottom resonator beam 2 will be in compression. VBA output can then be processed as a difference of the two resonator beam frequencies. The benefits in this scheme are further examined.

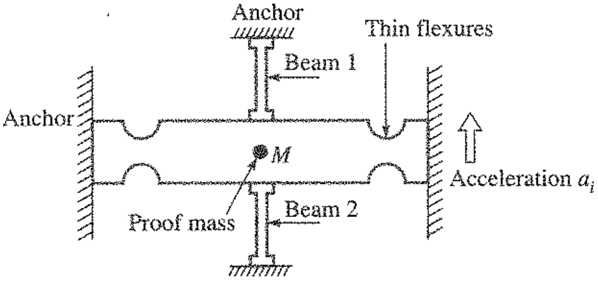


Figure 4.15 Push–pull scheme in a VBA.

Figure 4.16 shows a scheme where there is a common proof mass with a non-pendulous configuration. The lateral stiffness of the beam under acceleration a_i is given by

$$K = \frac{12EI}{L^3} + \frac{12}{\pi^2 L} F \tag{4.37}$$

where

E = Young's modulus of the beam material
I = moment of inertia of the beam
L = beam length
F = axial force (= Ma_i)

The resonator frequency f under acceleration can be expressed as:

$$f = \sqrt{\left[\frac{2}{m}\left(\frac{12EI}{L^3} + \frac{12}{\pi^2 L} F\right)\right]} \tag{4.38}$$

$$= f_0 \sqrt{\left(1 + \frac{L^2}{\pi^2 EI} F\right)} \tag{4.39}$$

where

$$f_0 = \sqrt{\left(\frac{24EI}{mL^3}\right)} \tag{4.40}$$

Equation (4.39) can be expanded by a series expansion to give:

$$f = f_0 \left[1 + \frac{1}{2} CF - \frac{1}{8} C^2 F^2 + \frac{1}{16} C^3 F^3 + \cdots\right] \tag{4.41}$$

where C is a constant. This can be expressed in the familiar form of accelerometer model as follows:

$$f = [f_0 + K_1 a_i + K_2 a_i^2 + K_3 a_i^3 + \cdots] \tag{4.42}$$

where the linearised scale factor K_1 is given by

$$K_1 = \frac{M}{2\pi^2} \sqrt{\left(\frac{24L}{mEI}\right)} \tag{4.43}$$

K_2 and K_3 are the coefficients describing nonlinearity in the accelerometer output.

Since there are two beams, when one is in tension, the other is in compression; Eq. (4.42) can be split into two equations. Also, if we take a realistic assumption that between these two beams, beam parameters will have some mismatches, then the output equations of the VBA become:

$$f_T = f_{0T} + K_1 a_i + K_{2T} a_i^2 + K_{3T} a_i^3 \tag{4.44}$$

$$f_C = f_{0C} - (K_1 + \Delta K_1) a_i + K_{2C} a_i^2 - K_{3C} a_i^3 \tag{4.45}$$

By taking a difference of the outputs, the net accelerometer output becomes:

$$f_T - f_C = [(f_{0T} - f_{0C}) + 2K_1 a_i + \Delta K_1 a_i + (K_{2T} - K_{2C}) a_i^2 + (K_{3T} + K_{3C}) a_i^3] \tag{4.46}$$

where

f_T = output of the beam under tension

f_C = output of the beam under compression

ΔK_1 = mismatch between the two scale factors

Equation (4.46) provides following conclusions for the push–pull mechanisation:

- Bias = $(f_{0T} - f_{0C})$ can be reduced substantially with matched resonators.
- Nominal scale factor = $2K_1$ has doubled.
- Second order nonlinearity = $(K_{2T} - K_{2C})$ has reduced appreciably.

Similarly, temperature dependency of bias and nonlinearity are cancelled or reduced.

The advantages of push–pull scheme are equally good for design with single beam resonator or with double ended tuning fork resonator, and as a result the push–pull scheme has become the standard vibrating beam accelerometer configuration. The scheme is independent of the choice of beam material. The push–pull mechanisation offers a choice for selecting a single proof mass or a dual proof mass. Single proof mass has a natural advantage of fabrication and size reduction. Figure 4.15 shows the single proof mass scheme, while Figure 4.16 shows a scheme with dual proof mass. Dual proof mass can offer some mechanical isolation between the two resonators.

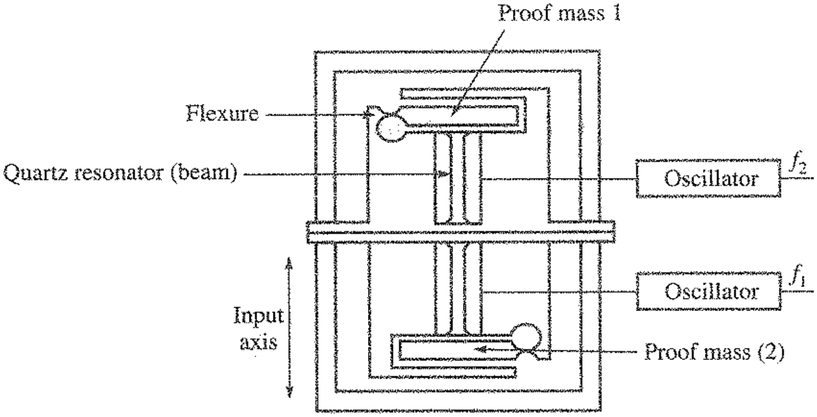


Figure 4.16 Dual proof mass VBA features.

EXAMPLE 4.7 (on VBA push–pull operation):

(i) The two resonators in a VBA have frequencies of 30,050 Hz and 30,025 Hz respectively. If the scale factor in each is 100 pps/g, calculate the VBA bias and its scale factor?

(ii) Now, VBA is exposed to variable ambient temperature and the no load frequency shifts by 0.01% and 0.015% respectively in the resonators. What will be the new bias?

Solution:

(i) VBA scale factor is $2 \times 100 = 200$ pps/g

Push–pull VBA bias is $30{,}050 - 30{,}025 = 25$ Hz or $25/200$ g $= 0.125$ g

(ii) The shifted frequency in one resonator is:

$$30{,}050(1 + 1 \times 10^{-4}) = 30{,}053 \text{ Hz}$$

The shifted frequency in the other resonator is:

$$30{,}025(1 + 1.5 \times 10^{-4}) = 30{,}029 \text{ Hz}$$

So, the new VBA bias is:

$$(30{,}053 - 30{,}029) = 24 \text{ Hz} = 0.120 \text{ g}$$

Thus we find that the effect of temperature has been practically cancelled by the push–pull operation. So, realising matched resonators becomes important where MEMS process helps.

Vibrating beam accelerometer feature involves *Pendulous or non pendulous* scheme and currently, both the schemes are operational. Preference of one scheme over the other depends on design and manufacturing considerations. Figure 4.15 shows a non-pendulous scheme, whereas Figure 4.16 shows a pendulous scheme.

Another aspect of VBA feature of importance is the *Suspension* of proof mass. The common form of suspension in a VBA uses thin flexures similar to that is used in rebalance accelerometer and is designed with similar considerations. Common flexure materials are, therefore, similar in nature. They are primarily made of maraging steel, beryllium copper, fused quartz and silicon. Switch-off tumbling related stress on the flexures, as seen in pendulous servo accelerometer, is not critical here due to the restraint effect of the resonator beam, which does not allow the flexure to deflect. However, the requirement of high Q in VBA allows minimal gaseous damping and this factor is borne in mind in protecting the flexures against shocks.

4.5.5 VBA Schematic and Operation

It is clear from the foregoing discussions that quite a few configurations of VBA are possible, and they are actually in operation. Figure 4.17 shows a typical of these operational configurations. The configuration depicts a single proof mass pendulous VBA having quartz as resonator and an electrical interface.

At the centre, the proof mass is supported with flexures. It has a gas damping system. Two dual tine resonators are connected with the proof mass. Crystal-controlled oscillator drives each resonator.

Resonator drive electronics provides for the following functions:

- Provides highly stable no load frequency
- Tracks the change in frequency under acceleration
- Provides highly stable voltage to the resonator

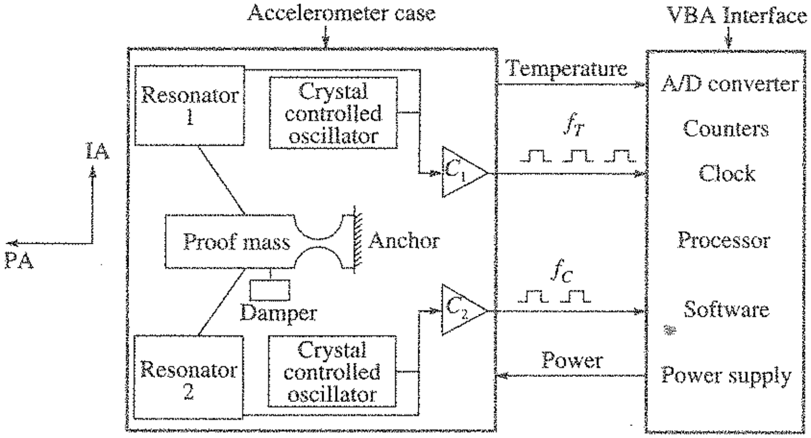


Figure 4.17 VBA schematic with electrical interface.

The oscillator output is a sine wave and this is converted to a square wave and buffered. The electronics blocks C_1 and C_2 carry out these functions. The square wave resonator pulse train outputs are denoted by f_T and f_C where the former is for beam under tension and the later for the beam under compression. Input axis (IA) and the pendulous axis (PA) are marked. Consistent with the current technology for sensor miniaturisation, the sensor electronics are miniaturised. This may be either in the form of hybrid microcircuit, or surface mount technology or for large number, in the form of low power ASIC and packaged inside the accelerometer case. This integral packaging of electronics leads to power dissipation and an effective scheme for removing the resultant thermal gradient across the sensor becomes important to provide stable sensor output in a short time. This is because the sensor enclosure is evacuated and hermetically sealed for high Q operation, which are not conducive to heat dissipation.

On energisation, the accelerometer provides stable output pulses within a short time. The output pulses are accumulated in up-down counters, which are processor controlled.

Scale factor improvement

The natural scale factor of the resonator is not high for navigation application. Quite often, digital electronic processing is employed using high frequency computer clock. More than an order of improvement is possible in scale factor with such scheme. As a result, no special efforts are usually made to improve on the stand alone resonator scale factor.

4.5.6 Errors

VBA can be modelled similar to a force-balanced accelerometer where the errors are described with similar terminologies like bias, scale factor error, nonlinearity in scale factor and temperature sensitivity of these model parameters. Besides these, there is one more error, which is specific to a VBA, which is termed as *frequency lock*.

Frequency lock

While the push–pull scheme using two resonators offer substantial benefits in a VBA, this configuration gives rise to an error in a VBA called *frequency lock*. The error manifests at the frequency cross over point when the two frequencies are found to remain locked over a small range of input acceleration. The error is explained along with the help of Figure 4.18.

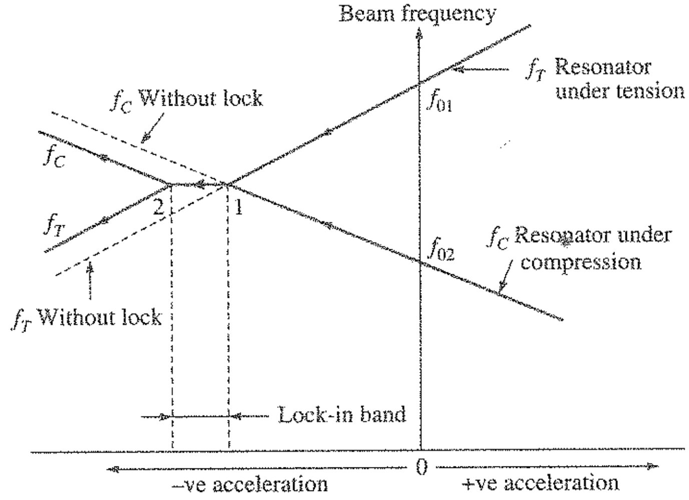


Figure 4.18 Frequency lock-in in a VBA.

The first resonator, initially under tension, is designated as f_T and has a no load frequency f_{01}. The second resonator, initially under compression, is designated as f_C and has a no load frequency f_{02}. When a slow changing negative acceleration is applied, the frequency of the resonator under tension starts coming down, while the frequency of the second resonator starts building up. They reach a cross over situation at a point 1, but the cross over does not take place immediately as the two frequencies get locked to each other till they reach point 2. From this point onwards, the normal characteristic resumes.

The acceleration magnitude between 1 and 2 is the lock-in band and is a manifestation of nonlinearity like any other dead band in a sensor output. This is called static lock-in. As this error cannot be modelled and compensated, magnitude of this error is minimised through appropriate design. One possible approach is to improve upon the Q-factor and use two proof masses in lieu of one, so that the coupling of energy, between the resonators, is minimised. However, it is possible to manipulate the zone of occurrence to make the problem manifest at a non-operating flight regime. For example, the no load frequency of the resonators can be so chosen that the lock-in zone is on the negative side of the flight acceleration, as shown in Figure 4.18. The lock-in magnitude can be established using multi-position test if it lies within ± $1g$.

4.5.7 Comparison of VBA with Closed Loop Accelerometer

A comparative analysis between VBA and the closed loop accelerometer technologies is described as under:

Acceleration range and power: All types of VBA are open loop in operation. The input power demand is independent of acceleration permitting design for high acceleration range at low power. Power is low due to high Q resonator. VBA acceleration range is normally limited to around 10 per cent of the no load resonator frequency. In a closed loop accelerometer with analog rebalance, higher acceleration normally consumes more power and puts a limit on the accelerometer operating range.

Output format: VBA output is digital and avoids the error and cost associated with analogue to digital converter, which is required in an analogue closed loop accelerometer.

Tumbling: The resonator beam supports VBA proof mass all the time, which prevents stress on the flexure during sensor un-powered phase. A closed loop pendulous accelerometer, during un-powered phase, tumbles due to gravitational pull to cause strain on the thin flexures.

Frequency lock: VBA suffers from frequency lock and this is not seen in closed loop accelerometer. However, this frequency lock zone can be manipulated to be away from the operating linear acceleration range.

Reliability: VBA has much less number of critical parts compared to a closed loop accelerometer. When reliability is calculated based on parts count basis, VBA offers higher reliability.

Producibility: Resonator, the most critical part in VBA, is produced in large number using micromachining process for a double-ended tuning fork configuration. This gives significant advantage to VBA production at a lower cost. Silicon VBA is further superior in this aspect.

4.6 Interferometric Fibre-Optic Accelerometer

In fibre-optic accelerometer, optical effects like micro bend losses, interference or strain in the fibre, are utilised for the detection of acceleration. The method of detection involves optical interferometric technique, that has been explained in Chapter 3, being capable of very high sensitivity.

A practical interferometric accelerometer, called Mach–Zehnder accelerometer is shown in Figure 4.19. In one of the split paths, called sensing arm fibre, a proof mass is attached to induce

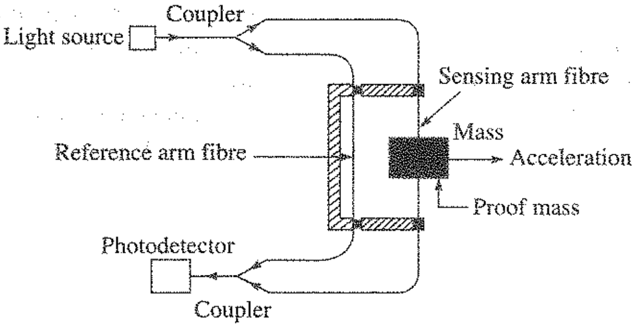


Figure 4.19 Mach–Zehnder interferometric accelerometer.

strain in the fibre under an external acceleration. The strain in turn causes a change in the optical path length compared to the reference arm fibre. This creates a relative phase change in the split beams, and when they are recombined in the coupler, it undergoes interference proportional to acceleration. This interference change is detected in the photo detector which converts it into an electrical output. In a variation to this design, the proof mass is attached to both the fibres in such a way that the effect of acceleration leads to increase in tension in one fibre and a decrease in the other. This is then used in push–pull mode to increase the sensitivity and reduction of common mode errors. Such optical accelerometers would find application where electricity may be hazardous to use or immunity is needed from RF or EMI interference.

Some more researches are in progress for alternative schemes and in one such scheme [**John et al., 2002**], in a micro fabricated silicon accelerometer, the detection is conceived through fibre optic sensing using Fabry-Perot interferometer.

Considerable research is in progress for micro accelerometer using detection through electron tunneling effect where tunneling current varies exponentially with the displacement of the proof mass and is thus highly sensitive [**Bose et al., 2008**]. This scheme suits very small structure as the sensitivity is independent of the size unlike capacitive detection where area of the plates influences the detection sensitivity as described earlier in this chapter.

4.7 Cold Atom Accelerometer

Current researches have been concentrated on precision navigation grade accelerometer based on cold atom interferometry whose intrinsic sensitivity is much higher compared to the navigation grade accelerometer discussed earlier. Considering its potential, a brief description of its principle is presented in the subsequent paragraphs.

4.7.1 Accelerometer Operation

In Section 3.9 (Chapter 3), the wave like property of cold atom has been discussed which can be further used to detect acceleration using interferometer. The method of constructing an interferometer is similar using a sequence of laser pulse $\pi/2 - \pi - \pi/2$.

In the absence of acceleration, the path lengths, path A and path B, of the two beams are straight and identical which results in zero phase difference between the beams. In the presence of acceleration acting perpendicular to the interferometer plane, the path lengths become curved and non equal which in turn results in a phase difference. These aspects are shown in Figure 4.20. Where Z-axis represents the position of atom at a time t.

[**Peter, Chung and Chu, 1991**], who have done commendable work, have shown that the phase difference $\Delta\phi_{\text{acc}}$ is related to acceleration a (gravitational acceleration in the measurement set up) as follows:

$$\Delta\phi_{\text{acc}} = kaT^2 \tag{4.47}$$

where T is the interrogation time, which is the time between the laser pulses and k is the effective laser wave number. The phase sensitivity equation shows the importance of long interrogation time within the interferometer as this gives quadratic improvement in sensitivity.

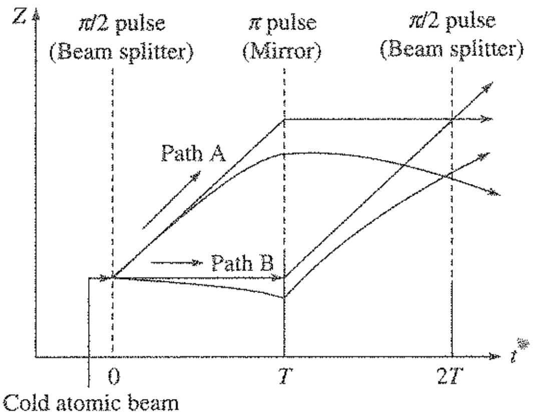


Figure 4.20 Principle of cold atom acceleration measurement.

The fringe of the interferometer can be read out by monitoring the relative population of the two states in the recombined atoms via laser induced fluorescence. Knowing the laser wave number and time T, the acceleration magnitude a can be determined. The demonstrated sensitivity of such cesium atom interferometer accelerometer to measure gravity in the laboratory, having transition wavelength of 852 nm, has shown $10^{-11}g$ with interrogation time $T = 1$ s. Readers can refer the paper of [**Kasevich, 2002**] to appreciate excellent prospect of this new development for high precision inertial navigation and as a gravity gradiometer.

SUMMARY

The chapter on accelerometer has brought out classification of modern aerospace accelerometers along with their principle of operation, features, configurations and various errors which are required to be minimised for navigation application. In this process, descriptive details have been made on the classical spring mass accelerometer operating either in open loop mode or in closed loop mode for high performance. The description has also touched upon the design aspects of such precision accelerometers. Thereafter, the descriptive details of modern vibrating beam accelerometer have touched upon some aspects of design considerations. These two types primarily constitute the majority in aerospace application. Fibre-optic accelerometer is introduced as it will have relevance for certain special application where electricity may be hazardous. Finally, cold atom accelerometer is introduced as it offers immense possibility on very high accuracy with unfolding application in the near future. Considerable number of worked examples are provided throughout the chapter for the benefit of the readers as well as problems given at the chapter end to test the understanding on these accelerometers.

EXERCISES

4.1 What is the difference between a pendulous accelerometer and a non pendulous accelerometer? Sketch a typical spring–mass open loop pendulous accelerometer and show the input axis (IA) and pendulous axis (PA).

(i) If accelerometer axes are governed by the right handed rule and if the accelerometer IA is vertical down while PA is towards east, what will be the direction of OA?

[**Ans:** (i) South]

4.2 In a pendulous force feedback accelerometer, an inertial acceleration of magnitude $2g$ is acting on a proof mass of 2×10^{-3} kg which is located 2 cm away from the flexure suspension with negligible stiffness. A forcer is located 1.5 cm from the flexure suspension. Sketch the problem and calculate

(i) Inertial force

(ii) Restoring force

Assume $g = 9.80$ m/s^2 [**Ans:** (i) 19.6×10^{-3} N, (ii) 26.1×10^{-3} N]

4.3 Sketch of a pendulous accelerometer with capacitive pick-off is shown in Figure E4.1. Nominal gap under zero acceleration is same for both. Explain how the gaps change with an applied acceleration shown in the sketch. If the upper capacitor value is C_1 and that of lower C_2, what will be $C_1 - C_2$

(i) +ve number

(ii) –ve number

(iii) 0

[**Ans:** (ii) –ve number]

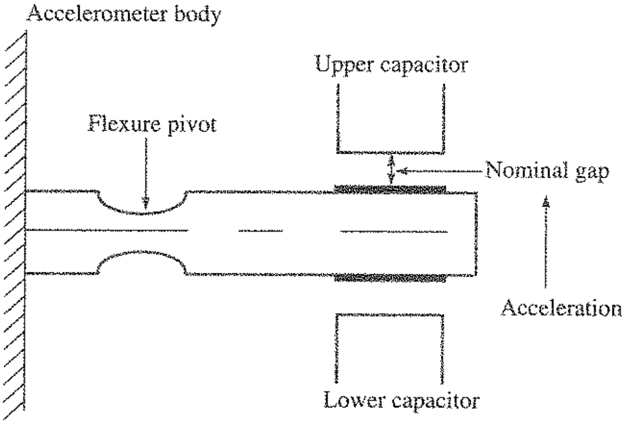


Figure E4.1

4.4 In a closed loop accelerometer, the accelerometer scale factor is 5 mA/g and a nominal bias of –10 mg. The accelerometer exhibits positive second order nonlinearity in its output. Write the accelerometer model and sketch the accelerometer output over the +ve and –ve acceleration. Calculate the output at $+10g$ with a nonlinearity of +100 μg/g^2.

[**Ans:** 9.995 g]

4.5 In an assembled VBA, the resonator under tension has a no load frequency of 30,000 Hz, while the other resonator has frequency of 30,200 Hz. Both resonators have scale factor of 100 Hz/g. Explain what happens if the VBA is used to measure thrust acceleration upto $5g$? What you would do on the VBA assembly so that the VBA works nicely in the thrust phase?

4.6 Sketch a vibrating beam accelerometer with DETF resonator showing the following:

(i) Resonator vibration direction

(ii) Proof mass suspension

(iii) A direction of acceleration along input axis

(iv) Resonator under tension

Explain briefly the function.

4.7 Sketch a DETF resonator and mention the advantages over a single tine resonator.

(i) Calculate the no load frequency of the DETF resonator given the following:
Beam mass = 0.01×10^{-3} kg, Beam length = 8 mm, Beam inertia = 1×10^{-7} m^4, $E = 2 \times 10^5$ Mpa **[Ans:** 48.8 kHz]

(ii) If the resonator is to be used to measure acceleration, what is its sensitive axis?

(iii) Mention advantages/disadvantages of single resonator accelerometer vis a vis two-resonator configuration.

4.8 Sketch of two separate VBA accelerometers (VBA1 and VBA2) is shown in Figure E4.2. Each one configured with a DETF.

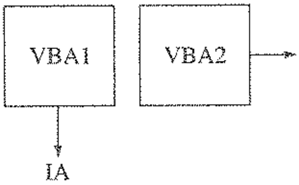


Figure E4.2

(i) How to realise a VBA in push–pull mode operation using these two? Sketch your answer.

(ii) If VBA1 has no load frequency of 35,000 Hz and scale factor of 125 Hz/g, while VBA2 has no load frequency of 35,020 Hz and scale factor of 125 Hz/g, what is the push–pull VBA bias and scale factor?

[**Ans:** Bias = 250 Hz/g, Scale factor = 0.08 g]

(iii) If VBA1 and VBA2 operate independently, what will be their bias?

[**Ans:** VBA1 = 280 g, VBA2 = 280.16 g]

4.9 An energised accelerometer, with its input axis vertical, is allowed to fall from an altitude so that during the time interval $0 - t_1$, it is under free fall, during the interval $t_1 - t_2$, it is encountering air resistance that is proportional to velocity and finally during the interval $t_2 - t_3$, the velocity remains constant. Sketch the output of the accelerometer.

REFERENCES

Bose, A., Puri, S., and Banerjee, P., *Modern Inertial Sensors and Systems*, PHI Learning, Delhi, 2008.

John, J. et al., Planar Silicon Accelerometer with Fiber-Optic Sensing-A System Level Study, *Proceedings of ISSS-SPIE*, International Conference On Smart Materials Structures and Systems, July 17–19, Bangalore, India, 2002.

Kasevich, Mark., Science and Technology Prospect for Ultra Cold Atom, Stanford University, CAMOS, November 2002.

Norling, B.L., An Overview of the Evolution of Vibrating Beam Accelerometer Technology, *DGON Symposium*, Stuttgart, Germany, 1991.

Peter, A., Chung, K.Y., and Chu, S., High Precision Gravity Measurement using Atom Interferometry, *Metrologia*, 2001.

Tai–Rau Hsu, *Mems and Microsystems*, Tata McGraw-Hill, 2002.

5

MEMS Based Inertial Sensors

Miniaturisation of silicon based electron devices and their integration has revolutionised the electronics industry. This has resulted in devices with nanometre feature size which led to high performance high speed integrated circuits capable of operation with supply voltage levels as low as 1.0 V. In most situations, silicon has been the material used as the substrate for Microelectronics. With the advent of Nanoelectronics and the associated quantisation effects several other materials such as silicon–germanium (SiGe), gallium–indium–arsenide (GaInAs), have found their way into this race to miniaturisation. Requirements of devices capable of operation at higher power levels and also in harsh environments especially temperatures in excess of 300°C have led to the use of semiconductor materials such as gallium nitride (GaN) and silicon carbide (SiC). Other materials which are gaining considerable attention in the nanoscience era are functional materials like the Carbon NanoTube (CNT). In addition, several other materials such as polymers are being considered for application in electronics.

The microfabrication technology, which has been the key to the success of microchips and microelectronics, is now revolutionising the Microsystems involving both microelectronics and micromechanical components. The basis for this revolution has been the excellent mechanical properties of silicon and its suitability for batch processing miniaturised mechanical devices using the already well established processing techniques for microelectronics devices and a few additional processes which are today referred to as micromachining. Etching out portions of silicon or any other material to realise the miniaturised mechanical structure is one of the several micromachining techniques. Miniature systems involving one or more micromachined micro scale devices are generally referred to as MEMS which stands for Micro–Electro-Mechanical Systems or simply the Microsystems. Miniature systems containing nanoscale mechanical devices, nano-devices or nanostructures are referred to as Nano–Electro-Mechanical Systems or NEMS. NEMS are MEMS scaled to submicron or nanoscale dimensions. MEMS/NEMS may have mechanical sensors and actuators in the micro/nanoscale and also micro/nano electronics integrated together. They also may have embedded biosensors, micro/nano channels and fluidic systems. Unlike conventional integrated circuits, these devices enable higher level functions including sensing, communication and actuation.

The MEMS field has grown initially on the firm footing of the existing silicon processing technology and infrastructure. MEMS devices have gradually evolved from the research labs and are now found in various technologies. Silicon micromachined mechanical devices are now widely used in aerospace, automotive, biotechnology, robotics and other applications and are opening avenues for the nanotechnology devices and systems. Pressure sensors, accelerometers, inkjet printer heads, and optical switches have all become commercial products. The integrated sensors are currently the largest application of MEMS. Other demonstrated MEMS applications include flow valves, electromechanical switches and relays, gyroscopes, inkjet nozzles, micromanipulators and connectors, as well as optical components such as lenses gratings, waveguides, mirrors, sources and detectors. For instance, a chip that contained over two million tiny mirrors, each individually addressed and moved by electrostatic actuation, has been produced by Texas Instruments. NEMS are powerful integrated systems that contain nanotechnology-enabled component or subsystems which are extremely miniaturised, broadly capable and highly reliable that it appears to be the almost impossible.

Moreover, MEMS and NEMS have also become enabling technologies for BioNEMS. The rapid growth of this technology is now widespread into other materials such as quartz, glass and polymer, and has extended into nano functional materials such as carbon nanotubes. The potential application of CNT for NEMS has been established and it has been noted as the forerunner of the nanoelectronics and nanotechnology.

This chapter initially describes the benefits of miniaturisation and scaling of mechanical components such as the sensors, and brings out the materials requirements for MEMS and the superior features of silicon as a mechanical material. The micromachining approaches such as bulk micromachining, surface micromachining and LIGA process are briefly discussed in Section 5.53. This chapter will also address micromachining of silicon, glass as well as polymer bringing out the relative merits and drawbacks of each technology. The piezoelectric, piezoresistive and capacitive sensor concepts are also covered. The operating principles and structures of some of the micro scale acceleration sensors and gyros are presented, discussing the issues specific to micromachined accelerometers and gyros.

5.1 Benefits of Miniaturising and Scaling Mechanical Sensors and Actuators

Miniaturisation of mechanical components to micro scale brings the same benefits to mechanical systems that are achieved in the scaled down microelectronics devices and systems using the microfabrication technology. These advantages can be summarised as follows: (i) Micromechanical devices and systems are inherently smaller and lighter. Hence they exhibit high performance in terms of speed of operation and high mechanical responsivity than their macroscopic counterparts and are invariably more precise. (ii) Cost of the device is very low because microfabrication using the photolithography and etching/deposition process paves way for batch processing. For instance, hundreds of devices can be realised on a single substrate such as silicon wafer. This also opens up the possibility of processing few tens of wafers simultaneously. (iii) Since the devices are fabricated with the microfabrication techniques, they can be integrated with electronics to develop high-performance closed-loop-controlled micro–electromechanical systems. (iv) On-chip integration of electronics and the sensors with microfabrication is particularly attractive to inertial sensors to enhance the reliability of the system. For instance, the wiring parasitic capacitances affect the performance and reliability of the capacitive accelerometers when the discrete sensor device is used along with a separate chip containing electronics. This problem can be overcome by on-chip integration of the accelerometer with electronics containing force balancing circuit. (v) When the sensors, actuators and electronics are miniaturised by micromachining technique and integrated in the MEMS, they provide lower power operation, compact and robust sensing capability. (vi) Electrostatic actuation force can be increased by almost two orders of magnitude by scaling because the breakdown strength increases from 3 V/μm (30 kV/cm) to 100 V/μm when the air gap is reduced close to about 1 μm which is approximately the mean free path length of air molecules.

Thus the reasons that accelerated use of MEMS technology are miniaturisation of existing devices, development of new devices based on the principles that do not work on a larger scale and the development of new tools to interact with the micro world. The cost reduction is achieved by decreasing the material consumption due to the miniaturisation. As the mass and size of the sensors can be reduced to micrometer level using micromachining technology, the applicability of the MEMS devices increases because the MEMS device can be placed where the traditional system does not fit in. Another advantage of MEMS is the system integration, which is made possible by integrating MEMS on silicon directly with electronics to include data acquisition, filtering, data storage, communication, interfacing and networking. Thus the MEMS technology not only makes the systems smaller but often makes them better. Often the filtering can be achieved with MEMS components such as mechanical resonators, and recent reports predict that future wireless designs will replace electronics with precision Mechanical Components.

The most important benefit of scaling down the mechanical components is the cube–square scaling effect. For instance when the length L is scaled down by a factor f, the volume of a mass M in an inertial system would reduce as L^3, whereas the surface area would reduce as L^2. As a result, the inertial force, $F = Ma$, is lowered by a factor f^3 for a fixed acceleration a. The resulting stress, given by the ratio of the force to the surface area, would, therefore, reduce by a factor f. Hence for a given material strength, miniaturisation permits the structure to withstand higher acceleration.

Scaling and miniaturisation also affect the properties of the material used to realise the sensors and actuators. Miniaturisation introduces constraints such as surfaces and interfaces, surrounding a thin film act to restrict dislocation formation and motion, resulting in very high strengths. This higher strength allows for an increased force transmission capability. However, the drawback is that in thin films the residual stress is invariably higher. This has to be tackled at the process level itself.

Increase in surface to volume ratio encourages surface diffusion mechanism such as creep. Therefore *ductile materials* cease to be tough at small sizes. (Toughness is determined by the plastic dissipation which is controlled by structural dimensions). However, miniaturisation reduces the probability of finding a 'flaw of given size' within the volume of the material under load. Therefore, the material strength of *brittle materials* (governed by the maximum flaw size) is higher when the structure is miniaturised, thus indicating that a brittle material such as silicon is more suitable for realising miniaturised structures. As the mechanical properties of the material being used for realising miniaturised sensors for any application, in the following section we examine them in detail.

5.2 Materials Requirements for Micromachined Inertial Sensors

Microfabrication process and material used to create the device must be scalable and suitable for batch processing to realise a low cost of production. Material and process must enable integration between electronic and non-electronic function. High performance, high strength and high reliability materials are required for mechanical elements. Materials for transducer elements which permit power or signal conversion from one physical domain to another are also necessary. As silicon is one of the most widely used material for microfabrication, we compare the mechanical properties of silicon with other popular mechanical material such as steel, iron as well as with other hard materials such as diamond, silicon carbide, sapphire, etc. Table 5.1 gives a summary of the results reported in the literature [**Petersen, 1982; and Sullivan et al., 2001**].

Table 5.1 Selected Mechanical Properties of Some Common Materials

Serial No.	**Material**	**Yield Strength** (GPa)	**Hardness** (GPa)	**Young's Modulus** (GPa)	**Density** (gm/cm^3)	**Thermal conductivity** (W/cm-K)
1	Diamond*	53	100	1035	3.5	20
	Diamond	—	90	800	—	—
2	SiC*	21	24.5	700	3.2	3.5
	SiC	—	30	350	—	—
3	TiC*	20	24	497	4.9	3.3
4	Al_2O_3*	15.4	20.5	530	4.0	0.5
	Al_2O_3	—	20	440	—	—
5	AIN	—	27	340	—	—
6	Si_3N_4*	14	34	385	3.1	0.19
	Si_3N_4	—	30	130	—	—
7	Iron	12.6	3.9	196	7.8	0.803
8	SiO_2(fibers)	8.4	10	70	2.5	0.014
9	Si*	7.0	12	190	2.3	1.57
10	Steel (max. strength)	4.2	14.7	210	7.9	0.97
11	W	4.0	4.66	410	19.3	1.78
12	Stainless steel	2.1	6.47	200	7.9	0.329
13	Mo	2.1	2.7	343	10.3	1.38
14	Al	0.17	1.27	70	2.7	2.36
15	Quartz	11 (compressive)	—	72	2.2	1.3 W/(mK)

*Indicates that the material is single crystal, Pascal (Pa) = 1 N/m^2 = 10 dynes/cm^2

5.2.1 Why Silicon is the Best Material for MEMS?

It is interesting to note that Silicon (Serial No. 9 in Table 5.1), which is the second most available material on the earth's crest, shows mechanical properties better than the most accepted mechanical material such as stainless steel (Serial No. 12 in Table 5.1) in terms of yield strength, hardness and Young's modulus. In addition, the density of silicon is lower than steel as well as aluminium. Thus one can say that silicon is harder than steel and lighter than aluminium. In addition, silicon is used as an electronic material in an already advanced microfabrication technology. Miniaturised mechanical devices such as sensors can be realised on silicon with high precision using batch processing, photolithography and microfabrication steps. As a result, on-chip integration of sensors with electronics can be achieved with very little additional effort.

Even though brittle, silicon exhibits higher strength when miniaturised and free from creep, fatigue and hysteresis. Other ductile materials suffer from thermally activated deformation processes such as creep, and are susceptible to hysteresis and fatigue when subjected to repeated stress cycle. The use of silicon microfabrication processes such as lithography, wet and dry etching methods, and the availability of single crystal silicon substrates with very low defect density allows the creation of structures with very fine surfaces, and therefore, with very high mechanical strength.

Additional benefits are achieved when miniaturisation is carried out with silicon. For instance, the processing routes such as deposition, etching and diffusion, doping, etc. primarily act on surfaces are more economically attractive at small scales due to the cube–square scaling of volume to surface area.

5.2.2 Merits of Silicon Micromachining Compared to Conventional Machining Techniques

As the silicon micromachining is suitable for batch processing, production costs of whole production is independent from number of components fabricated and the miniaturisation with finest details in the range of 1 μm to 10 μm is possible using standard photolithography technique to define the regions for micromachining. On the other hand, in the conventional micromachining approach, each component must be made piece by piece, and hence low price for large production volumes are the result of mechanisation. With the ultrasonic machining, sand blasting, laser ablation and spark erosion used for miniaturisation in the conventional approach, the finest details that can be machined are one to two orders larger than what photolithography makes possible in the silicon micromachining process.

5.2.3 Other Materials being Considered for Micromachining and Miniaturisation

Several other materials such as diamond, SiC, Al_2O_3, AlN and amorphous diamond have received attention due to their superior mechanical properties as shown in Table 5.1. Their application has been restricted to laboratory level because of the difficulties associated with micromachining them by simple etching process. However, deposited thin films of Si_3N_4 and

SiO_2 are very popular in the micromachining process and will be clear in the subsequent sections on micromachining.

Diamond for MEMS

Main attraction for diamond for MEMS applications is its highest hardness (100 GPa) and elastic modulus (1035 GPa), extreme wear resistance up to 10,000 times greater than silicon. The hydrophobic surfaces with inherent stiction resistance (parts do not stick together due to capillary forces from entrapped water) would assist in releasing cantilever beams. The chemical inertness of diamond allows their use in aggressive chemical environment. The primary challenge with this material is that it is tough to process, and it is difficult to integrate mechanical devices with electronics with diamond substrates.

Silicon carbide for MEMS

SiC in single crystal form is a high band-gap semiconductor capable of operation at high temperatures and high power levels compared to silicon. SiC offers much higher stiffness, hardness, toughness and wear resistance than the core CMOS material set. This is a major attraction of SiC for MEMS. At the same time the development of SiC based MEMS has been slow because it is difficult to process due to its relatively very low chemical reactivity and extremely high melting point (2300°C). In spite of these difficulties SiC MEMS pressure sensor has already been reported.

Quartz and polymer for micromachining application

Quartz is very stable dimensionally at higher temperatures and hence it has made inroads into micromachined accelerometers and gyroscopes. The coefficient of thermal expansion of quartz is $5.5 \times 10^{-7}/°C$ in the temperature range of 20°C to 320°C compared to the value of $2.6 \times 10^{-6}/°C$ at 27°C for silicon. However, quartz has not gained widespread applications like silicon because it is difficult for batch processing quartz, and integration with electronics is impossible.

Polymers such as SU-8 and PDMS, plastics, adhesives, plexi glass and so on have become increasingly popular materials for MEMS e.g. micro channels for micro fluidic systems and biomedical applications. Special Pyrex glass such as 7740 having sodium content and TCE matching with silicon is used very often for anodic bonding with silicon to achieve wafer level packaging.

In addition to the above materials, Nickel and Nickel alloys along with polymer mould by lithography for LIGA process are used for high aspect ratio MEMS structures as described in Section 5.3.3.

5.3 Micromachining for MEMS

As already elucidated in the previous sections, silicon is the major material used for micromachining and microengineering, particularly because there is a wealth of process knowledge and experience in the silicon semiconductor industry, in addition to the excellent mechanical properties of silicon. The basis of silicon micromachining is the photolithography which defines the regions on silicon wafers where machining is done. Machining includes etching, doping and deposition of thin films. Fabrication of three-dimensional structures with complex forms only in two dimensions is possible. Extension to 3D structures can be accomplished using wafer bonding of silicon wafers and etching process as will be discussed in the subsequent sections. Micromachining process can be broadly classified into three major categories: (i) Bulk micromachining, (ii) Surface micromachining and (iii) LIGA process and a collection of numerous and varied techniques that can produce structures and mechanisms on the micrometre scale such as Deep Reactive Ion Etching (DRIE) together with wafer bonding. We now take a look at these processes in some detail.

5.3.1 Bulk Micromachining Technology

Bulk micromachining of silicon involves fabrication of micromechanical devices by carving out the silicon wafer by (i) wet chemical etching and (ii) dry etching techniques using suitable protective etch mask layers.

Wet chemical etching of silicon

The wet chemical etching process can be isotropic as shown in Figure 5.1 where the etch rate in the X and Y directions is almost equal to that in the Z direction. Such isotropic profiles are obtained when single crystal silicon, or polycrystalline and amorphous silicon are etched in an etchant consisting of Hydrofluoric, Nitric and Acetic acids called the HNA solution. Evidently, this isotropic etching of silicon is not a suitable process for realising structures having well controlled geometries.

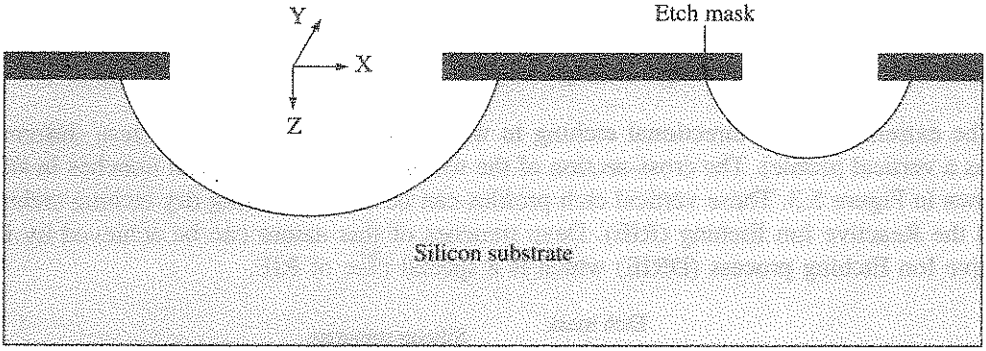


Figure 5.1 Isotropic etching of <100> silicon (cross section).

On the other hand, the use of anisotropic etching of silicon is the simplest approach for realising micromachined structures of silicon. This approach for bulk micromachining of silicon has been possible with the availability of several anisotropic (KOH, EDP and TMAH) etchants of silicon which etch single crystal silicon preferentially along given crystal planes. Anisotropic etchants for silicon are usually carried out elevated temperatures. The etch rates are orientation dependent and hence they etch the different crystal orientations with different etch rates. Anisotropic enchants of silicon etch the (100) and (110) crystal planes significantly faster than the (111) crystal planes. Hence as shown in Figure 5.2 during the anisotropic etching of (100) silicon, the etch rates in the Z direction are larger than in the lateral

(X or Y) direction. Etch rate ratios of about 400 : 1 for (100) to (111) orientations and about 600 : 1 for (110) with respect to <111> orientations have been observed. Typical silicon etch rates of about 60 μm/minute are achieved along the (100) direction with 30% KOH solution at 75°C temperature. Silicon dioxide, silicon nitride, and some metallic thin films (e.g., chromium, gold, etc.) provide good etch masks for typical silicon anisotropic enchants. These films are used to mask areas of silicon that are to be protected from etching and to define the initial geometry of the regions to be etched. Heavily-boron-doped silicon (above 7×10^{19} cm^{-3}), referred to as p^+ etch-stop, is effective in practically stopping the KOH etching some portions of the material. Thus using the etch-masks and etch-stop techniques to selectivity prevent regions of silicon from being etched, it is possible to fabricate microstructure in a silicon substrate by appropriately combining etch masks and etch-stop patterns with anisotropic enchants.

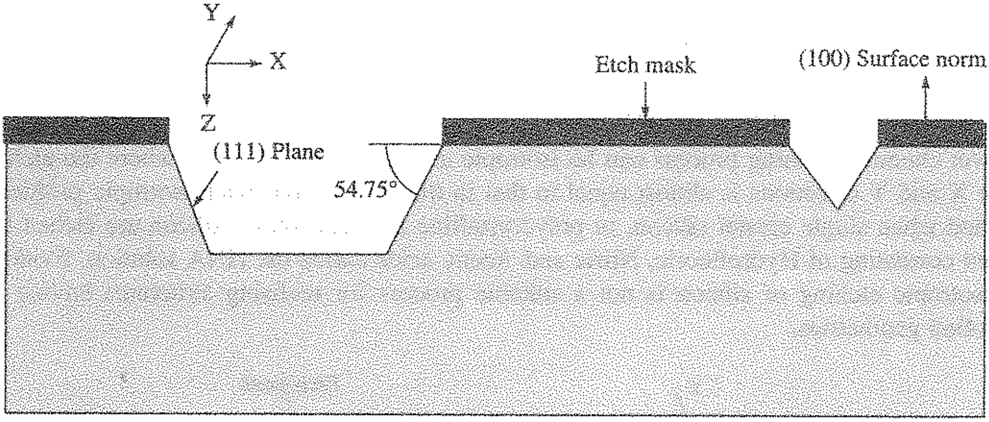


Figure 5.2 Anisotropic etching on <100> surface of silicon (cross-section).

The extreme case of directional etching in which the lateral etch rate is zero (referred to here as a vertical profile). The cross-section of the wafer having such vertical trenches in silicon is shown in Figure 5.3. These vertical etch profiles can be achieved using dry etching technique called the Reactive Ion Etching (RIE). Deep trenches of this nature can be achieved by Deep Reactive Ion Etching process (DRIE) which is a special case of RIE.

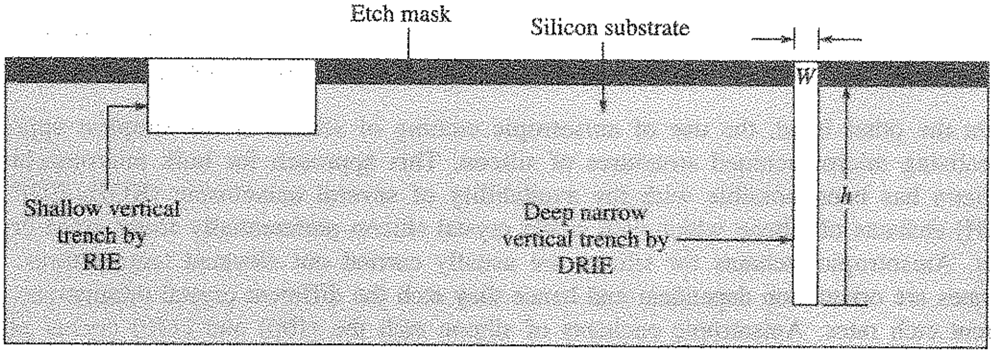

Etch mask
Silicon substrate
W
h
Shallow vertical trench by RIE
Deep narrow vertical trench by DRIE

Figure 5.3 Cross-section of silicon showing shallow vertical trench and deep narrow vertical trenches by Reactive Ion Etching (RIE) and Deep RIE (DRIE).

The concepts of bulk micromachining are illustrated in Figures 5.4 and 5.5 by showing the top view and cross-sections respectively of a through hole of predetermined size and a diaphragm of 5 micron to 10 micron thickness realised by bulk micromachining using anisotropic wet chemical etching of a (100) silicon substrate. As shown in these diagrams, for a (100) silicon substrate etching proceeds along the (100) direction, while it is practically stopped along the (111) planes which are at an angle of 54.75° with the (100) planes as shown in these figures. Due to the slanted (111) planes, the size W of the etch-mask opening, which is shown in Figure 5.5 determines the final size of the hole. For instance, as shown in Figure 5.5, if the etch mask openings are square of size W and the sides are aligned with the [110] direction [i.e. the direction of the intersection line between (100) and (111) planes], practically no undercutting of the etch-mask feature takes place, assuming the etch rate of the ($\bar{1}$11) planes is negligible. The size L of the square hole at the back side of the silicon wafer is related to the thickness H of the wafer by the trigonometric relation:

$$L = W - 2H \cot (54.75°) = W - 1.414H \quad (5.1)$$

A bulk micro-machined silicon diaphragm of thickness h and side width a, defined by a p^+ etch stop of thickness h, can be fabricated by etching from the topside of the (100) silicon wafer thickness H through an etch mask window of side W as shown in Figure 5.5. In this case, the bottom p^+ layer prevents this layer from etching by the anisotropic etchant, and hence the width of the diaphragm is given by the relation:

$$a = W - 2\,(H - h) \cot (54.75°) = W - 1.414\,(H - h) \quad (5.2)$$

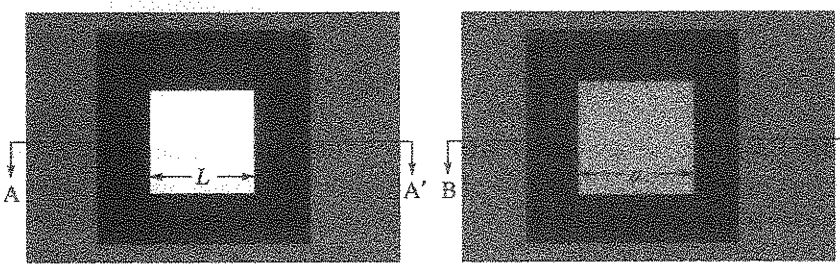


Figure 5.4 Top view of square hole of size L etched in KOH and the top view of square diaphragm of size a realised by KOH etch and p+ doped etch stop.

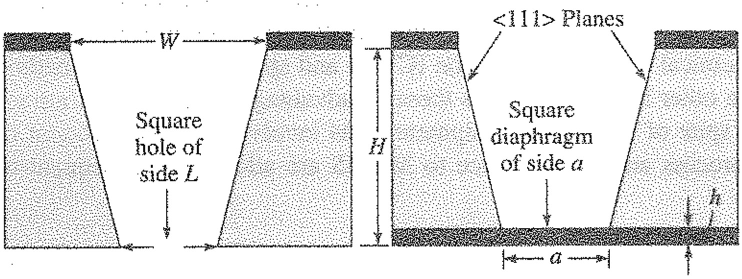


Figure 5.5 Cross-section A-A′ of KOH etched hole shown in Figure 5.4 and the cross-section B-B′ of diaphragm of Figure 5.4 showing the <111> planes delineated by KOH etching.

For convex corners, misalignment with the (110) direction, and curved edges in the etch-mask openings, significant under-etching may occur until etching is limited by the (111) planes. Therefore, the alignment of mask to crystal orientation is important in the anisotropic etching. Misalignment will change the etch-rate greatly. The under-etching characteristics can be utilised to fabricate suspended microstructure as shown in Figures 5.6 and 5.7. This gives an example of a bulk micro-machined silicon cantilever fabricated by undercutting of the convex corners of the beam geometry (which is defined by an etch stop) from the front side of a wafer. Anisotropic etchants for silicon are usually alkaline solutions used at elevated temperatures.

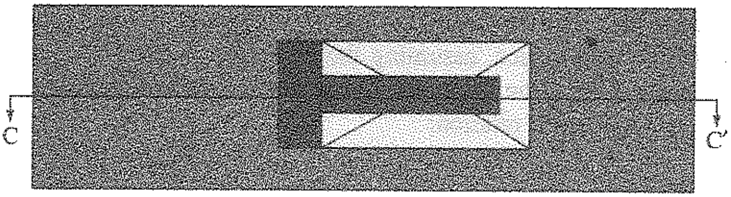


Figure 5.6 Top plan view of a cantilever beam realised by anisotropic etching using a boron etch stop layer and using anisotropic etching of silicon.

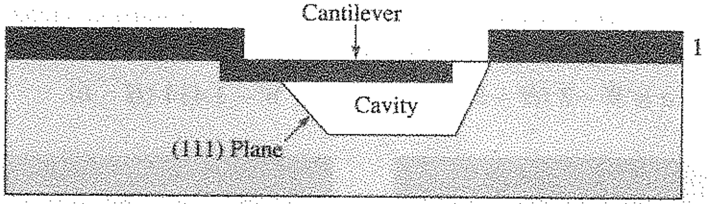


Figure 5.7 Cross-section C-C′ of the cantilever beam of Figure 5.6.

Mechanical structures of precise dimensions in the micro scale can be conveniently fabricated by the wet chemical anisotropic etching of single crystal silicon without the need for expensive equipment. However, the wet chemical etching process involves the use of corrosive acids or alkalis and requires disposal of waste products and this process is difficult to automate. In addition to the above disadvantages, the wet chemical etching process turns out to be isotropic when used for etching polycrystalline silicon, amorphous silicon and other such materials like SiO_2 and Si_3N_4. These disadvantages overwhelm the merits of wet chemical etching method, making this technique rather unattractive for processing thin film materials deposited by Chemical Vapour Deposition (CVD) and sputter deposition methods. Dry etching technique, on the other hand, overcomes these disadvantages and makes it attractive for MEMS related work, in spite of the higher equipment costs involved. A brief account of the principles of dry etching process and its relevance to MEMS are addressed subsequent to the following Example 5.1.

EXAMPLE 5.1 A (100) oriented p-type silicon wafer of thickness 300 μm is oxidised on both sides of the wafer and a square window of lateral dimensions 1 mm × 1 mm is opened in the oxide on one side of the wafer using photolithography. The edges of this window are along the <110> crystallographic directions. The wafer is immersed for 6 hours in KOH solution which etches silicon in the <100> direction at a rate of 50 μm per minute and the etch rate along <111> direction is zero. Assume that the oxide etch rate is negligibly small. At the end of 6 hours, the oxide layer is completely etched using a suitable chemical. Draw (i) the cross-section of the silicon wafer and (ii) the view of the wafer as seen from the top side of the wafer, at the end of the process.

Solution: The (111) plane intersects the (100) plane at 54.75° as shown in Figure E5.1 at the edges of the oxide window.

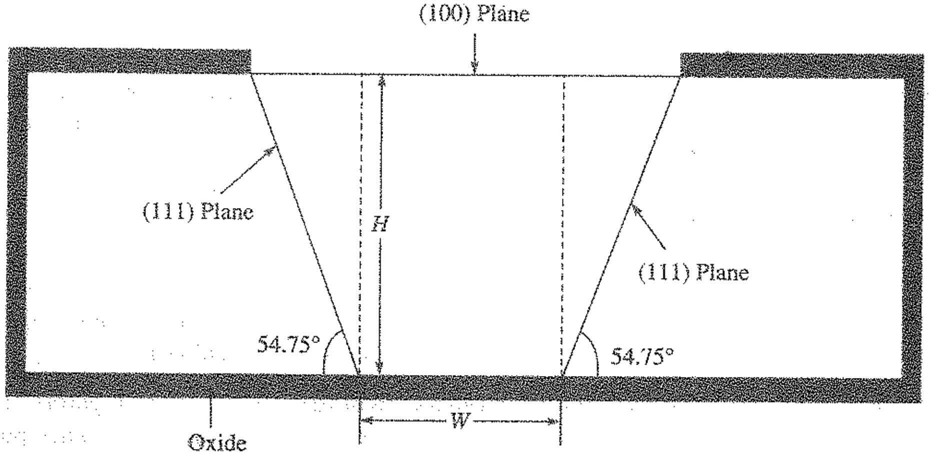


Figure E5.1

The oxide window width $W_o = W + 2H/\tan(57.5) = W + H\sqrt{2} = W + 1.414H$. In this problem, $H = 300$ μm and $W_o = 1$ mm $= 1000$ μm. The etch profile follows the (111) plane because the etch rate in the <111> direction is zero. As the etch rate along the <100> direction (i.e. perpendicular to the wafer plane) is 50 μm/minute, at the end of 6 hours, a square hole of size $W \times W$ will take place at the opposite side of the wafer $W = W_o - 1.414H = 1000 - 424 = 576$ μm. At the end of the process, we have cross-section and top view as shown in Figures E5.2 and E5.3.

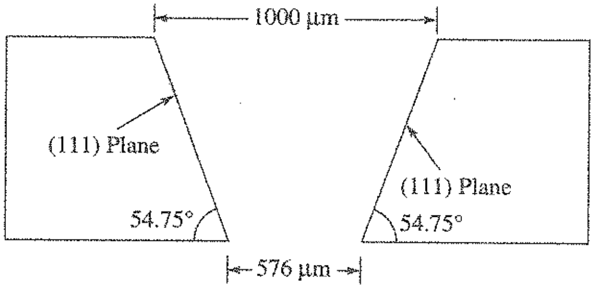


Figure E5.2 Cross-section.

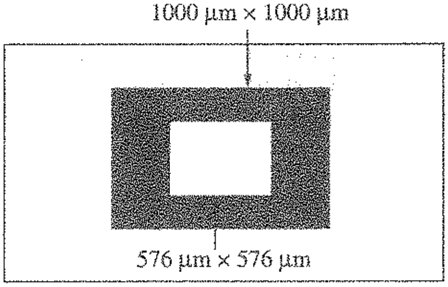


Figure E5.3 Top view.

The dark region in the top view is the (111) plane at an angle of 54.75° to the wafer plane. In fact, under the microscope, this region will appear dark because the light does not get reflected back from these angled surfaces.

Dry etching technique

Dry etching is invariably identified with plasma etching. This is due to the fact that all the dry etching process involves gas plasma of either a chemically active or inert species. Mainly there are three types of dry etching processes based on the etching mechanism, namely, (i) physical basis, (ii) chemical basis, (iii) process based on combination of (i) and (ii).

Dry etching based on physical removal (sputter etching): In the case of purely inert gas plasma, process involving heavy ions such as argon , the etching process takes place purely by physical sputtering of the host atoms either in *an Ion Beam Etching* (*IBE*) or *a glow discharge sputtering* process. This process is a directional etch, hence is highly anisotropic (as shown in Figure 5.8) and does not distinguish between different layers, thus making it useful for etching multilayer structures. Due to the poor selectivity of the process, since the etching takes place irrespective of the material type, it becomes difficult to adopt for selective etching some portions of the substrate. This difficulty has been successfully tackled by using thick masking layer of photoresist as shown in Figure 5.8. As the excitation energy of the heavy ions is very high in the range 1 keV to 3 keV, radiation damage is possible with this process and would lead to deterioration of the characteristics of the active devices.

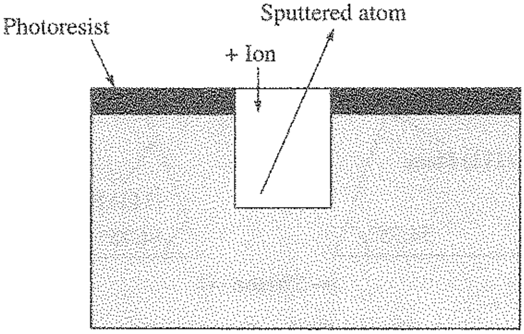


Figure 5.8 Cross-section of substrate after ion beam etching.

Dry etching based on chemical reaction (plasma etching): When the plasma etching takes place purely by the chemical reaction of the radical with the substrate, the etching process is called the *plasma etching*. Thus in this process, fluorine containing gas such as SF_6 or CF_4 are extensively used for etching silicon, silicon dioxide, silicon nitride, etc. The most abundant ionic species found in CF_4 plasma is $CF_3{}^+$, and such ions are formed by the ionisation and dissociation reaction given by

$$e^- + CF_4 \longrightarrow CF_3^+ + F + 2e^- \tag{5.3}$$

$$e^- + CF_4 \longrightarrow CF_3 + F + e^- \tag{5.4}$$

Radicals exist in plasmas in much higher concentration than ions because they generate at faster rate, and they survive longer than ions. The free radical such as F or CF_3 is neutral and exists in a state of incomplete chemical bonding making it very reactive. With silicon, the following etching reactions takes place:

$$SiF_x + F \longrightarrow SiF_{(x+1)} \cdot (x = 0 \text{ to } 3) \tag{5.5}$$

In the plasma etching process, an RF glow discharge produces chemically reactive species (atoms, radicals and ions) from a relatively inert molecular gas. The etching gas is selected to generate species which react chemically with the material to be etched, and whose reaction products are volatile so that they can be vented through the exhaust. Thus the process is based purely on chemical reaction. Hence it is highly selective against both mask layer and the underlying substrate layer. As this process relies only on chemical mechanisms for etching, the reaction is isotropic and the process does not provide a solution to the problem of undercutting as shown in Figure 5.9. However, it is suitable for photoresist stripping using oxygen gas plasma in a simple barrel type plasma reactor.

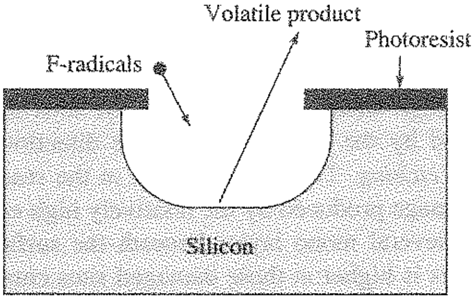


Figure 5.9 Cross-section of silicon wafer after isotropic plasma etching.

Reactive Ion Etching (RIE): The reactive ion etching is a plasma etching system which uses chemical reaction for etching in the presence of energetic ions. In this process, the substrate to be etched is placed on an electrode C (cathode) of area A_c which is driven by the RF power supply. The area A_a of the other electrode A (anode) is the chamber wall and is larger compared to the electrode C. This larger electrode is grounded. A schematic diagram of the system is shown in Figure 5.10.

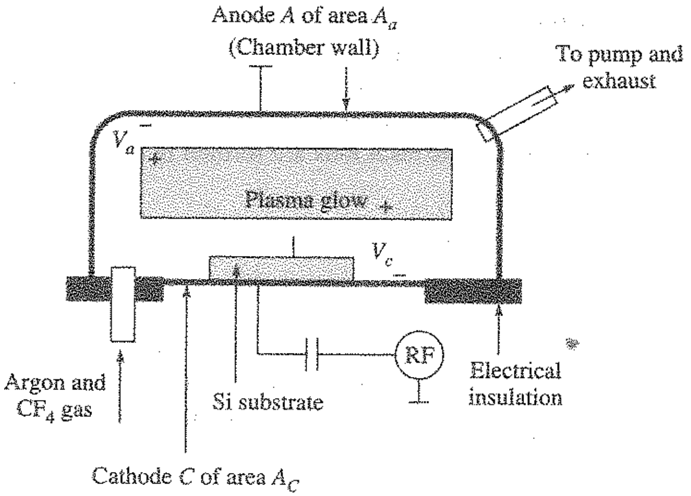


Figure 5.10 Schematic diagram of reactive ion etching.

When an RF voltage is applied as shown in Figure 5.10, a plasma is generated in the gas mixture containing reactive gas (CF_4 or SF_6) and an inert gas like argon, and anode A and cathode C acquire negative voltages V_a and V_c with respect to the plasma due to the relatively faster movement of the electrons and the higher mass of the ions which almost get confined to the plasma core region. As the current flowing through these electrodes are same, the current density in the smaller area A_c of cathode is higher than that in the larger area A_a of anode. Hence $V_c > V_a$ as related to the area of the electrodes by the relation:

$$\frac{V_c}{V_a} = \left(\frac{A_a}{A_c}\right)^n \tag{5.6}$$

In practice n takes a value between 1 and 2. The substrate to be etched is placed on this cathode which develops high sheath voltage where the ions acquire energy and assist in enhancing etch rates produced by the reactive gas in the directional etching. Therefore, this process is called *reactive ion etching*. This enhancement in the reaction rate is experimentally demonstrated to the lattice damage produced by the relatively high energy impinging argon ions (>50 eV) at the surface and several mono layers beneath the surface being etched. Chemical reaction by the reactant species is higher at these damaged sites compared to the regions where no damage has occurred. As the ion bombardment is highly directed perpendicular to the cathode where the substrate is placed, the side walls receive much smaller flux of the bombarding high energy ions and do not get damaged. Hence the reaction rate at the side walls is considerably smaller than in the vertical direction where the ion bombardment is received. This difference in the lateral and vertical etch rates makes the RIE process highly anisotropic. Etch rates in the range 40 nm–100 nm to microns per minute can be achieved with this process. This RIE process is restricted to shallow trench applications. When used for deeper trenches, the lateral etching occurs as shown in Figure 5.11, due to long exposure of the side walls to the reactive species. For nanoscale applications requiring deep trenches and excellent anisotropy over a long duration of etching, the RIE is modified by a process called the Deep RIE (DRIE).

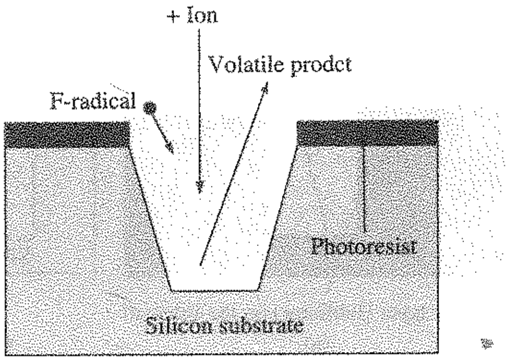


Figure 5.11 Cross-section of silicon substrate.

Deep Reactive Ion Etching (DRIE): The deep reactive ion etching process is usually referred to as the Bosch process and is based on patents currently owned by Robert Bosch GmbH and Texas Instruments. This process employs repeated cycles that consist of an etching step followed by a coating (passivation) step to protect the side wall during the etching cycle. The process is carried out in an Inductively Coupled Plasma (ICP) to achieve high density of plasma at high power levels of several hundreds of watts, so that very deep trenches of 450 microns having high aspect ratio features with near vertical sidewalls can be produced in short durations of 15 minutes. High plasma density of 10^{11}–10^{12} per cm^3 which is about two orders of magnitude higher than in the conventional RIE systems can be achieved with ICP to create a magnetic envelope inside the etch chamber which reduces the loss of charged species to the surrounding.

In the DRIE process, during the etching cycle, the ICP plasma in a gas mixture of SF_6 and argon is used. A negative bias (–5 V to –30 V) is applied during the etching step so that the positive ions generated in the plasma are accelerated vertically into the substrate being etched. During the passivating cycle, all the exposed surfaces (sidewalls and horizontal surfaces) are coated with teflon—like polymer layer of thickness 50 nm, by switching the gas mixture to C_4F_8. During the subsequent etching cycle, the polymer formed on the horizontal surfaces of the etched trench get etched due to ion bombardment, whereas the coatings on the sidewalls do not get etched because the ions are highly directed on to the horizontal surfaces. Hence during the etching cycle, the sidewalls do not get etched due to the presence of polymer layer on the sidewalls as shown in Figure 5.12(a). This is called the *Bosch process* named after the company which invented this process. In some of the recent commercial DRIE systems, the etching and coating cycle time can be made as small as 1–2 seconds to achieve vertical sidewall with minimum scallop on the sidewalls. The size of the scallops depends on the cycle time. Typical scallop depths are 10–20 nm. Figure 5.12(b) shows a schematic diagram that can be seen when viewed with a high magnification SEM.

With DRIE, silicon etch rate of 30 μm per minute have been achieved using thermal oxide of 1 micron thickness as masking layer to etch 350 micron deep trenches in silicon. Etch depths of the order of 500 μm to 1000 μm can be obtained using 4-micron to 5-micron thick photoresist as the masking layer. Aspect ratio defined as trench height *h*, width *w* (see Figure 5.13) up to 30 : 1 (with sidewall angles 90° ± 2°) can be achieved. Photoresist selectivity of 50 to 100 : 1

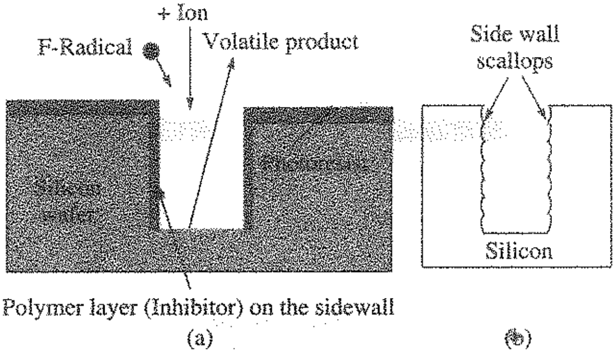


Figure 5.12 (a) Schematic cross-section during the DRIE, showing the etched trench and the passivating polymer coating (inhibitor) (b) Schematic cross-section showing the scallops in DRIE.

and silicon dioxide selectivity of 120 to 200 : 1 with silicon have been achieved due to the high plasma density of 10^{11}–10^{12} per cm^3 which is about two orders of magnitude higher than that in the conventional RIE systems. DRIE and RIE are very useful techniques for fabricating MEMS structures such as accelerometers and gyroscopes on polycrystalline silicon layers without any undercut etching.

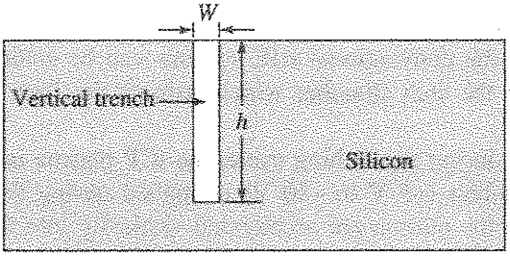


Figure 5.13 Schematic illustrating the concept of aspect ratio of trench in Si.

Wafer bonding techniques

Wafer-level bonding of a silicon wafer to another silicon substrate or to a glass wafer is an important technique used in bulk micromachining and plays a key role in all the leading-edge MEMS devices. When used along with the wet or dry etching techniques, the wafer bonding technique can be used to realise (i) membranes of thickness varying from couple of microns to several microns, suitable for pressure sensors over a wide range of pressures, (ii) complicated three-dimensional structures for accelerometers for sensing acceleration, (iii) multilayered device structures such as micro pump suitable for biomedical and micro-fluidic applications, and (iv) other high aspect ratio structures. The manufacturers of MEMS require wafer-level bonding of one silicon wafer to another silicon substrate or a glass wafer. This provides a first level packaging solution that makes these processes economically viable. In this section, the various silicon wafer bonding techniques are presented and their role on MEMS devices such as accelerometers and gyros are illustrated in Sections 5.5 and 5.6.

Silicon wafer bonding for MEMS is achieved by several different approaches such as *anodic bonding*, *direct bonding* and *intermediate layer bonding* which includes eutectic and glass-frit bonds. Even though, the process conditions used for all the three-bonding techniques vary, the general process of the wafer bonding follows a three step sequence consisting of surface preparation, contacting and annealing.

Anodic bonding: It involves bonding a silicon wafer and a glass wafer having high content of sodium. Figure 5.14 shows the schematic of the anodic bonding arrangement. The anodic bonding is carried out at 450°C by applying a high negative voltage in the range 500–1000 V to glass with respect to the silicon wafer which is held in contact with it. The Na^+ ions move away from the interface towards the negative electrode where they are neutralised. This leads to the formation of a space charge at the glass silicon interface, and creates a strong electrostatic attraction between glass and silicon wafer and enables the transport of oxygen from glass to the glass–silicon interface and converts silicon to SiO_2 creating a permanent bond. Processes for anodic bonding of silicon to bulk glass and silicon to silicon using thin glass layer have also been reported. Typically Pyrex 7740 or Schott 8330 glass are used. The Thermal Expansion Coefficient (TEC) of these glasses match closely with the TEC of silicon, resulting in low stress in the bonded devices.

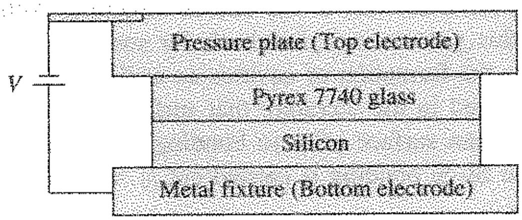

Pressure plate (Top electrode)
V
Pyrex 7740 glass
Silicon
Metal fixture (Bottom electrode)

Figure 5.14 Silicon–glass anodic bonding arrangement.

Silicon Direct Bonding (SDB): It is usually referred to as *Silicon Fusion Bonding (SFB)* and is used for bonding two or more silicon wafers. It is based on the initial bonding by hydroxyl radicals present on the silicon wafer surfaces prepared by standard cleaning process prior to bonding. Mechanical spacers are placed at the edges of the wafers (Figure 5.15) to maintain physical separation, so that pressing the middle of the wafers creates an initial point contact that originates the bond. Removing the mechanical spacers (Figure 5.16) allows a single bonding wave to propagate from the centre of the wafers. The mechanical spacers are important in establishing a single bond wavefront that propagates outward because multiple bonding

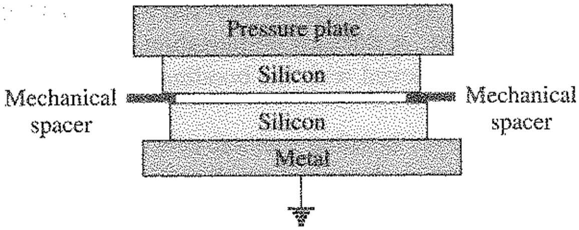


Figure 5.15 Silicon fusion bonding step with wafers placed in position with spacer.

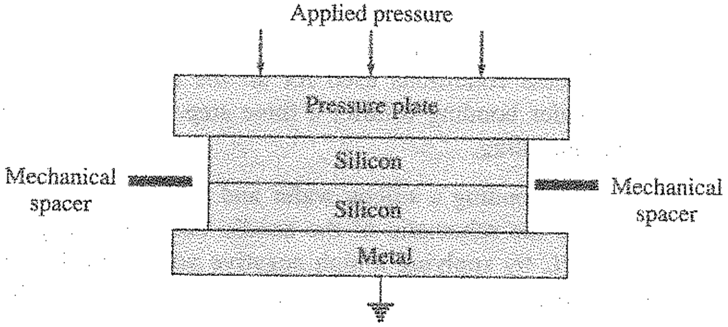


Figure 5.16 Silicon fusion bonding after the two wafers are bonded by removing the mechanical spacers and pressurising.

waves lead to warpage and gases can be trapped in pockets formed by multiple waves, and result in areas of poor bonding. After this pre-bonding step, subsequent annealing is carried out at temperatures in excess of 1000°C for durations ranging from two hours to five hours. During this annealing step, the hydroxyl groups from water molecules create Si–O–Si bond as hydrogen diffuses away. Oxygen also diffuses into the crystal lattice to create a bond interface that is not distinguishable from the rest of the silicon structure. Although the high annealing temperature involved in this process is a drawback for some applications, the silicon fusion bonding technique permits the formation of cavities as well as all—silicon, stress free bonded structures. It has been reported in the literature that surface activation methods such as argon beam etching to create a clean surface prior to bonding result in excellent bond strengths of 10–12 MPa even when the Si–Si bonding is carried out at room temperature.

Intermediate-layer bonding: This technique involves deposition of either glass or metallic intermediate films prior to bonding two silicon wafers, and they are referred to as glass-frit bonding and the eutectic bonding. The eutectic bonding makes use of the existence of a eutectic melting temperature which is considerably lower than the melting point of individual constituent elements. For gold and silicon system, the eutectic melting point is 363°C and corresponds to a eutectic composition of 3.16 per cent silicon and 96.84 per cent gold by weight (19 per cent silicon and 81 per cent gold by atomic per cent). The eutectic bond is performed by evaporating and plating gold on to one of the silicon wafers and then exposing the gold to UV light just before bonding to remove organic contaminants that preclude gold surface contact with the second silicon wafer into which it is bonded. To accomplish good bond, the second silicon wafer surface preparation must remove any oxide film that can hamper diffusion of gold into silicon. The eutectic bonding method uses pressure applied with the wafers held at a temperature slightly higher than the eutectic temperature. A detailed optimisation study has revealed that maximum bond strength of 18 MPa can be achieved with the bonding temperature of 400 °C and the gold layer thickness of 1.0 μm.

In the another intermediate-layer bonding method called the glass-frit bonding process, a thin glass layer such as lead borate is deposited on the silicon substrate. The wafers are then brought into contact under pressure at the melting temperature of the glass, which is generally < 600 °C.

5.3.2 Surface Micromachining Technology

As pointed out in the previous section, it is rather difficult to adapt the bulk micromachining method to MEMS processing with the conventional top-down approach. Contrary to this, the surface micromachining technique, which we briefly describe in this section, overcomes this problem and also makes it easy for integration with electronics. Figure 5.17 describes the concept of surface micromachining to fabricate an oxide anchored polysilicon cantilever beam. In this approach 'structural layer', typically, polycrystalline silicon is deposited on an easily removable material called the 'sacrificial layer' (e.g. $\mathrm{SiO_2}$) grown or deposited on silicon substrate as shown in Figure 5.17(a). The structural layer is then patterned as shown by the top view and cross-section in Figure 5.17(b) to form the cantilever and the anchor region. In the next step, called the release step (e.g. etching $\mathrm{SiO_2}$ the sacrificial layer in hydrofluoric acid) is performed, creating a cantilever suspended over the substrate at a height equal to the thickness of the sacrificial layer. The top view and the cross-section of the released cantilever beam are shown in Figure 5.17(c).

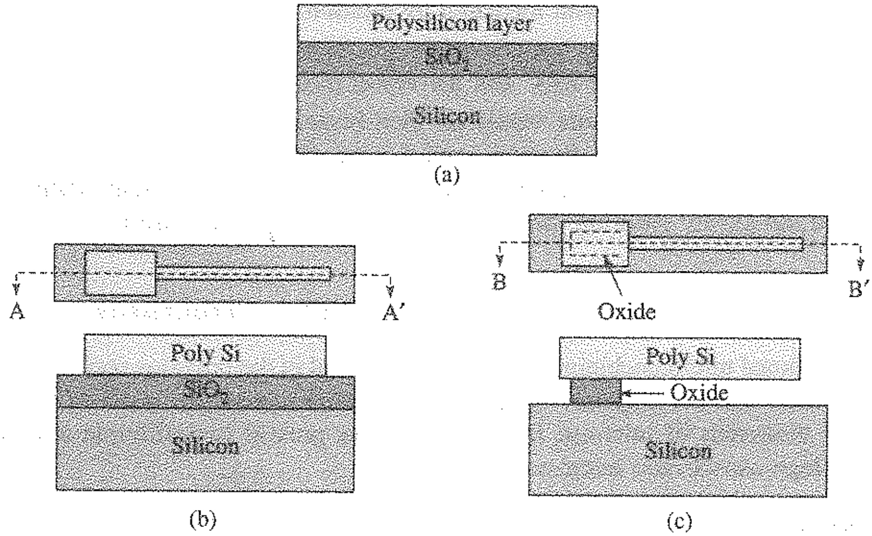


Figure 5.17 (a) Cross-section after depositing $\mathrm{SiO_2}$ and polysilicon on silicon substrate (b) Top view and cross-section A-A′ after patterning polysilicon (c) Top view and cross-section at B-B′ after beam release.

The key to the success of this one-mask process is the proper selection of the lateral dimensions of the structural parts to be released such that the sacrificial oxide layer is fully removed from under these parts, while the anchor regions are only partially under-etched after the release as shown in Figure 5.17(c). This is achieved by selecting larger lateral dimensions for the anchor regions as compared to the beam width that must be fully released. Since the anchor regions are also undercut during release, the release process should be carefully controlled to ensure that the anchor region is not fully undercut. Care must be taken during the design stage itself to ensure that the anchor size is larger than the widths of all devices on the same wafer or die so that during the release process the anchor regions do not experience serious undercut problems.

In the surface micromachined process, the lateral dimensions of the structural layer depends upon the ability to generate patterns in the micro or nanoscale and the ability to etch the structural layer and release. With the electron beam lithography process and the use of dry etching techniques, this has indeed been made possible to achieve resonating structures with micro/nano scale geometries. However, a major problem in the surface micromachining approach is 'stiction' which is due to the released beam getting pulled to the substrate and getting stuck to the substrate by the surface tension force arising when the water used for rinsing the HF acid begins to dry.

Super critical drying process

The 'stiction' problem has been tackled using super critical drying process or simply called the Critical Point Drying (CPD) and involves the following steps:

(i) After HF etching the sacrificial oxide layer, the structure is rinsed in de-ionised water.

(ii) The water is exchanged with methanol by dilution.

(iii) The silicon wafer is then transferred to pressure vessel in which the methanol is replaced by liquid CO_2 at 25°C and the vessel pressure is raised to about 1200 psi.

(iv) The contents of the pressure vessel are then heated to 35°C. At this temperature and pressure, the CO_2 reaches the super critical stage. In this state, the liquid and gas phase are indistinguishable and hence the surface tension force is absent when the CO_2 is vented out slowly ensuring that it exists in the gaseous form itself.

5.3.3 Processes for High Aspect Ratio MEMS Structures

With the conventional micromachining techniques discussed in the previous section, it is difficult to achieve metallic and polymer microstructures with thickness of the order of few hundreds to few tens of thousands of microns. In this section we present two different opproaches for realising such high aspect ratio microstructures.

LIGA process

LIGA is a German word for *Lithographie Galvanoformung Abformung* which stands for *Lithography, electroplating* and *moulding*. This process is used mainly for realising high aspect ratio metal microstructures (i.e. thickness is high compared to the lateral dimensions). This process involves X-ray lithography and electroplating. High energy collimated X-rays from synchrotron is used for this purpose. The process is best illustrated in Figure 5.18.

A major benefit of LIGA process is that high aspect ratio metal microstructures can be built. The requirement of synchrotron, complicated mask for X-rays and long exposure times of photoresists such as Poly Methyl MethAcrylate (PMMA) makes LIGA unsuitable for large scale use. LIGA is good for small parts, with additional requirement of assembly to realise most of the useful devices. As a result, the use of photosensitive polyimide and special types of photoresists such as SU-8 are becoming popular.

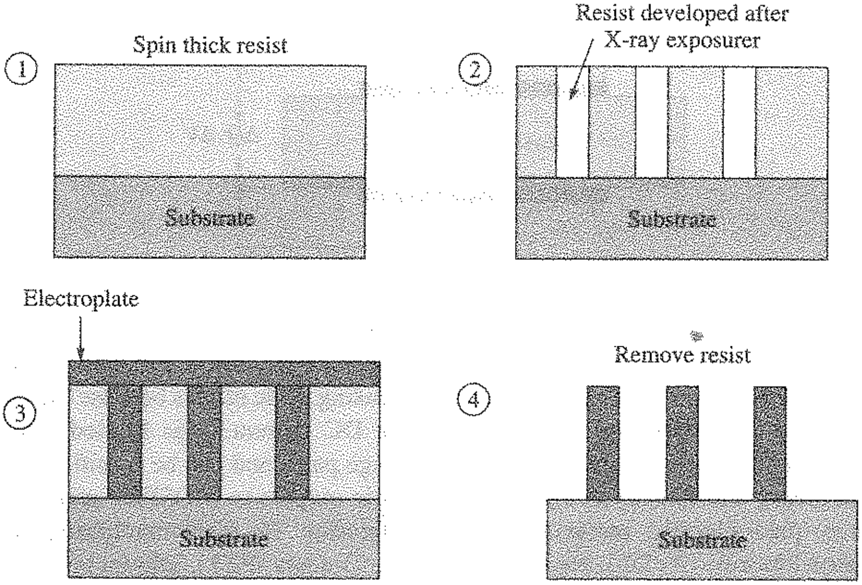


Figure 5.18 Illustration of LIGA process.

SU-8 Process (an alternative to LIGA): SU-8 is a low-cost negative resist for MEMS developed by IBM. Processing uses standard lithography equipment: standard UV light source and standard photo-masks on glass plate. Structures up to several microns thick have been demonstrated with high aspect ratio greater than 15. SU-8 process has several merits compared to LIGA:

- Processed layers as thick as 100 μm can be achieved.
- They offer new capabilities for masking moulding and building high aspect ratio structures at low-cost.
- The cost of SU-8 lithography is considerably lower than that of other techniques such as LIGA process and DRIE for realising high aspect ratio microstructures.
- SU-8 has been integrated in number of microdevices, including micro-fluidic devices.

5.4 Micromachined Sensors

Micromachined sensors can generally be classified into the following three categories based on the principle of operation: (i) Piezoelectric (ii) Piezoresistive and (iii) Capacitive.

5.4.1 Piezoelectric Sensors

In the piezoelectric effect, a mechanical stress on a material produces electrical polarisation and reciprocally, an applied electric field produces a mechanical strain. Figure 5.19 is an illustration showing the generation of incremental charge, and hence voltage on metallised electrodes on opposite sides of a slab of piezoelectric material in response to the application of force.

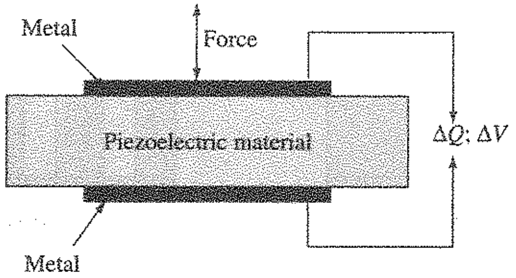


Figure 5.19 Illustration of piezoelectric effect.

This effect can thus be used to sense mechanical stress, force and also as an actuation mechanism. However, this effect cannot be used for sensing static effects because the piezoelectric effect produces a DC charge (polarisation) but not DC current. This limited low frequency response is primarily due to the parasitic charge leakage paths and can be significantly improved by micromachining and directly coupling piezoelectric outputs to MOSFET gates.

Common piezoelectric materials with application in micromachining are quartz, Polyvinylidene Fluoride (PVDF), lead (Pb) Zirconate Titanate (PZT), lithium niobate, ZnO, etc. Among these materials, PZT is a ceramic with a high value of piezoelectric constant. However, it is difficult to deposit as a thin film. PVDF and ZnO are often used in the microfabrication of piezoelectric sensors and actuators.

A major disadvantage of piezoelectric material which makes it unattractive for microfabrication is that it is not used in the standard microelectronics technology. In addition to this, piezoelectric effect can be used effectively in dynamic conditions only and cannot be used for sensing static pressure or force.

5.4.2 Piezoresistive Sensors

In the piezoresistive sensors, the physical phenomena such as pressure, acceleration, etc. are sensed by a change ΔR in the resistance R of a piezoresistor when it experiences a strain ε, due to the stress developed in the membrane or the beam on which the resistor is placed. The sensitivity of a piezoresitive sensor depends upon its gauge factor G which is related to strain as follows:

$$G = \frac{\Delta R}{\varepsilon R} \tag{5.7}$$

The resistance R is related to its resistivity ρ and the physical parameters such as length, L and the lateral dimensions viz., width W and thickness t, by the relation:

$$R = \rho \frac{L}{Wt} \tag{5.8}$$

When subjected to strain, the resistance change ΔR takes place due to the changes in all the three parameters. Therefore, we can write:

$$\frac{\Delta R}{R} = \frac{\Delta \rho}{\rho} + \frac{\Delta L}{L} - \frac{\Delta W}{W} - \frac{\Delta t}{t} \tag{5.9}$$

It is well known that when a beam is stretched along the length, the increase in length (ΔL) is accompanied by a decrease in the lateral dimensions and is related to ΔL by the Poisson ratio v and is written as

$$\frac{\Delta W}{W} = \frac{\Delta t}{t} = -v\frac{\Delta L}{L} \tag{5.10}$$

The Poisson ratio is less than 1.0 and is usually in the range 0.2 to 0.5.

Using Eq. (5.10) in Eqs. (5.9) and (5.7), the expression for gauge factor G turns out to be as follows:

$$G = \frac{\Delta\rho}{\varepsilon\rho} + (1+2v) \tag{5.11}$$

In a metal strain gauge, the resistance change ΔR takes place mainly due to the change in its length and area of cross-section when it is subjected to strain and negligible change takes place in the resistivity ρ. Hence the gauge factor in metallic strain gauges is in the range 2 to 5. On the other hand, in a semiconductor piezoresistors, the change in resistance ΔR occurs due to the change in the resistivity ρ of the material in addition to the change in the physical dimensions of the resistor. As a result, the gauge factor of semiconductor like single crystal silicon is in the range 100 to 200 depending upon the doping density. At higher doping densities, the gauge factor is lower than at lower doping concentrations. The semiconductor piezoresistors, therefore, are superior to the metallic strain gauges in several ways:

(a) As the gauge factor is high, the sensitivity of a semiconductor piezoresistive based sensor is considerably higher than its metallic strain gauge.

(b) The piezoresistors can be laid out on semiconductor beams or membranes. Hence the process is compatible with the conventional microelectronics process.

(c) The semiconductor piezoresistors can be miniaturised by the microfabrication process by defining the resistor region by photolithography and implantation or diffusion of the dopants. Thus it is possible to achieve miniaturised sensors which have several merits elucidated in Section 5.1.

High sensitivity is achieved by connecting four piezoresistors in the form of a Wheatstone bridge which is balanced, when the stress is absent. The locations of the resistors on a diaphragm (in the case of a diaphragm based pressure sensor) and on a beam (in the case of accelerometer having a mass suspended by beams) are selected such that the resistors of the opposite arms of the bridge increase by ΔR and the other two opposite arms decrease by ΔR when sensing the stress in either as a pressure sensor or as an accelerometer. It can be easily shown that the output V_o of such a sensor is given by the relation:

$$V_o = \frac{\Delta R}{R} V_i \tag{5.12}$$

The advantage of a piezoresitive sensor is that it is simple to fabricate, it is compatible with the microfabrication process and the output is electrical. Hence it can be used straight away. Piezoresitive pressure sensors and accelerometers are well developed and widely used in industry and consumer applications due to their high sensitivity, high linearity and ease of signal processing. However, piezoresistive sensors have some drawbacks:

(a) The output voltage of a piezoresistive sensor is temperature sensitive due to the Temperature Coefficient of Resistivity (TCR) and also the Temperature Coefficient of the Sensitivity (TCS) itself due to the temperature dependence of the gauge factor which again is related to TCR. This can be minimised by using piezoresistors with doping concentrations in the range of $10^{19}/\text{cm}^3$. However, at such high doping concentrations, the gauge factor is low, and hence the sensitivity gets reduced.

(b) Piezoresistors are usually fabricated by implantation or diffusion of p-type dopants on to a n-type diaphragm, and these junctions provide the isolation between the resistors. Therefore, they are susceptible to junction leakage and can cause stability problems particularly at higher operating temperatures. This can be overcome using oxide isolation between the resistors with Silicon On Insulator (SOI) approach.

(c) Piezoresistive sensors are basically sensitive to stress. Therefore, their performances are closely related to the packaging technologies and may lead to large offset voltage and a temperature drift of the output voltage.

EXAMPLE 5.2 p-type piezoresistors R_1, R_2, R_3 and R_4, each of them equal to $R = 1$ kohm, are arranged (Figure E5.4) on oxide grown on a thin square membrane anchored at the edges. Assume that these resistors have longitudinal gauge factor $G_L = 30$ and that the transverse gauge factor G_T is negligibly small ($G_T = 0$). These resistors are connected in the form of a Wheatstone bridge as shown in Figure E5.5. For this pressure sensor, (i) derive an expression for the output voltage V_o in terms of R, ΔR and V_i, mark the polarity of V_o for the case when the membrane is subjected to a uniform pressure directed into the plane of the paper. (ii) Determine V_o for an input voltage of 10 V when the longitudinal strain on the resistors R_1 and R_3 are 10^{-3} and the longitudinal strain on R_2 and R_4 are zero.

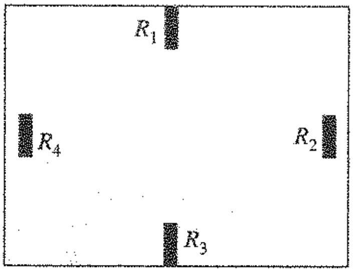


Figure E5.4

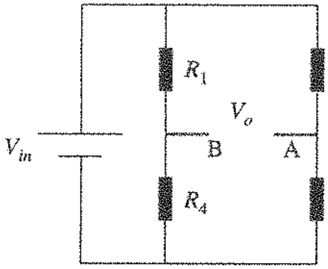


Figure E5.5

Solution: Referring to the Wheatstone bridge in Figure E5.5, $R_1 = R_2 = R_3 = R_4 = R$. As the p-type resistor marked R_1 and R_3 experience longitudinal tensile strain, their resistance increase by ΔR determined by the longitudinal tensile strain and the gauge factor in the longitudinal direction (i.e. along the direction of the current flow). Thus, for these two resistors

$$\frac{\Delta R}{R} = G_L \varepsilon$$

Here ε is the longitudinal tensile strain on resistors R_1 and R_3. Substituting $G_L = 30$ and $\varepsilon = 10^{-3}$, we obtain

$$\frac{\Delta R}{R} = 30 \times 10^{-3} = 0.03$$

(i) The resistance values of R_2 and R_4 do not change because the longitudinal strain on them is zero and the transverse gauge factor is zero.

$$\therefore \quad V_o = V_A - V_B$$

$$= V_i \left[\frac{R_3}{R_2 + R_3} - \frac{R_4}{R_1 + R_4} \right] = V_i \left[\frac{R + \Delta R}{2R + \Delta R} - \frac{R}{2R + \Delta R} \right]$$

$$= V_i \frac{\Delta R}{2R + \Delta R} \approx V_i \frac{\Delta R}{2R}$$

(ii) Substituting $V_i = 10$ V and $\dfrac{\Delta R}{R} = 0.03$, we obtain

$$V_o = 150 \text{ mV}$$

5.4.3 Capacitive Sensors

The capacitive sensor detects deflection of moving plate or mass with respect to a fixed plate. by sensing the change in capacitance The change in the capacitance can take place either due to the change in the gap or the change in the overlap area between the plates.

Merits of capacitive sensors

(i) The capacitive sensing schemes are not related to the mechanical properties of the material which are more stable than the piezoresitive properties. The absence of temperature coefficient of sensitivity of capacitive sensors makes it very attractive particularly for space and inertial navigation applications.

(ii) High sensitivity and high resolution accelerometers can be achieved with capacitive sensing using bulk micromachining technologies.

(iii) The capacitive accelerometers based on surface micromachining allow low cost mass production and have been widely used in applications such as air-bag deploying systems in automobiles.

Drawbacks of capacitive sensing

(i) Capacitive sensors are inherently nonlinear.

(ii) Measurement of small capacitors of a miniaturised structure is very difficult due to the parasitic and stray capacitances and the electromagnetic interference from the environment.

(iii) As the output is capacitance change, electronics circuit is a mandatory requirement for converting the capacitance to voltage.

5.5 MEMS Accelerometers

Accelerometers are sensors that measure and monitor linear acceleration along one or several axes. Over the past several years, these sensors were mainly used in the inertial navigation systems and cost of a sensor was the inhibiting factor. Today the micromachined accelerometers find their place in almost all walks of life and mainstream products because of their miniaturisation and the resulting cost reduction and enhanced reliability. These applications include the automobile industry for front and side air bag crash sensing. The accelerometers are also used as vibration sensors for health monitoring for engine management, shock and impact monitoring as well as for monitoring seismic activity. The accelerometers work on different transduction principles and are available in different configurations. In this section we discuss the principle of operation of these accelerometers and their design considerations and configurations.

5.5.1 Operation Principle and Parameters

All accelerometers consist of an inertial mass suspended from a spring as shown in Figure 5.20. The inertial mass gets displaced from its equilibrium position due to the effect of an externally applied acceleration. This displacement is a measure of the acceleration.

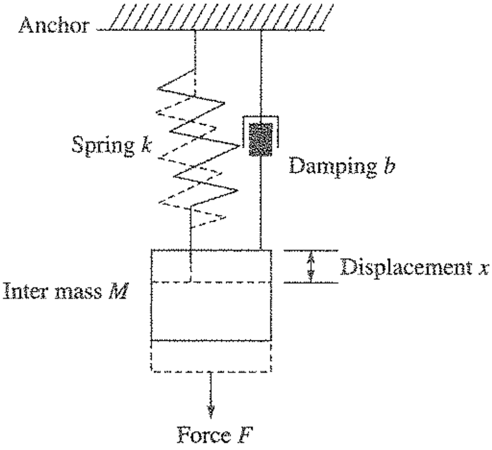


Figure 5.20 Spring–mass system illustrating the accelerometer concept.

The equation governing the response of the spring–mass system in general to an applied force F or the acceleration a is governed by mass M, spring constant k and the damping coefficient b as follows:

$$M\frac{d^2x}{dt^2}+b\frac{dx}{dt}+kx=F=Ma \tag{5.13}$$

Solving this differential equation, the resonance frequency f_r of the spring–mass system is given by the relation:

$$f_r=\frac{1}{2\pi}\sqrt{\frac{k}{M}} \tag{5.14}$$

From Eq. (5.13), we can see that under steady state conditions, the displacement $x = \delta$ is related to force F as follows:

$$F = Ma = k\delta \tag{5.15}$$

The sensitivity S of the accelerometer is expressed by the relation:

$$S = \frac{\delta}{a} = \frac{M}{k} \tag{5.16}$$

From Eqs. (5.14) and (5.16), it can be seen that the sensitivity can be increased by reducing the spring constant or/and increasing mass M, and thus by reducing the resonance frequency. Thus for an accelerometer useful for inertial navigation systems, the resonance frequency will be low. In addition to the sensitivity and the resonance frequency, the Total Noise Equivalent Acceleration (TNEA) a_{noise} is an important parameter. This parameter is defined as follows:

$$a_{\text{noise}} = \sqrt{\frac{8\pi k_B T f_r}{QM}} \qquad (f_r < B) \tag{5.17}$$

where

k_B = Boltzmann constant
T = temperature in degree Kelvin
B = bandwidth
Q = quality factor $\left(Q = \frac{2\pi f_r M}{b}\right)$

It is clear that for monitoring low acceleration levels, a large proof mass M and high Q are required.

The main specifications of an accelerometer are the range, expressed in terms of the earth's gravitation unit $g = 9.81$ m/s^2, sensitivity, resolution (mg), bandwidth (Hz), cross axis sensitivity and immunity to shock. The range and bandwidth vary depending upon the application. The full range of acceleration for airbag crash-sensing accelerometers is ±50g and the bandwidth of about one kilohertz. On the other hand, for measuring engine knock or vibration, the range is only about one g with capability to resolve very small accelerations below 100 μg and a bandwidth greater than 10 kHz. In certain applications such as those which need to be implanted in human body should operate with very low power to save battery life.

In general, all the three approaches, viz, piezoelectric, piezoresitive and capacitive sensing methods have been used for realising the spring–mass system of the accelerometer.

In the piezoelectric approach, either the spring itself is piezoelectric or it contains a piezoelectric thin film to provide a voltage directly proportional to the displacement. In the second approach, piezoresistors are used to sense the inertial stress induced in the spring during the displacement of the mass. Piezoresistive sensing approach is very commonly used because of its simplicity. Capacitive sensing is yet another commonly used method, where the mass forms one side of a parallel plate capacitor. In this approach, special electronics circuit is required to detect very small changes in capacitance of the order of femto farads and convert them into voltage and supply amplified output voltage. We focus on the piezoresitive and capacitive type acceleration sensors.

5.5.2 Piezoresistive Silicon Micro Accelerometer

A schematic structure of the first silicon piezoresistive micro accelerometer proto type developed in Stanford University in the year 1979 is shown in Figure 5.21. This sensor was a cantilever beam–mass structure fabricated in single crystal silicon by the bulk micromachining process. As shown in the figure, this accelerometer chip consists of a seismic mass of silicon and a narrow, thin silicon beam supported on to a silicon frame. Two piezoresistors are formed by selectively diffusing boron on the chip, one on the beam and the other on the frame and they are connected by metallisation to form a half bridge.

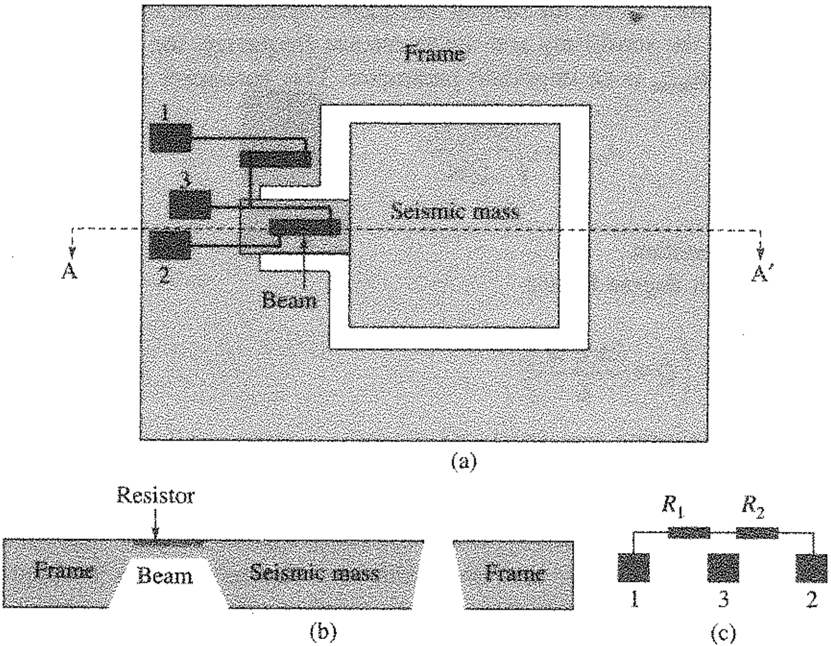


Figure 5.21 Schematic structure of the first silicon piezoresistive accelerometer (a) top view (b) cross-section at A-A′ (c) the half bridge circuit connection (R_1 is on the frame and R_2 is on the beam and its value depends on the strain on the beam due to acceleration).

The basic principle of the piezoresitive accelerometer can be understood referring to this structure. When the device moves up with acceleration normal to the chip plane, the inertial force on the mass forces the beam to bend down. This results in stress in the beam. The resultant strain in the beam is transferred as a longitudinal strain to the resistor on the beam and causes a change in the resistance value of the piezoresitor R_2. The change in output of the half bridge across R_2 is directly proportional to the acceleration.

Structural improvements on the above device have been subsequently implemented to improve the sensitivity and reduce the cross axis sensitivity. The piezoresistive accelerometer was successful in mass production and industrial applications in the 1980s after sorting out several problems experienced, by providing over-range protection with bumpers and damping control with air gap as shown in Figure 5.22. The dimensions of the air gap were chosen such that the critical damping ratio $\xi = 1.0$. The top and the bottom caps are silicon wafers etched by

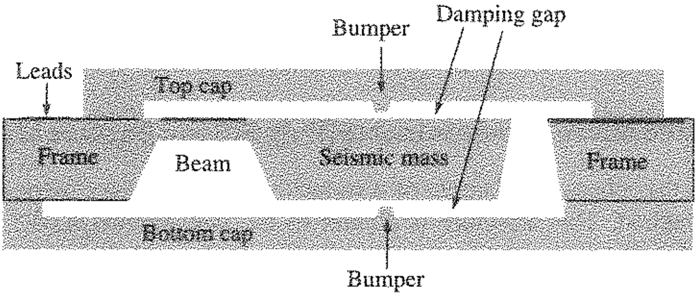


Figure 5.22 Microstructure of an accelerometer with air damping and bumpers to serve as over-range stop mechanism. The frame and seismic mass are fabricated by back machining (etching) the silicon wafer.

bulk micromachining process and then bonded on to the two sides of the middle silicon wafer by fusion bonding technique to provide the air gap. The middle Si wafer contains the seismic mass suspended by the Si beam.

5.5.3 Capacitive Micro Accelerometer (Silicon Bulk Micromachined)

Figure 5.23 shows an example of capacitive accelerometer fabricated by bulk micromachining of silicon. This consists of a spring–mass system realised using a seismic mass supported by cantilever beam anchored at one end to the main body of silicon. This structure is sandwiched between two Pyrex glass plates by anodic bonding technique. The two fixed electrodes are placed on the glass plates as shown. When the system is subjected to acceleration in Z direction, the seismic mass moves in the minus Z direction, resulting in change in the capacitance between the mass and the fixed plates. The capacitance difference ΔC, between the two capacitances can be used as a measure of acceleration. The symmetric design and differential sensing reduce the effect of thermal mismatch, if any, and linearises the ΔC versus capacitance relationship. The differential capacitive sensing also opens up the possibility of using force balance technique.

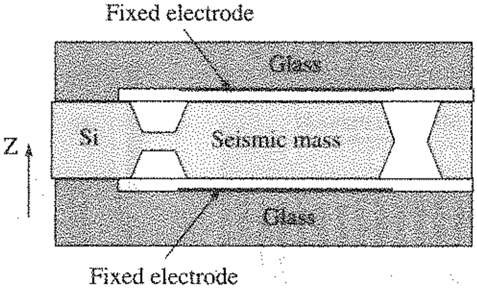


Figure 5.23 Capacitive accelerometer fabricated by bulk micromachining technology.

The gap between the movable electrode and fixed electrode need to be made small in the range 2 μm–5 μm and the mass area is made large to make the sense capacitance large as compared to the parasitic capacitances. A large mass and a small gap result in a large squeeze film air damping force. This is not desirable as it will reduce the quality factor Q and the bandwidth of the accelerometer. The damping can be reduced by providing holes in the seismic mass. In addition, a force balanced sensing measurement scheme will also reduce the damping effect.

5.5.4 Capacitive Micro Accelerometer (Poly Silicon Surface Micromachined)

Figure 5.24 shows the schematic diagram of a capacitive accelerometer fabricated using surface micromachining method. This structure is similar to the well known ADXL50 accelerometer developed by Analog Devices and Siemens. In this approach as discussed in Section 5.3.2, the structural material is doped polycrystalline silicon of about 2 μm thickness deposited on the SiO_2, and patterned to realise the mass plate having several fingers on both sides as shown. This structure is released by etching the sacrificial oxide layer from its underneath except in the regions below the four anchor regions so that the mass is suspended over the substrate by four thin beam flexures. This ensures that mass M is electrically isolated from the substrate to which it is anchored and that it is free to move along its central line as indicated by the arrow marked on the movable mass when subjected to an inertial force in the X direction. The doped polysilicon fingers marked A and B are not released from the substrate and are electrically isolated from the substrate because the oxide underneath them is not etched. They serve as fixed fingers on both sides of any finger of the seismic mass plate at distances of the order of 1 μm to 1.5 μm. The small holes etched in the mass plate are designed to assist in etching the sacrificial layer.

The fingers marked A connected together into one group form capacitance C_1 with respect to the fingers of the movable mass M. Similarly, fingers marked B connected together constitute capacitor C_2 [Figure 5.24(a)]. The equivalent lumped model of the capacitances is shown in Figure 5.24(b).

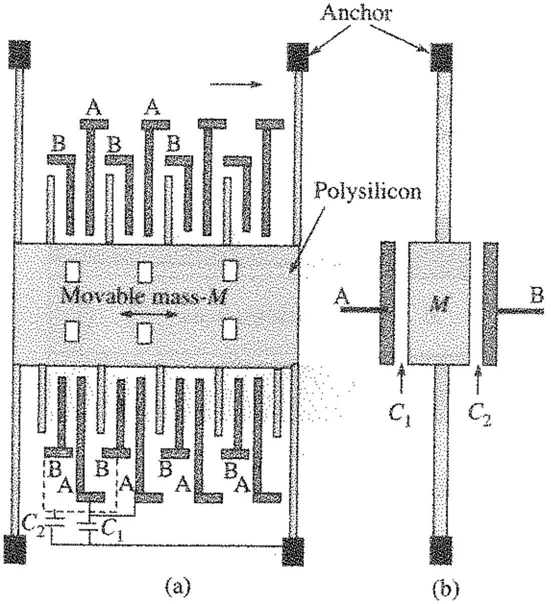


Figure 5.24 Capacitive accelerometer using surface micromachining showing the schematic (a) top view (b) equivalent lumped model.

In the quiescent condition, in the absence of external force or acceleration, $C_1 = C_2$, and hence the output signal proportional to the difference in capacitance is zero. If the suspended structure experiences an inertial force, when the displacement of the mass is to the left side in Figure 5.24, the capacitance C_1 between the fixed electrodes A and the movable electrodes attached to the mass increases, and the capacitance C_2 between the fixed electrodes B and the movable electrodes attached to the mass decreases. These capacitors are very small because the polysilicon layer is usually only about 2-µm thick. The overall capacitance in the ADXL05 rated at ±5*g* is only of the order of 0.1 pF even with several tens of fingers, and the change in capacitance for a 1*g* is about 0.1 fF (= 100 aF). As this is very minute change in capacitance, it cannot be measured directly. Therefore, an on-chip integrated electronics is essential to ensure that the effect of parasitic capacitances is minimised. We describe below the force balanced technique which has been adapted in the ADXL05 by Analog Devices for on-chip electronics.

5.5.5 Force Balanced Accelerometer

The force balancing for measuring the acceleration is achieved using a closed-loop feedback circuit to electromechanically force the movable mass to its equilibrium position when it is subjected to acceleration. The schematic diagram is shown in Figure 5.25.

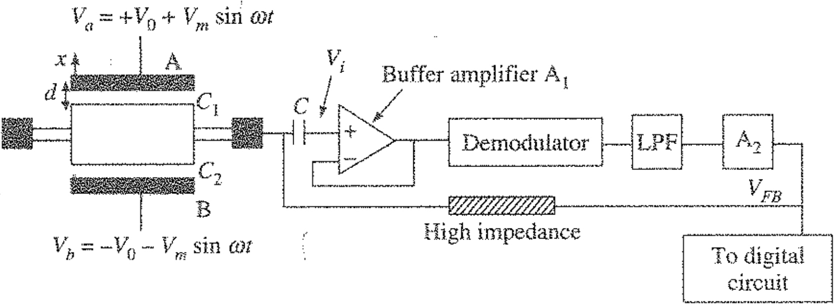


Figure 5.25 Accelerometer with schematic of the force balanced electronics.

We show that the feedback voltage in this force balanced circuit is a measure of the inertial acceleration. In this arrangement, the voltages V_a and V_b applied to electrodes A and B respectively, consist of DC and AC components of equal magnitude, but of 180° out of phase. The capacitor *C* blocks the DC component and allows only the AC signal V_i. The signal output V_i is taken from movable mass and fed to an unity gain buffer stage A_1 followed by a synchronous demodulator, Low Pass Filter (LPF) and an amplifier A_2. When the acceleration is zero, the movable plate is in a balanced position at equal distance *d* between the two fixed plates A and B. The signal voltage V_i for this situation is zero and hence the feedback voltage V_{FB} is zero.

When a downward acceleration *a* acts on the frame supporting the accelerometer, the mass moves towards the upper fixed plate A by a small distance *x*, due to an upward inertial force acting on it, the signal voltage V_i will no longer be zero and noting that $x << d$ due to the presence of feedback condition, this voltage V_i can be shown to be equal to:

$$V_i = \frac{C_1 - C_2}{C_1 + C_2} V_m \sin \omega t = \frac{x}{d} V_m \sin \omega t \tag{5.18}$$

If the combined open loop gain from the buffer, demodulator and operational amplifier is A_0 and noting that the demodulator is a peak detector, the feedback DC voltage through a high impedance path is given by

$$V_{FB} = +A_0 V_m \frac{x}{d} \tag{5.19}$$

This positive DC voltage appears at the central plate, and therefore, the effective DC voltage between the fixed top electrode A and the moving plate reduces from V_0 to $\left(V_0 - A_0 V_m \frac{x}{d}\right)$. At the same time the DC voltage between the fixed electrode B and the movable middle plate increases from $-V_0$ to $-\left(V_0 + A_0 V_m \frac{x}{d}\right)$. As a result, the net direction of electrostatic force on the movable plate caused by the feedback voltage V_{FB} is towards electrode B, thus bringing the mass towards the initial equilibrium position. The displacement of the central plate thus decreases by the electromechanical feedback, and this decrease is significant if the open loop gain A_0 is large enough. In such a situation, the displacement is negligibly small compared to the initial distance d. As $x << d$, the feedback voltage has forced the movable plate back to its neutral position, the feedback voltage can be used as a measure of acceleration a. Hence this is referred to as the force balance technique.

Assuming that $V_m = \alpha V_0$ and using the conditions that $\frac{x}{d} << 1$, it can be easily shown that the electrostatic force F_e applied on the movable plate having an area A can be expressed as:

$$F_e = -\frac{2A\varepsilon_0 V_0^2 \alpha A_0 x}{d^3} \tag{5.20}$$

The net force acting on the movable plate in steady state condition is given by the equation:

$$Ma - kx - \frac{2A\varepsilon_0 V_0^2 \alpha A_0 x}{d^3} = 0 \tag{5.21}$$

Solving for x in Eq. (5.21) and assuming that A_0 is very large, and that 'kx' is small we obtain,

$$x = \frac{Mad^3}{2A\varepsilon_0 V_0^2 \alpha A_0} \tag{5.22}$$

Using Eq. (5.22) in Eq. (5.19), and noting that $V_m = \alpha V_0$, we show that the feedback voltage V_{FB} is related to the acceleration a by the expression:

$$V_{FB} = \frac{d^2}{2A\varepsilon_0 V_0}(Ma) \tag{5.23}$$

Thus by monitoring the feedback voltage V_{FB}, the inertial force (= Ma) and the acceleration a can be estimated and the accelerometer can be calibrated using this V_{FB}.

EXAMPLE 5.3 A capacitive sensing accelerometer is shown in Figure E5.6.

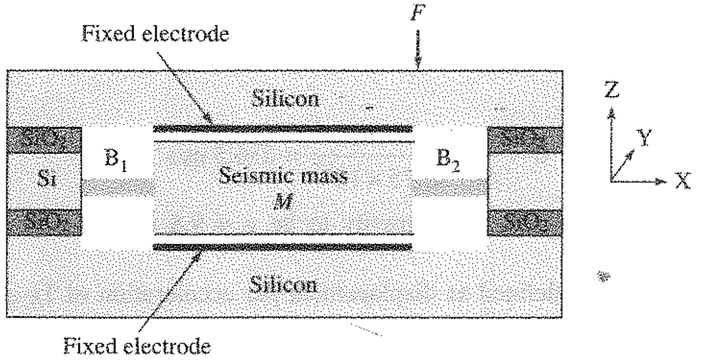


Figure E5.6 Cross-section of a capacitive sensing accelerometer.

The dimensions of the silicon seismic mass are as follows: Thickness t_m = 250 μm and the top view is a square of 1000 μm × 1000 μm. This mass is suspended by two silicon beams of thickness t_b = 15 μm, length L = 300 μm and breadth b = 16 μm and anchored on to the frame as shown, with a layer of SiO_2 for electrical isolation. The density of silicon 2300 kg/m^3 and the Young's modulus E of silicon = 170 GPa. The gap d between the fixed electrode and the seismic mass is 2 μm. The mass is heavily doped silicon so that the conductivity of the mass is very high, and the frame is lightly doped so that its conductivity is low. The frame is subjected to a force F in the –Z direction as shown in the figure. As a result, the mass gets deflected by a vertical distance of 0.2 μm. Determine (a) the steady state capacitances C_1 and C_2 of the mass with respect to the upper electrode and lower fixed electrode respectively. (b) the steady state force F and the acceleration experienced by the mass (c) the sensitivity and the resonance frequency of the accelerometer.

Solution:

(a) When the frame is accelerated in the –Z direction, the mass will get deflected towards the +Z direction by an extent Δd. In this example, Δd = 0.2 μm. As a result, the capacitance C_{01} increases above its equilibrium value C_{01} to a value C_1 and C_{02} decreases to C_2.

$$C_1 = \frac{\varepsilon_0 A}{(d - \Delta d)} = \frac{(8.854 \times 10^{-12}) \times 10^{-6}}{1.8 \times 10^{-6}} = 4.9188 \text{ pF}$$

$$C_2 = \frac{\varepsilon_0 A}{(d + \Delta d)} = \frac{(8.854 \times 10^{-12}) \times 10^{-6}}{2.2 \times 10^{-6}} = 4.0245 \text{ pF}$$

(b) In steady state conditions, $F = Ma = k_{\text{eff}}\,(\Delta d)$. In the example,

$$M = \text{Volume} \times \text{density} = (10^{-6} \times 250 \times 10^{-6}) \times 2300 = 5.75 \times 10^{-7}\,\text{kg}$$

Spring constant k per beam,

$$k_b = \frac{E b\, t_b^3}{L^3} = \frac{1.7 \times 10^{11} \times 16 \times 10^{-6} \times (15)^3}{(300)^3} = 340 \text{ N/m}$$

As the mass is supported by two beams, the effective spring constant is $k_{eff} = 2k_b = 680$ N/m.

$$F = k_{eff}\Delta d = 680 \times 2 \times 10^{-7} = 1.36 \times 10^{-4}\text{ N}$$

The acceleration is:

$$a = \frac{F}{M} = \frac{1.36 \times 10^{-4}}{5.75 \times 10^{-7}} = 236.5\text{ m/s}^2$$

Taking the acceleration due to gravity to be $g = 9.8$ m/s^2 the above acceleration is $24.13g$.

(c) The sensitivity S is defined as displacement per acceleration of $1g$ and is equal to

$$S = \frac{2 \times 10^{-7}}{24.13} = 8.3\text{ nm/g}$$

The resonance frequency of the accelerometer $f_0 = \dfrac{\omega_0}{2\pi} = \dfrac{1}{2\pi}\sqrt{\dfrac{k_{\text{eff}}}{M}} = 5.47$ kHz

EXAMPLE 5.4 In Example 5.3, a force feedback circuit, shown in Figure 5.30 is used with $V_0 = 20$ V and $V_m = 1$ V. Determine the feedback voltage V_{FB} necessary to achieve the force balance condition.

Solution: Electrostatic force F_{es} on the mass when the feedback voltage V_{FB} is applied to the mass is given by the relation (taking A = area of the electrodes):

$$F_{es} = \frac{\varepsilon_0 A}{2}\left[\frac{(V_0 - V_{FB})^2}{(d-\Delta d)^2} - \frac{(V_0 + V_{FB})^2}{(d+\Delta d)^2}\right]$$

$$= \frac{\varepsilon_0 A V_0^2}{2d^2}\left[\frac{\left(1 - \dfrac{V_{FB}}{V_0}\right)^2}{\left(1 - \dfrac{\Delta d}{d}\right)^2} - \frac{\left(1 + \dfrac{V_{FB}}{V_0}\right)^2}{\left(1 + \dfrac{\Delta d}{d}\right)^2}\right]$$

Noting that $\dfrac{\Delta d}{d} << 1$, $A_0 = \dfrac{V_{FB}}{V_i}$, $V_{FB} = V_m A_0 \dfrac{\Delta d}{d}$, the above equation can be simplified and expressed as follows:

$$F_{es} = -\frac{2A\varepsilon_0 V_0}{d^2} V_m A_0 \frac{\Delta d}{d}$$

When the mass is forced back to the equilibrium position, $F = Ma = F_{es}$ in magnitude. We obtain:

$$Ma = \frac{2A\varepsilon_0 V_0}{d^2} V_m A_0 \frac{\Delta d}{d}$$

This gives us:

$$V_m A_0 = \frac{d^2(Ma)}{2A\varepsilon_0 V_0(\Delta d/d)} = \frac{4\times 10^{-12}\times 1.36\times 10^{-4}}{2\times 10^{-6}\times 8.854\times 10^{-12}\times 20\times(0.2/2)} = 15.36$$

$$V_m = 1\text{ V},\qquad \text{Therefore } A_0 = 15.36$$

$$V_{FB} = V_m A_0 \frac{z}{d} = 15.36\times 0.1 = 1.536\text{ V}$$

5.5.6 Damping Factor in Micromachined Inertial Grade Accelerometers

As discussed in Chapter 4 and also depicted by Eq. (5.13), the dynamic performance of accelerometers is governed not only by the mass and the spring constant, it also depends mostly on the damping coefficient or the damping factor denoted by the symbol b. The quality factor and the noise equivalent acceleration, defined in Eq. (5.17), as well as the step response of the accelerometer depends upon the damping factor. A more convenient term related to the damping factor is the damping ratio ξ which is related to the damping factor b in the following equation:

$$\xi = \frac{1}{2Q} = \frac{b}{2\omega_r M} \tag{5.24}$$

The coefficient of damping b is proportional to the dimensions of the mechanical structure and the coefficient of viscosity of the surrounding fluid. Hence in the accelerometers made of conventional mechanical structures, which are invariably quite large, the ξ is very small in air. However, in practice, this problem has been circumvented and damping ratio $\xi = 0.7$ has been achieved by immersing the structure with silicon oil which has high viscosity. Contrary to this, accelerometers formed by micromechanical structures by micromachining technology, the damping ratio can be easily raised to around 0.7 in air by using some mechanical structures to increase the damping force in a controlled way. Using air damping is advantageous compared to oil damping because of the lower temperature coefficient of air and the ease of packaging the device in the absence of the oil based damping. Squeeze film damping and slide film damping are the mechanisms of air damping in micromechanical structures.

In the case of devices such as resonant sensors and gyroscopes which require high Q-factor, air damping should be reduced by evacuating the air from a hermetically sealed package containing the device.

In this section, we provide a bird's eye view of the squeeze film air damping and the damping in rare air as well as in vacuum.

Damping factor and damping force for a rectangular plate in air

Squeeze film damping factor in the normal direction for a rectangular plate of length L and width B with an air gap h, in general, for B/L ratio varying from 0 to 1 is shown to be:

$$b \approx \frac{\mu L B^3}{h^3}\times f\left(\frac{B}{L}\right) \tag{5.25}$$

Here,
$$f\left(\frac{B}{L}\right)=1-\frac{192}{\pi^5}\times\frac{B}{L}\times\sum_{n=1,3,5}^{\infty}\left(\frac{1}{n^5}\tanh\frac{n\pi L}{2B}\right) \tag{5.26}$$

μ = coefficient of viscosity = 1.8×10^{-5} Pa-s for air at 20°C. μ for water is 1.0×10^{-3} Pa-s.

From Eq. (5.26), we have

$$f\left(\frac{B}{L}\right)=0.42 \qquad \text{when } B = L \text{ (square plate)}$$

and
$$f\left(\frac{B}{L}\right)\approx 0.7 \qquad \text{when } B/L = 0.5$$

Thus for a square plate of side length, $B = L$, the air damping force F_n for movement in the direction (z) perpendicular to the plane of the plate is thus given by

$$F_n = b\frac{dz}{dt} = 0.42\frac{\mu L B^3}{h^3}\frac{dz}{dt} \tag{5.27}$$

Thus, for a square plate of 1 mm × 1 mm and at h = 2 microns away from the neighbouring substrate, the damping factor from Eq. (5.25) is b = 0.945 Pa-s-m.

For the same plate, if h = 20 microns, it turns out that $b = 9.45 \times 10^{-4}$ Pa-s-m

Damping force F_s and damping factor b for a comb type accelerometer

Air damping force on comb type accelerometers such as the one discussed in Section 5.5.4, consists of several components such as the slide film damping forces damping on the bottom, top and side walls respectively. The total side film damping force F_s is the sum of all these components and can be shown to be given by the relation:

$$F_s = \mu\left[\frac{A_p}{d_p}+\frac{A_p}{\delta}+\frac{A_s}{d_s}\right]\frac{dx}{dt} = b\frac{dx}{dt}$$

Here

A_p = area of the plate

d_p = distance between the plate and the moving part and the substrate

μ = coefficient of viscosity

A_s = area of the side walls that are parallel to the moving direction

d_s = distance between the side walls and their neighbouring structures

δ = effective distance which is defined as:

$$\delta = \sqrt{\frac{2\mu}{\rho\omega}} \tag{5.28}$$

where ρ is the density of air and $\omega = 2\pi f$. δ decreases form about 70 μm at resonance frequency f = 1 kHz and saturates at 20 μm for frequencies $f \gtrsim 10$ kHz.

Effect of air pressure on the damping force

The damping force is independent of air pressure over a wide range of pressures around and above atmospheric pressure because the viscosity of air is practically constant in this pressure range. However, it has been observed that the air damping force on microstructures decreases significantly when the air pressure is below several hundreds of Pascal. From a detailed analysis based on the molecular model, it has been shown that the squeeze film damping force F_{rn} due to rarified air at pressure P, acting on a plate of surface area A_p moving in the normal direction, can be expressed as follows:

$$F_{rn} = 4\sqrt{\frac{2M}{\pi RT}}PA_p\frac{dz}{dt} = b\frac{dz}{dt} \tag{5.29}$$

where the universal molar gas constant R = 8.31 kgm^3/s^2/°K, M is the molar mass of the gas and T is the temperature in °K.

Equating the F_{rn} and F_n in Eqs. (5.29) and (5.27), the transition pressure P_r can be determined.

From the above discussion, it is evident that the damping factor decreases proportional to the pressure and hence the quality factor Q increases at lower pressures. However, the Q-factor levels of 10^4–10^5 at very high vacuum levels P_0 (as shown in Figure 5.26) when the energy dissipation is governed by other factors such as the internal friction and support losses. The pressure P_0 at which this levelling takes place is in the range of several Pascal.

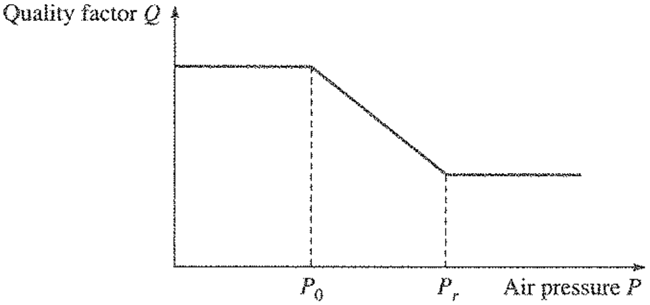


Figure 5.26 Quality factor versus air pressure for microstructures.

5.6 MEMS Gyro

The gyroscope is an inertial sensor, which measures the angular motion about one or more axes. Invariably, It is part of a larger control system, and it gives information on correcting the changes in the angular motion so that the system is stabilized. For example, in an automobile, a gyroscope detects the angular motion of the car, and gives input to the safety system to actively control the steering angle and the brakes at each wheel to prevent the vehicle from overturning, if the angular motion is above a critical level. Miniaturised MEMS based gyroscopes have now become all-pervasive and find applications in automotive, platform stabilisation for video camera, stability and active control system and robotics. This section gives a brief concept of gyroscope along with an example of a 'Tuning Fork MEMS Gyro'.

5.6.1 Operating Principle

Guidance, navigation and control systems in aircrafts and spacecrafts require gyros to maintain orientation in flight. Gyro based systems provide the most direct method for sensing inertial angular velocity. The requirement on the dynamic operating range of gyro depends upon the particular application. It is typically around 10^4 in inertial systems for land transport systems. In the case of precision inertial navigation systems such as military airplanes, launch vehicles, missiles and satellites, dynamic operating range value touches as high as 10^8. Such a large dynamic range typically covers maximum operating range of 100°/s (3.6×10^{5}°/h) to minimum detectable threshold of 0.01°/h. Currently, macro sized gyros, discussed in Chapter 3, such as hemispherical resonator gyro, dynamically tuned gyro, ring laser gyro and fiber-optic gyro, meet the high dynamic operating range. Whereas, MEMS gyros are primarily fulfilling the requirements on the lower dynamic range, but showing further improvement over the years as sustained research and development continues. Depending upon the technology and/or the physics used, micromachined gyros can be categorised under MEMS gyros and the MOEMS (Micro Optical Electro Mechanical Systems) gyros. In this section we focus on the MEMS gyros.

Currently, MEMS gyros use a vibrating structure rather than the traditional rotating disc to determine inertial orientation. These gyros measure angular rate by detecting the Coriolis acceleration or force taking advantage of Coriolis effect, which can be described as follows.

Consider a rotating disc where an object near the centre of the disc moves slowly, whereas an object near the outer edge moves much faster. Both the objects are also subject to tangential acceleration due to rotation. The acceleration which is proportional to the objects distance from the centre of the disc is known as the *Coriolis acceleration*. Thus the object near the outer edge experiences a greater Coriolis acceleration than the object near the centre of the disc. Vibratory gyros make use of the fact that if a mechanical member is vibrating along one reference axis, rotation of gyro frame will couple some of the vibrational energy into another axis due to the Coriolis effect. Mathematically, the Coriolis force F_c, exerted on a mass M moving (vibrating) with a velocity v, and subjected to a rotation (by mounting the mass on a rotating gyro frame) at angular velocity Ω, is expressed by the vector cross product relation as:

$$\mathbf{F}_c = 2M(\mathbf{v} \times \mathbf{\Omega}) \tag{5.30}$$

This force is thus at right angles to both the velocity of the mass and the angular velocity of the frame. Figure 5.27 shows a simplified model for a vibratory gyro. In this figure, velocity v is along +ve X-axis (on the plane of the paper), angular rate Ω along +ve Z-axis (vertical up and perpendicular to the plane of the paper). Then the Coriolis force F_c is along +ve Y-axis, and it will be on the plane of the paper.

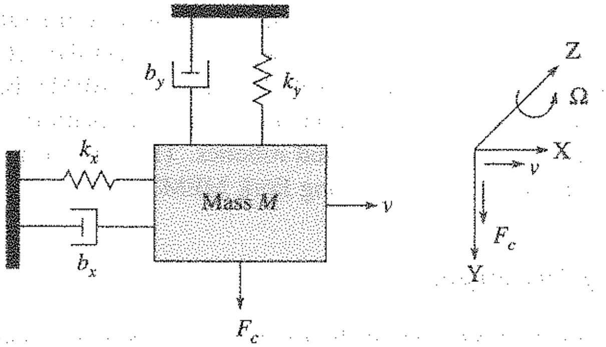


Figure 5.27 A simple model for vibratory gyroscope.

This vibratory system has two orthogonal vibration modes which are vibration of mass M, in the driving mode (X direction), and the second is the sensing mode (Y direction) with the corresponding resonance frequencies as ω_x and ω_y.

When the mass is driven into vibration in the X direction shown in the figure with a frequency, $\omega_d \approx \omega_y$, if the frame on to which the mass is anchored through the flexures rotates with an angular rate of Ω around the Z-axis (normal to the plane of paper) as shown, an alternating Coriolis force appears in the Y direction driving the system into vibration in the Y direction. If the vibration in the X direction can be represented as $x = A_{dm}$ sin ωdt, the Coriolis force F_{cy} is in the Y direction is obtained from Eq. (5.30). Thus, we have:

$$F_{cy} = 2M\frac{dx}{dt}\Omega = (2MA_{dm}\omega_d\Omega)\cos\omega_d t \tag{5.31}$$

The differential equation for the mass movement in the Y direction is given by

$$M\frac{d^2y}{dt^2} + b_y\frac{dy}{dt} + k_y y = F_{cy} \tag{5.32}$$

where

b_y = damping term

k_y = flexure spring stiffness

Using Eq. (5.31) in Eq. (5.32) and solving it can be shown that, when $\omega_d = \omega_y$, the amplitude A_{ym} of vibration in the Y direction is given by

$$A_{ym} = \frac{2A_{dm}\,Q_y\Omega}{\omega_y} \tag{5.33}$$

Thus it is seen that the amplitude A_{ym} is proportional to the angular rate Ω. Hence Ω can be measured by determining the displacement amplitude A_{ym}. This can be done by any one of the techniques, namely, capacitive or piezoresistive or piezoelectric sensing method. Capacitive sensing method is often used because high sensitivity and temperature insensitivity can be achieved with this approach. It is to be noted from Eq. (5.33) that the quality factor Q_y in the sensing mode and the driving amplitude A_{dm} should be as large as possible, while the vibration frequency ω_y should be small to achieve high sensitivity. This is particularly important because $\Omega << \omega_y$.

The driving of the micro gyro can be done by electrostatic activation, piezoelectric or electro-thermal means. The sensing can be done using capacitive, piezoresistive or piezoelectric methods. Even though many designs for MEMS vibratory gyro have been attempted, the most popular design is the tuning-fork gyro design. Therefore, the working of this type of MEMS gyro is presented in the following Section 5.6.2.

5.6.2 Tuning Fork MEMS Gyro

The structure of an operating tuning fork gyro, described further, is not like a conventional tuning fork. It consists of a pair of mass plates marked as the moving plates A and B which are connected using wide and stiff cross bar as shown by a schematic diagram in Figure 5.28. These two mass plates can be excited with electrostatic forces by applying a voltage on the fixed electrode marked C and oscillate in phase opposition in the X direction. The natural vibration frequency for the X and Y directions are ω_x and ω_y respectively. The operation of this type of gyro is described further.

The two mass plates—A and B are driven into vibration by electrostatic actuation of the comb drives in an out-of-phase pattern as shown by the arrows on the two mass plates in Figure 5.28. In this diagram, the +ve velocity is along X, angular rate is about Z-axis which is perpendicular to X, but both these axes are lying on the plane of the paper. When the device is rotated about the Z-axis, the Coriolis forces act on the two mass plates—A and B along Y-axis, where Y-axis will be perpendicular to the plane of the paper. Since the plates are vibrated in out-of-phase mode, the Coriolis force on plate A will also be out-of-phase with respect to the Coriolis force on plate B. For the velocity directions of the plates, marked by arrows (solid lines) in the figure, when the Coriolis force causes plate A to deflect into the plane of paper, the Coriolis force causes plate B to bulge out from the plane of paper. With electrodes provided underneath of each plate, the plate deflections can be detected by the differential capacitance of the two plates using a custom CMOS ASIC.

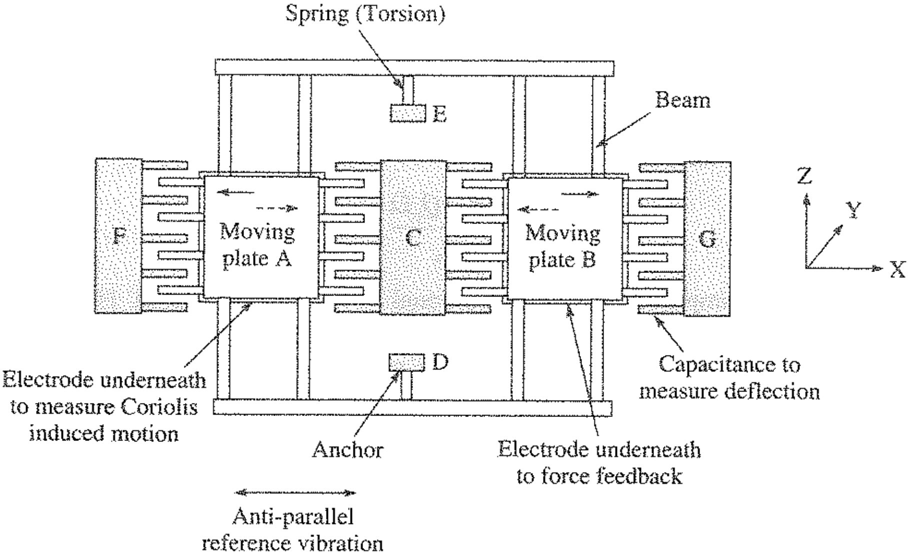

Spring (Torsion)
E
Beam
F
Moving plate A
C
Moving plate B
G
Z
Y
X
D
Electrode underneath to measure Coriolis induced motion
Capacitance to measure deflection
Anchor
Electrode underneath to force feedback
Anti-parallel reference vibration

Figure 5.28 Schematic view of a tuning fork gyro.

The tuning fork is fixed to the substrate at the anchor points D and E. The outer electrodes F and G are used to control amplitude and phase of the mechanical reference vibration. The gyro will be exited at resonance frequency to achieve highest amplitudes. Capacitor plates are located underneath each of the electrodes A and B for measuring the Coriolis induced movement, i.e. the gyro signal, for force feedback usage and to apply a static voltage to trim the torsional resonance frequency.

As the differential capacitance of two comb drive resonators is used as a measure of the angular rate, the gyro will have much higher immunity to the interference from environment because the interference are usually common mode signals. As the two masses are oscillating out of phase, the dual resonator structure is often referred to as a tuning fork structure.

For a practical gyro device, the design, fabrication, packaging and signal conditioning circuitry are quite involved. The design is related to the driving schemes, damping control and sensing details. The fabrication is rather involved as the requirement for the resonant frequencies of the structure is rigid, and quite often, the structure will have to be encapsulated in vacuum to achieve high Q. The signal detection is difficult because of the extremely weak signal and the phase differences among the electrical driving signals, driving vibration and the detection vibration caused by mechanical reasons and the air damping effect.

One of the most important requirements in a MEMS gyro is to control the drive frequency and drive amplitude. This is normally achieved using Phase Locked Loop (PLL) for such function.

SUMMARY

The benefits of miniaturisation of mechanical components such as the inertial sensors and actuators were first presented. The implications of using silicon for this purpose was then described pointing out that the batch processing and miniaturisation are possible when silicon is used for this purpose. The micromachining processes such as the bulk micromachining, surface micromachining for mechanical sensors and actuators are detailed; the LIGA and SU-8 processes suitable for high aspect ratio devices were presented. Both wet chemical etching and dry etching processes have been described in detail. The applications of these micromachining processes for realising inertial sensors such as accelerometers and gyros are discussed at length with examples of the device structures.

EXERCISES

5.1 In the worked Example 5.1, if the etch duration is only 3 hours, draw the cross-section and top view showing the dimensions. [**Ans:** $H = 150$ μm; $W = 787.9$ μm]

5.2 In the worked Example 5.1, draw the cross-section and top view if the oxide window size were 2 mm × 1 mm.

[**Ans:** $H = 300$ μm; Top surface: 2 mm × 1 mm; Bottom surface: 1576 μm × 576 μm]

5.3 A (100) silicon wafer is 210-μm thick. A square window of unknown size is created by photolithography. The sides of the window are parallel to <110>. After through wafer etch, a hole size of 100 μm × 100 μm is formed on the other side of the wafer. Determine the size of the oxide window, assuming that the etch rate along the <111> direction is negligible. [**Ans:** 397 × 397 μm]

5.4 In the worked Example 5.2, the longitudinal strain on R_1 is 2ε and on R_3 it is ε with $\varepsilon = 10^{-3}$, determine the output voltage V_0 of the bridge. [**Ans:** $V_0 = 0.22$ V]

5.5 Work out the details same as for the worked Example 5.3 keeping all the details same except that the mass is supported by four flanges B_1, B_2, B_3 and B_4 as shown by the top view in Figure E5.1.

[**Ans:** (a) C_1 and C_2 are same as in Example 5.3 (b) Acceleration = 473 m/s^2 = 48.26 g (c) Sensitivity $= \frac{M}{k_{\text{eff}}} = 4.15\ \text{nm/g}$ (d) $f_r = 547\sqrt{2}$ kHz]

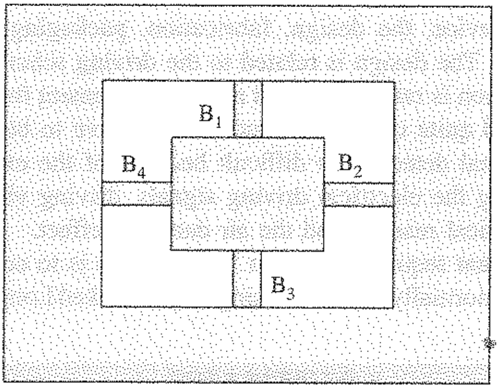


Figure E5.1

REFERENCES

Petersen, K.E., Silicon as a Mechanical Material, *Proc. IEEE*, **70**, pp. 420–457, 1982.

Sullivan, J.P., et al., Diamond and Amorphous Carbon, *MEMS*, **26**, pp. 309–311, *MRS Bulletin*, Special Issue on Micromechanical Systems: Technology and Applications, April 2001.

6

Satellite Navigation

Satellite Navigation provides extremely accurate three-dimensional position and velocity information to users anywhere in the world. Currently, there are two proven systems: the first one is called Global Positioning System (GPS), while the second one is called Global Orbiting Navigation Satellite System (GLONASS). The evolution and the basics of satellite navigation have been introduced in Chapter 1. It provides accurate position, velocity and time information. Additionally, it can provide attitude information also. Hence, satellite navigation is an independent system and capable of providing all the useful data to any user across the globe, covering land, air and space. However, it is not autonomous in the sense an INS is described to be autonomous, as major ground and space segments are involved in its operation.

GPS and GLONASS are military funded programmes of USA and Russia respectively. The basic operational aspects of the two systems have considerable similarity between them and because of this similarity, currently receivers can be realised so that all the satellites belonging to these two systems can be tracked and navigation data can be computed which increases the robustness of the Satellite Navigation System (SNS). However, the system is very costly to initially operationalise and later to maintain it over the years. This happens as these orbiting satellites have limited life time requiring replacement at regular intervals. Fortunately, electronics technology has made the receiver to become progressively cheaper and has enabled growing commercial usage across the globe. Its growing application has triggered other space powers, like India, China and Japan, to configure their own satellite navigation system which will, instead of global coverage will cater for regional cover. As considerable publications have already been made on GPS, its description in the following sections will effectively serve in the understanding of SNS, in general.

6.1 Global Positioning System

Global Positioning System (GPS) consists of three major segments (refer Figure 6.1) known as the space segment, the control segment and the user segment, which is also known as the receiver. These three segments are further described in the sections to follow.

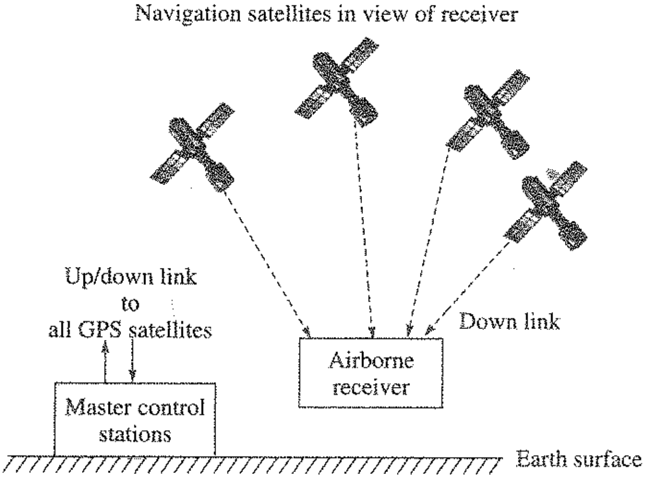


Figure 6.1 Architecture of satellite navigation system.

6.1.1 Space Segment

Presently, the GPS *space segment* comprises a constellation of 28 operational Navstar satellites with four as spares. These satellites orbit the earth with a period of one-half a sidereal day, which is 11 hour 58 minutes to be precise, in nearly circular orbit of radius approximately 26560 km from the centre of the earth. This is at an altitude of approximately 20,200 km. There are six orbital planes with four satellites in each plus four orbital spares. A view of the constellation is shown in Figure 6.2.

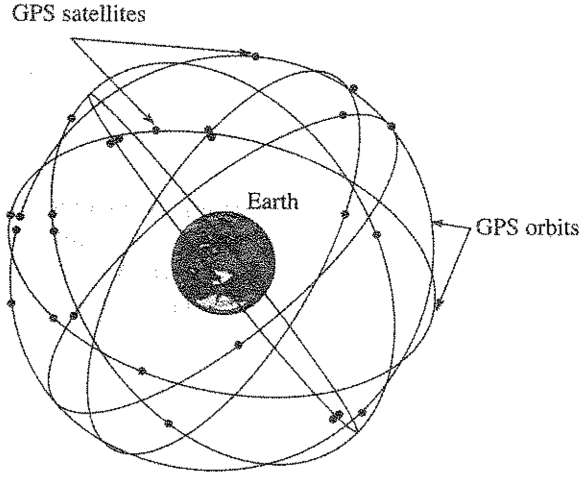


Figure 6.2 GPS constellation of orbiting satellites.

The full constellation ensures global coverage with 6 to 11 simultaneously observable satellites to users located anywhere in the world at any time of day, thus ensuring considerable redundancy over the minimum four satellites needed for position computation. Each of these satellites continuously transmits coded L-band radio signals that the receiver will decode to determine important satellite parameters. The receiver tracks the RF signals of selected satellites and calculates three-dimensional navigation data and time. The satellites have various systems of identification such as (i) Launch sequence number (ii) Assigned vehicle Pseudo Random Noise (PRN) code (iii) Orbital position number (iv) Catalogue number (v) International designation and so on. Each satellite carries a highly accurate Cesium or a Rubidium atomic clock to provide timing for the satellite signals. Internal clock correction is provided for each satellite clock. All signal components are precisely controlled by the atomic clock. The satellite employs transmitting antenna whose shaped-beam gain radiates near uniform power to system users.

6.1.2 Control Segment

The *control segment* comprises a Master Control Station (MCS), an alternate master control station, six worldwide monitor stations and four dedicated ground antenna stations. The ground monitoring stations measure signals from the overhead satellites at fixed interval of time and corrected data is transmitted to master control station. MCS determines the orbital model, clock performance and the health of the satellites, and these are then relayed to the uplink ground antennas for transmission to the satellites, which is further broadcast to the user segment. Main operations and tasks of MCS can be listed as:

(a) Tracking of satellites
(b) Orbit determination
(c) Prediction, modelling and time synchronisation of satellites
(d) Upload of the data for broadcast to the user segment
(e) Monitoring the health status of the satellites.

6.1.3 User Segment

The user segment, normally called a receiver, consists of an antenna along with the receiver electronics that receives and decodes satellite transmissions. The receiver also converts satellite signals to computed position, velocity and time (P, V, T) estimates. The receiver performs the following primary tasks:

(a) Selection of one or more satellites
(b) To acquire satellite signals, measuring the range to the satellite and tracking more satellites
(c) Processing of measurements in real time to compute navigational data in a navigation frame that is needed by the user application.

The receiver maintains a time reference used to generate a replica of the code transmitted by the satellite. The amount of time the receiver must apply to correlate the replica with the satellite clock referenced code received from the satellite, provides a measure of the signal propagation time between the satellite and the receiver. This time when multiplied by the velocity of light provides the range distance.

6.1.4 GPS Signal and Navigation Message from Satellite

Prior to the modernisation of GPS and also as not all the satellites have changed over to new scheme, each of the older satellites transmits simultaneously on two L-band frequency carriers known as L_1 and L_2. The L_1 carrier has an in-phase as well as a quadrature component. The in-phase component is biphase modulated by a 50 bps data stream and a pseudorandom code known as C/A code. The quadrature phase is modulated by the similar 50 bps data stream, but with a different pseudorandom code known as P-code. The L_2 signal is modulated by the 50 bps data stream and P-code. Only, C/A code is available to the civilian user, whereas, P-code is for defence and is not available to the civilians through a process of encryption.

Navigation message from satellite contains the data, which the receiver requires to perform the operations and compilations for successful navigation. The requirements and characteristics of the data are shown in Table 6.1.

Table 6.1 Requirements and Characteristics of SNS Navigation Data

Requirements	Information provided by SNS navigation data
Precise satellite position at the time of transmission	Satellite ephemeris using a modified Kebler model (sinusoidal perturbations) in an Earth Centred Inertial (ECI) co-ordinate frame with transformation to Earth Centred Earth fixed (ECEF) co-ordinates
Precise satellite time at time of transmission	Satellite clock error models and relative correction
Select the best set of satellites for lowest appropriate GDOP within elevation angle constraint (requires approximate knowledge of satellite position)	A moderate accuracy almanac that gives approximate position, time and satellite health for the entire SNS constellation
Time transfer information	SNS time to Universal Co-ordinate Time (UTC) conversion data
Ionospheric correction for single frequency users	Approximate model of ionosphere vs. time and user location
Quality of satellite signal/data	User Range Accuracy (URA) index N is transmitted that gives a quantised measure of space vehicle accuracy available to the civil users

The receiver determines the satellites visible to the antenna and searches for the signal both in frequency and in C/A code phase and confirms detection of both. Thereafter, the receiver locks onto the C/A code as well as locks onto the carrier. After necessary processing, the receiver demodulates the 50 bps navigation data. The code tracking loop is important for the precise measurement of the time of arrival of the received code for the purpose of range estimation. Time of arrival cannot be made directly from the received signal due to the presence of noise. The code tracking loop process aligns the received code with a generated reference code, and this enables the range measurement to be done using the reference code rather than the noisy received signal code.

Modern aerospace receivers are increasingly designed with multi-channel capability, some times called all-in-view. This means that with around 6–8 satellites visible most of the time to a receiver antenna, the multi-channel feature of 12-channel receiver enables one receiver channel to lock onto one satellite and still having spare channels to look for new satellites appearing on the horizon.

6.2 Position and Velocity Determination

The principal observation in SNS is the measurement of time interval between the receiver and the satellite. This time interval t_s multiplied by the velocity of light c can be converted to a distance, which is defined as range. Neglecting the clock and other measurement errors, the true range ρ_r to satellite is then given as:

$$\rho_r = ct_s = c(T_S - t_u) \tag{6.1}$$

where

T_S = time tagged with the satellite position information received by the receiver

t_u = receiver clock time when satellite data is received

Taking the range measurements $\rho_r^{(i)}$ with i satellites (i = 1, 2, 3), the user position location can be expressed as:

$$\rho_r^{(i)} = \sqrt{[(X^{(i)} - x)^2 + (Y^{(i)} - y)^2 + (Z^{(i)} - z)^2]} \tag{6.2}$$

$X^{(i)}, Y^{(i)}, Z^{(i)}$ = ith satellite (i = 1, 2, 3) position co-ordinates in ECEF frame at the time of transmission, known from ephemeris data

x, y, z = receiver position, in ECEF frame, to be determined

Equation (6.2) represents equation for three spheres, and for each, the radius is given by the observed range and the centre at the corresponding satellite. These spheres intersect at the user location to be established that has been shown in Figure 1.13 in Section 1.5. This aspect of position determination is explained further. True range is also known as the *geometric range*.

If the surfaces of two spheres intersect at more than one point, they intersect in a circle. This means that with two observations, the user lies anywhere on the circle. When a third observation is taken and it is known that the user lies anywhere on its spherical surface, then the intersection of this third sphere with the first two results in two user locations where one of them is true. Appropriate algorithm can then find out the true one. When a fourth observation is added, then it can uniquely define the user position. This process of finding the true position is normally called *Trilateration*.

Similarly, the Doppler measurement of range rate $\dot{\rho}_r$ provides means for the determination of receiver velocity as follows:

$$\dot{\rho}_r = \frac{1}{\rho_r}[(X - x)(\dot{X} - \dot{x}) + (Y - y)(\dot{Y} - \dot{y}) + (Z - z)(\dot{Z} - \dot{z})] \tag{6.3}$$

where

$\dot{\rho}_r$ = range rate (known)

$\dot{X}, \dot{Y}, \dot{Z}$ = satellite velocity (known)

$\dot{x}, \dot{y}, \dot{z}$ = receiver velocity (to be determined)

x, y, z = receiver position [known from solution of Eq. (6.2)]

The receiver position will be thus available in the ECEF frame. But, for terrestrial application, it will be necessary to convert the solution to the user defined frame, typically a geographic frame. Figure 6.3 represents this aspect of the navigation.

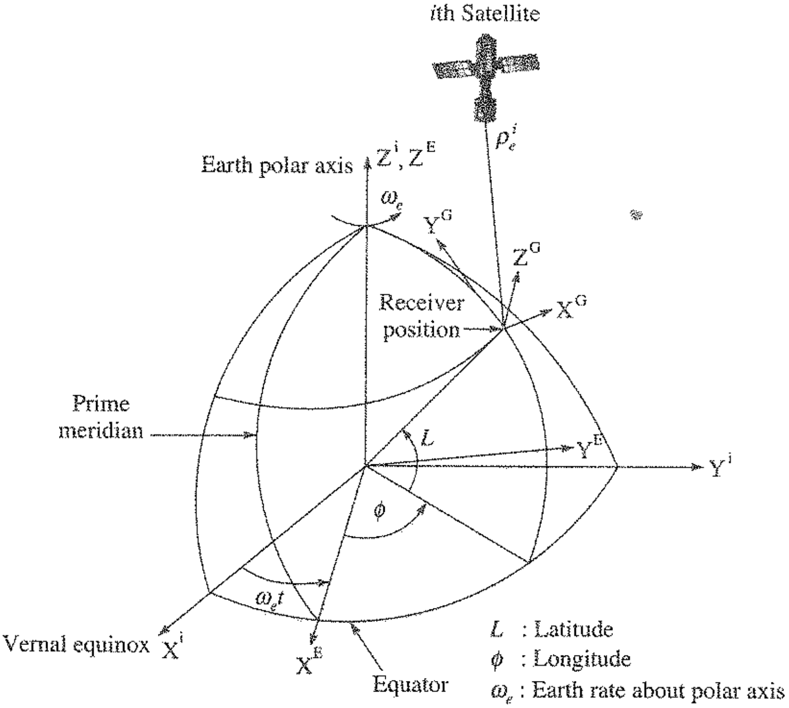


Figure 6.3 Position determination in the user defined geographic frame.

The transformation equation that converts to the user defined *G-frame* can be expressed as:

$$\mathbf{R}^G = C_E^G\ \mathbf{R}^E + \mathbf{R}_0^G \tag{6.4}$$

where

$\mathbf{R}^G$ = position vector transformed to user defined *G-frame*

C_E^G = transformation matrix from the ECEF (*E-frame*) to the *G-frame*

$\mathbf{R}^E$ = solution of position vector in *E-frame*

$\mathbf{R}_0^G$ = transformation necessary to *G-frame* due to shifting of origin between the frames

These transformations, after defining the *G-frame* as ENU, have been worked out, and the readers can refer the book by [**Grewal et al., 2001**].

6.2.1 Range Equations with Errors

In the actual range measurement, several errors crop up and foremost error amongst them is that the receiver clock is not an accurate clock, unlike the highly accurate atomic clock used in satellites. Even 1 μs error leads to an error of 300 m, which can be seen from Eq. (6.1). This error is known as the bias error b_u in the receiver clock. While all the atomic clocks of the satellites are synchronised, the receiver clock is not synchronised with the satellite clock which further results in the receiver clock bias value b_u. The bias value will also vary due to receiver clock drift and needs to be assessed. The clock bias leads to error in the measurement of ranges and these observed range measurements are then known as *pseudo ranges*. This modifies Eqs. (6.1) and (6.2) as follows:

$$\tilde{\rho}^{(i)} = \rho^{i} + cb_u$$

$$\tilde{\rho}^{(i)} = \sqrt{[(X^{(i)} - x)^2 + (Y^{(i)} - y)^2 + (Z^{(i)} - z)^2]} + cb_u \tag{6.5}$$

where

$\tilde{\rho}^{(i)}$ = pseudo range measured to *i*th satellite (i = 1, 2, 3, 4)

$\rho^{(i)}$ = geometric range to the satellite

cb_u = error in range due to receiver clock bias with respect to *i*th satellite clock

To eliminate the large error associated with receiver clock bias, the bias is modelled as an unknown parameter and estimated along with estimates of the three-dimensional position from the pseudo range measurements taken from four satellites. Since the clock bias is estimated at each measurement update, the requirement of a highly stable clock, stable over a longer time, is thereby eliminated. Besides the clock bias, there are other parameters, which corrupt the pseudo range measurements. The total pseudo range error ρ(error), due to these remaining sources, is collectively defined in the following equation:

$$\rho(\text{error}) = cb_s^{(i)} + ct_a^{(i)} + E^{(i)} + MP^{(i)} + \eta^{(i)} \tag{6.6}$$

where

$cb_s^{(i)}$ = range error due to satellite clock error

$ct_a^{(i)}$ = atmospheric delay error

$E^{(i)}$ = error in the broadcast ephemeris data

$MP^{(i)}$ = multi path error

$\eta^{(i)}$ = receiver error tracking noise

i = superscript that denotes the satellite number

These pseudo range error sources are depicted in Figure 6.4.

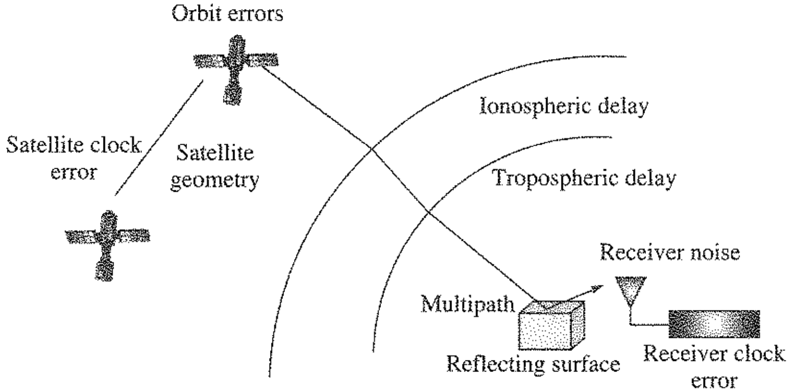


Figure 6.4 Errors in pseudoranging.

The GPS control segment monitors and determines the satellite clock bias with respect to GPS time and upload to the satellite. The receiver receives the correction information from the satellite and uses it to correct the measured pseudo range.

An estimate of these pseudo range errors is brought out in Table 6.2 [**Kintner et al.**]

Table 6.2 Typical Estimated Values of Range Error

Error sources	User equivalent range error (C/A code)
Satellite clock error	2 m
Atmospheric delays	5–15 m
Group delay (satellite equipment)	1 m
Multi-path	3–5 m
Receiver noise + Resolution	1 m

6.2.2 Geometric Dilution of Precision

Geometric Dilution Of Precision (GDOP) is a measure of the goodness of navigation measurement due to the quality of satellite geometry that is tracked by the receiver antenna, and it is roughly interpreted as ratio of position error to the range error precision. The geometry aspect can be expressed mathematically and from there the dilution of precision is also derived.

If, H is the satellite geometry observation matrix, mathematically, GDOP is defined as the square root of the trace of the co-variance matrix $[H^T \cdot H]^{-1}$. Imagine that a square pyramid is formed by lines joining the four satellites and receiver at the point. The larger the volume of this tetrahedron, the better is the value of GDOP; while the smaller its volume, the worse the value of GDOP will be.

The geometry aspect of the satellites under view of the receiver is shown in Figure 6.5 to explain a good GDOP or a poor GDOP. The most optimum value of GDOP is 1.6, while a realistic value may lie between 2 and 3, and in worst case it can exceed 10. When the number of

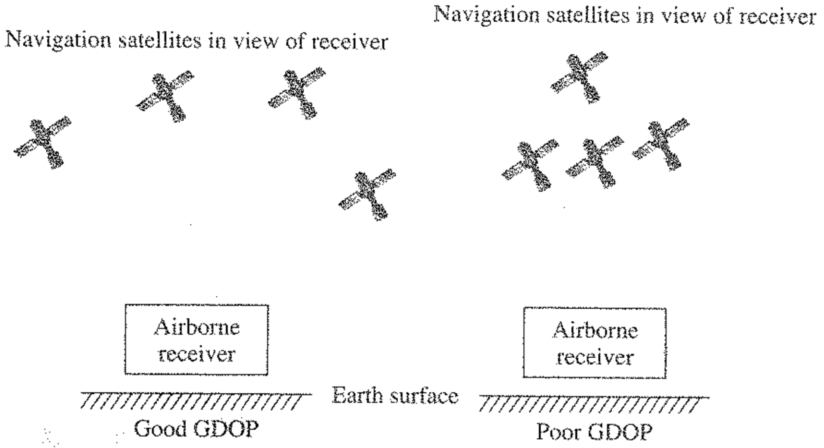

Navigation satellites in view of receiver
Navigation satellites in view of receiver
Airborne receiver
Airborne receiver
Earth surface
Good GDOP
Poor GDOP

Figure 6.5 Illustration of GDOP.

satellites in view is much more than four, GDOP provides a criterion for selecting the optimum four.

The co-variance matrix $[H^T H]^{-1}$ is a (4 × 4) matrix and the diagonal terms define some important aspects of the different types of dilution of precision which are normally associated with satellite navigation. Some of these useful expressions are as follows:

GDOP = $(a_{11} + a_{22} + a_{33} + a_{44})^{1/2}$ is called the Geometric Dilution Of Precision (6.7)

PDOP = $(a_{11} + a_{22} + a_{33})^{1/2}$ is called the Position Dilution Of Precision (6.8)

HDOP = $(a_{11} + a_{22})^{1/2}$ is called the Horizontal Dilution Of Precision (6.9)

VDOP = $(a_{33})^{1/2}$ is called the Vertical Dilution Of Precision (6.10)

TDOP = $(a_{44})^{1/2}$ is called the Time Dilution Of Precision (6.11)

where $(a_{11}, a_{22}, a_{33}, a_{44})$ are the diagonal terms of the co-variance matrix.

The significance of the above non-dimensional DOPs is that if the pseudo range measurement errors σ_{PR} are assumed to be uncorrelated, then they can be transformed into estimation errors as:

One sigma error of (geometric, position, horizontal, vertical, time)
= (GDOP, PDOP, HDOP, VDOP, TDOP)σ_{PR} (6.12)

6.2.3 Position Computation

The position computation, using the measured pseudo ranges, is required to be done fast and accurately using the user processor. As a result, numerous schemes have evolved over the years. Equation (6.5) describes a set of four nonlinear equations which have four unknowns (x, y, z and b_u), to be determined. These nonlinear equations are typically solved by either iterative or non-iterative technique. In the iterative approach, an initial estimate of the user position and clock bias is made. If the first estimate of this user position is x_1, y_1, z_1, then Δx, Δy, Δz and Δt constitute the displacement from actual position x, y, z and from time b_u.

Typical steps **[Sarin et al., 2012]** involved in such iterative method are as follows:

1. Set initial position components as (0, 0, 0) and the clock bias as 0 if no other initial estimates are available.
2. Calculate range r_1^i to all satellites with the approximate user position.
3. Calculate the satellite geometry matrix H. This is obtained from the matrix coefficients a_{xi}, a_{yi}, a_{zi} where these coefficients denote the direction cosines of the unit vector $\mathbf{a}_i$ pointing from the approximate user position to the ith satellite. Mathematically, the equations are:

$$r_1^i = \sqrt{[(X^i - x_1)^2 + (Y^i - y_1)^2 + (Z^i - z_1)^2]} \tag{6.13}$$

$$a_{xi} = \frac{X^i - x_1}{r_1^i}; \quad a_{yi} = \frac{Y^i - y_1}{r_1^i}; \quad a_{zi} = \frac{Z^i - z_1}{r_1^i} \tag{6.14}$$

$$H = \begin{bmatrix} a_{x1} & a_{y1} & a_{z1} & 1 \\ a_{x2} & a_{y2} & a_{z2} & 1 \\ a_{x3} & a_{y3} & a_{z3} & 1 \\ a_{x4} & a_{y4} & a_{z4} & 1 \end{bmatrix} \tag{6.15}$$

4. Compute H^{-1} and $\Delta\rho$

$$\Delta\rho = \rho_1^i - \rho^i \tag{6.16}$$

where $\Delta\rho$ is the difference in the pseudo range estimate from the first approximate user position and the actual measured pseudo range.

5. Compute ΔX from:

$$\Delta X = H^{-1}\ \Delta\rho \tag{6.17}$$

6. Add the values of Δx, Δy, Δz and Δt obtained from ΔX matrix to the initial estimate.
7. Continue the iteration till the root sum square of these four variables become less than a specified threshold. At this stage, the iteration is stopped and final values of the user position (x, y, z) and clock bias b_u are declared.
8. If more than four satellites are available with acceptable GDOP, then a least square solution is applied to ΔX matrix computation and this can be expressed as:

$$\Delta X = (H^T H^{-1})H\Delta\rho \tag{6.18}$$

Iteration is repeated till a convergence is reached.

In the Bancroft method [**Bancroft, 1985**], the method adopted is to solve the nonlinear equations using non-iterative techniques. This method has the advantage of reducing the computational time when compared with the iterative method described earlier. The choice of navigation frame for position computation depends on the user requirement, and for the majority of earth based users; the frame is geographic such as ENU or NED.

6.3 Range Error Minimisation Schemes

The accuracy with which the computed position is available depends on, besides the GDOP factor, the error in range measurement, and these aspects of errors have been described earlier. The receiver clock bias is the largest error contributing to range error. With the correction for receiver clock bias implemented in the position computation, the error parameters which remain to contribute to the ranging error in GPS is shown in Table 6.2. Some schemes have been evolved which eliminate or reduce some of these errors and the schemes are known as:

1. Differential operation
2. Dual frequency measurement scheme
3. Carrier phase measurement scheme.

6.3.1 Differential Operation

Differential operation is a method by which measurement errors inherent to the same satellites, particularly the ephemeris errors and those caused by reduction in the velocity of light during travel through the atmospheric layers, are eliminated to a great extent. Earlier, when Selective Availability (SA) was in operation to intentionally degrade position accuracy, differential operation provided significant improvement in position determination by completely eliminating the error due to SA. This was, perhaps, the biggest motivation to conceive differential operation.

The differential operation consists of partial compensation of measurement error of pseudo ranges by means of their evaluation at correcting stations located at points of known co-ordinates and transmission with the help of special instrumentation of corresponding corrections to the receiver. It is also to be noted that as the object moves away from the reference station, the effectiveness of compensation becomes less, thus creation of a network of monitoring stations becomes important. From earlier days of GPS usage, correction methods and systems have progressively evolved through local area DGPS and wide area DGPS involving ground stations to sophisticated Wide Area Augmentation System (WAAS) and Satellite Based Augmentation System (SBAS). Each of the six space faring nations has evolved its own SBAS to serve its own requirement especially for civilian aircrafts and the airports. Indian GAGAN is also such a SBAS to serve its airports and aircrafts. While actual SBAS configuration is country specific, a general configuration involves a set of adequately dispersed reference stations, a master station along with a stand by, up linking station with dish antenna and all these are located on ground. There is one space based geo stationary satellite and all linked through a system of communication (Figure 6.6).

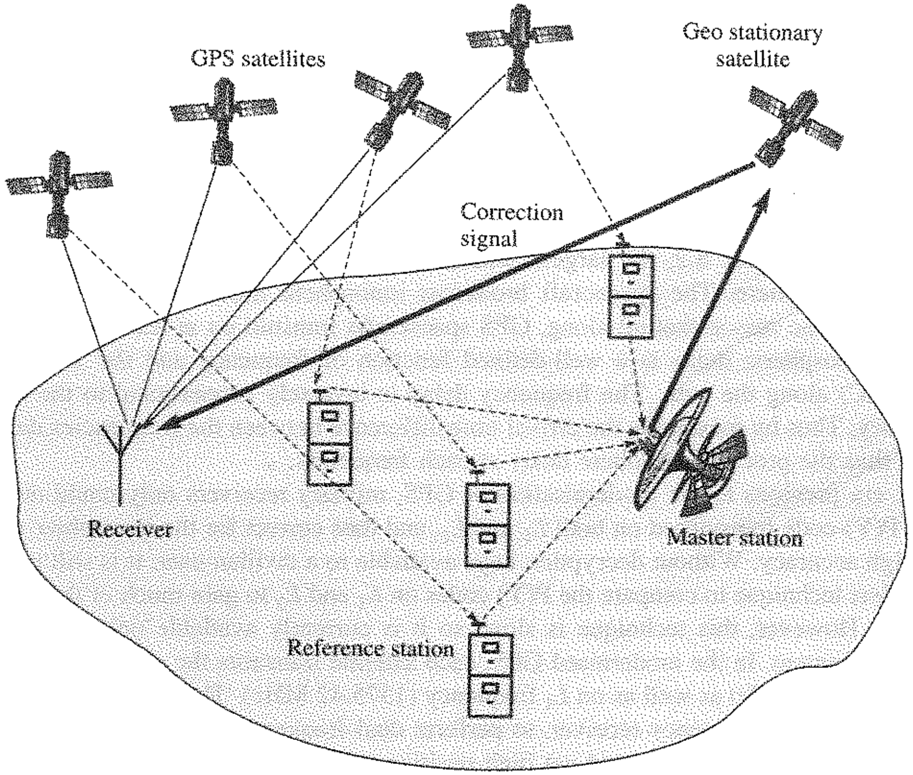


Figure 6.6 Space Based Augmentation System (SBAS).

All the reference stations receive data from the corresponding GPS satellites and the data is forwarded to the master station, which processes the data to determine the integrity, differential correction required and ionosphere related information for each of the monitored satellite. Master station sends the data to the up linking antenna station to uplink the message to the geo stationary satellite. This data, called correction signal, is then downloaded to the receiver in the same L-band ranging signal frequency with GPS type modulation. The receiver then corrects the received GPS data to get an enhanced accuracy.

During the operation in the differential regime, considerable improvement in accuracy for the estimation of the co-ordinates is achieved due to the use of phase measurement of satellite navigation system. Accuracy of single frequency receiver has considerably improved under the organisation of the differential regime as shown in Table 6.3.

Table 6.3 Regimes of Use and Accuracy of Differential Operation

Receiver feature	Regimal accuracy	
	Position	*Velocity*
Conventional, single frequency	15–30 m	0.1–0.3 m/s
Differential, single frequency	2–10 m	0.05–0.1 m/s
Use of phase measurements with DGPS	20–30 cm	

6.3.2 Dual Frequency Measurement Scheme

Accuracy can be improved through precise monitoring and measurement of existing GPS signals in additional or alternate ways. One such scheme is the dual frequency receiver that takes advantage of the dual frequency signal spectrum for civilian use that the modernised GPS system is providing since the year 2005.

As seen from Table 6.2, the largest remaining error is usually the unpredictable delay through the ionosphere. The spacecraft broadcast ionosphere model parameters, but errors remain. This is one reason that all along, GPS spacecrafts transmit on at least two frequencies, L_1 and L_2. Ionospheric delay is a well-defined function of frequency and the Total Electron Content (TEC) along the path. The frequency delay is inversely proportional to the square of the frequency. Thus by measuring the arrival time difference between the frequencies determines TEC, and thus the precise ionosphere delay at each frequency.

In the era previous to the modernisation of GPS, military receivers only could decode the precision P(Y)-code transmitted on both L_1 and L_2, and thus correct for the ionosphere error and provide high accuracy. Without decryption keys available to a civilian user, it is still possible to use a *codeless* technique to compare the P(Y) codes on L_1 and L_2 to gain much of the same error information. However, this technique is slow, so it is currently available only on specialised surveying equipment. In the modernised GPS, each satellite transmits the same C/A code on L_2 frequency (1227.6 MHz) as well as on L_1 frequency (1575.42 MHz). This enables all civil users, equipped with a dual frequency receiver, to perform dual-frequency measurements and directly compute ionosphere delay errors and get high position accuracy. Old receivers can continue to work with L_1 frequency. Addition to the above modernisation, the new GPS is also providing a new L_5 signal (1176.45 MHz) modulated by a new code structure, different from C/A code, that will also be available to civilian user especially for higher accuracy application.

6.3.3 Carrier Phase Measurement Scheme

While pseudo range measurement using the C/A code provides an unambiguous measurement scheme for user position, a much higher level of precision measurement is seen to be possible by measuring the received phase of the carrier, either L_1 or L_2. This happens as the carrier wave has a very short period of 0.63 ns (for L_1), and the noise induced error in measuring the signal delay is around 100 times smaller compared to that encountered in code delay measurement. However, what is observed is the fraction part of the phase as the phase angle repeats at 2π intervals, and this result in an integer ambiguity number N^i for each of the observed satellites at any point of time as shown in Figure 6.7.

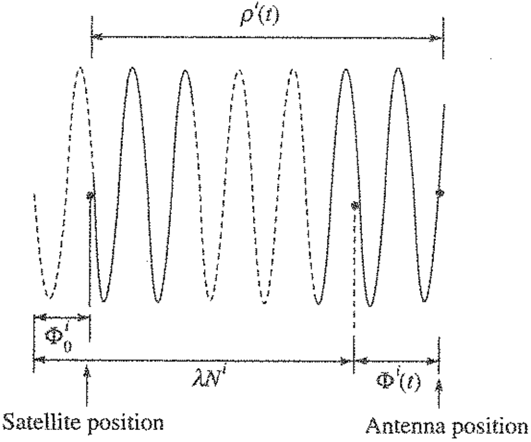


Figure 6.7 Carrier phase measurement and the ambiguity.

The carrier phase observation (variables observed directly) model of the signal sent by the satellite which is received by the receiver antenna at time t, is expressed as:

$$\Phi^i(t) = \rho^i(t) - \lambda N^i + \Phi_0^i + \text{Error terms} \tag{6.19}$$

where

i = observed satellite number (1, 2, 3, 4, ...)
λ = carrier wavelength (19 cm for L_1 carrier)
ρ^i = geometric range from the antenna to the ith satellite
Φ_0^i = initial phase at the time of transmission
N^i = integer carrier phase ambiguity cycles contained in the range between the satellite and the antenna.

Error terms contain all the errors described earlier in Eq. (6.6). Various schemes have been evolved to narrow down the uncertainty range of integer values, such as the use of pseudo range measurements from the same set of satellites or the use of dual frequency receiver, to assist in establishing unambiguously the integer value N^i. But it is seen that with a single receiver, it is not certain to completely eliminate the ambiguity. As a result, currently, high position accuracy with the carrier phase measurement is obtained only with differential scheme discussed in Section 6.3.1 and extent of achievable high accuracy is shown earlier in Table 6.3.

6.3.4 Attitude Determination

The principle of differential carrier phase measurement has been extended to determine the attitude of a vehicle. Various schemes of differential observations have evolved over the years primarily to improve upon the accuracy of attitude measurement and these are known as single difference scheme and double difference scheme. In single difference scheme, refer Figure 6.8, there are two antennae which are separated by a baseline distance. Each antenna is connected to its receiver. Carrier phase measurements are taken from each antenna to the ith satellite and then difference is taken. With the difference, some of the common mode errors, such as satellite clock error and initial phase error, are cancelled out. The single difference scheme can be represented by

$$\Delta\Phi_{21}^{i}(t) = \Delta\rho_{21}^{i}(t) - \lambda\Delta N_{21}^{i} + \text{Error terms (remaining)} \tag{6.20}$$

where Δ represents the difference in observations made at a time, for example

$$\Delta\Phi_{21}^{i} = \Phi_{2}^{i} - \Phi_{1}^{i};\quad \Delta\rho_{21}^{i} = \rho_{2}^{i} - \rho_{1}^{i};\quad \Delta N_{21}^{i} = N_{2}^{i} - N_{1}^{i}$$

Remaining errors are those which do not get cancelled out by single difference.

There is another model [Garcia et al., 2005] to represent the single difference observation model for attitude determination, and it is called carrier phase interferometric model. This model is based on the premise that the distance to satellite is extremely high in comparison to the antenna baseline distance that makes the received signal as a planar wavefront. This means that the line of sight vector from antenna 1 to the ith satellite, $\mathbf{u}_1^{i}$ is parallel to the line of sight vector $\mathbf{u}_2^{i}$ from antenna 2 to the same satellite. So, the difference in distance between the satellite with each of the antennas can be approximated as the vector dot product between the line of sight unit vector $\mathbf{u}^{i}$ and the baseline vector $\mathbf{r}_{21}^{i}$. The interferometric scheme is shown in Figure 6.9.

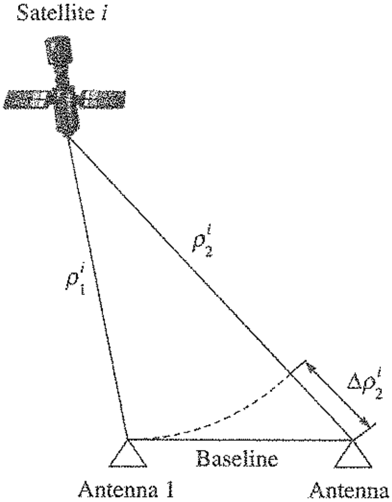


Figure 6.8 Single difference observation scheme for attitude.

The interferometric observation can be represented as:

$$\Delta\rho_{21}^{i}(t) = [\rho_{2}^{i}(t) - \rho_{1}^{i}(t)] \approx \mathbf{r}_{21}^{e}\mathbf{u}_{1}^{i} = (\mathbf{r}_{21}^{e})^{T}\mathbf{u}_{1}^{i} \tag{6.21}$$

The superscript **e** in $\mathbf{r}_{21}^{e}$ implies that the baseline 21 is referred to reference co-ordinate system of the GPS which is ECEF (**Kaplan, 1996**). Since the antennas are fixed on the vehicle body, a transformation is needed from body to ECEF system and this can be represented as

$$\mathbf{r}_{21}^{\mathbf{e}} = C_b^e \, \mathbf{r}_{21}^{\mathbf{b}} \tag{6.22}$$

where C_b^e is the *b-frame* to *e-frame* transformation matrix and $C_b^e = [C_e^b]^T$

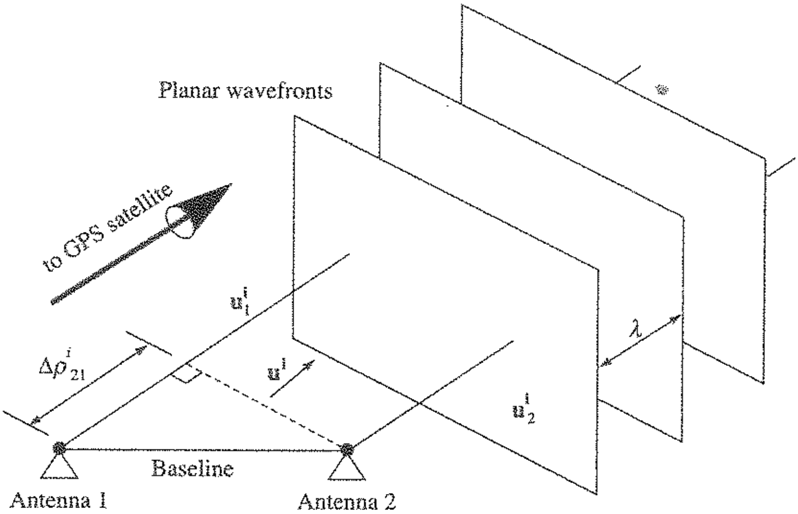


Figure 6.9 Interferometric model for single difference scheme.

In the *b-frame* orthogonal co-ordinate system, $\mathbf{r}_{21}^{\mathbf{b}}$, the baseline vector can be represented as:

$$\mathbf{r}_{21}^{\mathbf{b}} = \begin{bmatrix} x_2 - x_1 \\ y_2 - y_1 \\ z_2 - z_1 \end{bmatrix}_b$$

The matrix is known through measurements at the time of antenna installation. Normally, the receiver navigation frame is a local level *G-frame* and initially, the vehicle body axes (pitch, yaw and roll) are aligned to this *G-frame*. The attitude of the vehicle is defined as rotation about pitch, yaw and roll with respect to the *G-frame* and then the transformation takes the shape as:

$$C_b^e = C_b^G C_G^e \tag{6.23}$$

Finally, using Eqs. (6.21), (6.22) and (6.23), the differential phase Eq. (6.19) can be rewritten. The integer uncertainty ΔN_{21} remains to be resolved as the baseline distance most likely to be more than $\lambda/2$ (~ 8.5 cm), the length up to which integer resolution is not an issue. However, since the baseline distance is known in advance, it is possible to know the maximum number of λ which the baseline can accommodate and which is one factor that helps in resolving ΔN_{21}. Depending on application, attitude algorithms have been developed that cater for the resolution of integer uncertainty as well as reduction of multipath error, receiver clock bias and noise. Currently, the phase error has been reduced to around 5 mm rms which provides a guideline on the choice of baseline distance with respect to attitude accuracy desired. For three-axis rotational measurement, three non-collinear antennae are required.

Double difference scheme is an extension of the single difference scheme. In this scheme, two single difference observations are made, one for the *i*th satellite and the other for the *j*th satellite and then a further difference is taken. The objective is to cancel out the receiver clock error.

6.4 Velocity Measurement

Precise velocity measurement is important for all applications involving aerospace and also in several land based usages. The conventional way of measurement of velocity is either by directly estimating from the Doppler shift of the received carrier frequency or by differencing the position solution. The Doppler shift relating to the user velocity can be expressed as:

$$f_{di} = \frac{1}{\lambda}(\mathbf{v} \cdot \mathbf{u}_i - \mathbf{v}_i \cdot \mathbf{u}_i) + f_b \tag{6.24}$$

where

f_{di} = Doppler shift in Hz with respect to *i*th satellite carrier frequency where $i = 1, 2, 3, \ldots,$
$\mathbf{v}$ = unknown velocity vector (v_x, v_y, v_z) to be determined
$\mathbf{v}_i$ = known velocity vector of the *i*th satellite
$\mathbf{u}_i$ = unit vector from the receiver antenna to the *i*th satellite
f_b = receiver clock bias.

If $\mathbf{p}_i$ is the displacement vector from the receiver to the *i*th satellite, then unit vector $\mathbf{u}_i$ is determined from the relation:

$$\mathbf{u}_i = \frac{\mathbf{p}_i}{|\mathbf{p}_i|} \quad \text{where } \mathbf{p}_i = [(x_i - X), (y_i - Y), (z_i - Z)]^T \tag{6.25}$$

The receiver position can be established using the pseudo range equations described earlier.

Another method of determining the velocity can be made by computing the Doppler shift from the carrier phase measurements using the time difference technique which can be expressed as:

$$f_{di}(t) \approx \frac{\Phi_i(t + \Delta t) - \Phi_i(t - \Delta t)}{2\Delta t} \tag{6.26}$$

where

t = epoch when the velocity is calculated
Φ_i = carrier phase measured with *i*th satellite
Δt = sample time of measurement.

The Doppler shift is approximate when acceleration is not constant during the measurement sample time. Considering the dynamic limitation of the approach in Eq. (6.26), new algorithms **[Ding et al., 2009]** have been developed which computes 'delta position vector' to take into account change of acceleration during the interval. The acceleration change is approximated as piece-wise constant between the two epochs t_1 and t_2. The position is calculated using the carrier phase measurement scheme that has been discussed earlier.

6.5 GPS Time

GPS is a versatile and global tool to distribute time to an arbitrary number of users and synchronise clocks over large distances with high degree of precision and accuracy. The accuracy comes due to the use of atomic clock in each satellite, and each receiver is able to receive the accurate time tagged signal. As a part of the navigation solution, the computer within the GPS receiver can determine the difference between the user clock and either (i) GPS time, or (ii) the reference time for GPS which is UTC (USNO). The local receiver can be programmed to display either the UTC (USNO) as broadcast by GPS or the GPS time since the navigation solution yields the difference between the local clock and the GPS time. The UTC (USNO) is steered to UTC determined by Bureau International des Poids et Measures (BIPM) and usually kept within about 10 ns of UTC. The GPS synchronises time over large distances by a technique known as Common View (CV) where two stations observe the same satellite. The two users record their differences between local time and GPS time at the same instant using the same satellite in order to minimise the effect of errors. Then one can calculate the difference between the two local clocks.

6.6 Features of GLONASS

Over the last few years, considerable efforts have been made on the design and manufacture of dual GPS and GLONASS receivers. This is likely to increase the acceptance of GLONASS system and increase the robustness of Global Navigation Satellite System (GNSS) for world wide navigation. Comparative features of the GLONASS with respect to GPS is shown in Table 6.4.

Table 6.4 Comparison of GPS and GLONASS Features

Sl No.	GPS	GLONASS
1	Developed and maintained by US Government	Developed by USSR now maintained by Russian Republic
2	NAVSTAR satellite system	Global Orbiting Navigation Satellite System
3	24 Satellites	24 Satellites
4	6 Orbits of 4 satellites each	3 Orbits of 8 satellites each
5	Orbit inclination: 55°	Orbit inclination: 64.80°
6	Orbit radius: 26, 560 km	Orbit radius: 25, 510 km
7	Non-geostationary	Non-geostationary
8	Orbit right ascension 60°	Orbit right ascension 120°
9	Period: 1/2 sidereal day	Period: 8/17sidereal day
10	17 Revolutions in 8 days	16 Revolutions in 8 days
11	L-band carrier signals L_1 with f_1 = 1575.42 MHz L_2 with f_2 = 1227.6 MHz	L-band carrier signals L_1 with f_1 = (1.597–1.617) MHz L_2 with f_2 = (1.240–1.260) MHz
12	Receiving and analysing similar	Receiving and analysing similar

By November, 2011, GLONASS reached its full 24-satellite constellation to provide world wide coverage. According to Russian system of differential correction and monitoring of data as of 2010, precisions of GLONASS navigation definitions (for $p = 0.95$) for latitude and longitude were 4.46–7.38 m with mean number of navigation satellite visibility equals 7–8 (depending on station). In comparison, the precisions of GPS navigation definitions were 2.00–8.76 m with mean number of visible satellites equals 6–11 (depending on station). Civilian GLONASS, when used alone, is therefore slightly less accurate than GPS. On high latitudes (north or south), GLONASS accuracy is better than that of GPS due to the orbital position of the satellites.

Some modern receivers are able to use both GLONASS and GPS satellites together, providing greatly improved coverage in urban canyons and giving a very fast time to fix due to over 50 orbital satellites being available. In indoor, urban canyon or mountainous areas, accuracy can be greatly improved over using GPS alone. For using both the navigation systems simultaneously, precisions of GLONASS and GPS navigation definitions were 2.37–4.65 m with mean number of visible satellites equals 14–19.

6.7 Comparison between INS and SNS

1. INS is autonomous in nature and thus the method of INS operation protects itself against jamming or similar other interference during war. SNS operation is not autonomous and susceptible to jamming or other interference as the operation of the system is in the hand of the country that is responsible for the system.
2. INS shows a time growing low frequency error characteristic, while the SNS error is of high frequency characteristic devoid of time growing propagation.
3. Standalone INS can provide navigation output at a high update rate greater than 50 Hz, whereas a standalone SNS normally provides navigation output at 1 Hz to 10 Hz.
4. All strapdown INS provide vehicle attitudes and body rates at equally high update frequency, whereas some SNS receivers have been designed with widely spaced multiple antennae to provide attitude but at a slow update rate.
5. INS operates below the sea, inside a tunnel or inside oil drilling rig. SNS does not work under these conditions.
6. Loss of lock can be a serious problem in a standalone SNS under high dynamic condition where time to regain locking may be too large for the safe working of the vehicle.

These comparisons show some of the deficiencies in a standalone SNS inspite of its superior position performance and considerably lower cost in comparison to INS. All these shortfalls in SNS are overcome by integrating with a low cost inertial system.

6.8 Spaceborne SNS

An emerging area of application of SNS is in spacecraft navigation to provide its current position and velocity which is not feasible with INS as the spacecrafts are in freefall condition. The high velocity of the orbiting spacecrafts, typically 7–8 km/s, poses interesting problems for the receiver design, where the signal Doppler and Doppler rate are high, which are normally not encountered in earth based usage including aircrafts. Also the visibility of a SNS satellite under tracking from a spacecraft receiver, is only around 45–50 minutes as against around 6 hours for earth based usage. These requirements show that for space borne receiver, considerable design augmentation is needed on a receiver that is qualified for usage over land and in air. Another requirement arises is to provide three-axis attitude of spacecraft which is currently met with gyros, star sensor and sun sensor. Attitude information from SNS is an evolving technology, especially for use in spacecraft as it provides independence from the use of the costlier attitude sensors mentioned earlier. It is an evolving technology as obtaining three-axis attitude with adequate accuracy, low noise and the required attitude update rate, require continuous research and development efforts. Radiation tolerance of space borne receiver electronic devices is another key requirement to be ensured.

The development of SNS receiver has been extended for use in geostationary orbit, the orbit that is higher compared to SNS satellite orbit. This has additional and different sets of problem as it can receive signal with adequate signal-to-noise ratio from SNS satellites which are located on the other side of the earth, and which are also not obstructed by the earth. The number of such visible satellites at any time are currently around 6, with GPS coverage alone, having signal-to-noise ratio >20 dB-Hz. However, these visible satellites offer poorer GDOP, and the computer simulated position error is >750 m primarily due to poor GDOP of the observed satellites [**Chibout et al., 2007**].

With the cost and the size of the receivers reducing, considerable number of emerging applications [**Bauer et al., 1998**] are predicted for space borne SNS covering areas, e.g.

1. Autonomous onboard navigation, operation and orbit control
2. Attitude determination and control
3. Rendezvous and proximity operations
4. Formation flying and co-ordinated platforms.

SUMMARY

The chapter has described the working of satellite navigation system to provide position, velocity, time and attitude. The errors associated with the range measurement process and the measures to reduce the error are described, which include the current operation of differential schemes, two-frequency receiver and the carrier phase measurement.. The typical position computational steps are brought out with iterative scheme along with explanation for dilution of precision and their effect on navigation error. The new features of the modernised GPS signals have been brought out that aims to give much higher accuracy through the use of C/A code. Features of GLONASS have been brought out as this satellite navigation system is already operational and receivers are manufactured which can track both GPS as well as GLONASS. A comparative characteristic between INS and SNS has been brought out indicating how an integrated INS and SNS can provide a low-cost solution to a large number of applications. Lastly, how the space borne SNS is finding newer application has been brought out.

EXERCISES

6.1 If H is the satellite observation matrix, then

(i) If number of satellites observed is four, what is the size of the H matrix?

(ii) Explain the significance of the term DOP. How do you mathematically express the GDOP of the observed satellites?

(iii) The covariance matrix H is given as:

$$\begin{bmatrix} 0.672 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.672 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.6 & -0.505 \\ 0.0 & 0.0 & -0.505 & 0.409 \end{bmatrix}$$

Calculate the GDOP. Is it good or poor? [**Ans:** (i) 4 × 4, (iii) 1.831, Good]

6.2 Two GPS type RF transmitters and the receiver are located on the same plane X–Y.

(i) Sketch the problem and write the range equations to obtain the receiver location.

(ii) What is the difference between pseudo range and geometric range? If the receiver clock has a bias, how many more transmitters are needed to locate the receiver co-ordinates? Write the range equation under this scenario.

6.3 Co-ordinates of a navigation satellite in ECEF frame is given as:

$$Z_s = 26{,}560 \text{ km}; \quad X_s = Y_s = 0$$

Similarly, the co-ordinates of the receiver position are given as Latitude = Longitude = 45° and altitude as 0. Compute

(i) Geometric range to the satellite

(ii) Elevation angle of the line of sight to the satellite

(Assume a spherical earth model and nominal earth radius of 6378 km).

[**Ans:** (i) 22,506 km, (ii) 101.58°]

6.4 (i) In the computation of pseudo range with observation from four satellites, should the receiver clock bias estimation vary from one to the other? Give reason to your answer.

(ii) One form of error in pseudo range is associated with electromagnetic wave propagation through atmosphere. Briefly describe the reason for this error and describe the methods adopted to reduce it.

(iii) If the number of observable satellites with acceptable GDOP is more than four, what is normally done for computing the position?

6.5 For the Problem 6.1, compute PDOP, HDOP and TDOP. What are the significance of these terms? [**Ans:** PDOP = 1.71, HDOP = 1.15, TDOP = 0.64]

REFERENCES

Bancroft, S., An Algebraic Solution of the GPS Pseudo Range Equations, *IEEE Transactions on Aerospace and Electronic Systems*, AES-21, pp. 55–59, November 1985.

Bauer, Franc H, Hartman Kate., and Lightsey, E. Glenn., Spaceborne GPS Current Status and Future Visions, NASA Goddard Space Flight Centre, Greenbelt, Maryland, USA, 1998.

Chibout, B., Macabiau, C. et al., Comparison of Acquisition Techniques for GNSS Signal Processing in Geostationary Orbit, *Proceedings of the 2007 National Technical Meeting of The Institute of Navigation*, CA, San Diego, pp. 637–649, January 2007.

Ding, W. and Wang, J., Precise Velocity Estimation with a Stand-Alone GPS Receiver, School of Surveying and Spatial Information Systems, The University of New South Wales, Australia, 2009.

Garcia, J.G., Mercader, P.I., and Muravchik, C.H., Use of Carrier Phase Double Differences, *Laboratorio de Electronica Industrial Control e Instrumentacion* (LEICI), UNLP CC No. 91, Argentina, 2005.

Grewal, Mohinder S., Weill, Lawrence R., and Andrews, Angus P., Global Positioning Systems, *Inertial Navigation and Integration*, John Wiley & Sons, New York, 2001.

Kaplan, E.D., *Understanding GPS: Principles and Applications*, Artech House Telecommunication Library, Boston, USA, 1996.

Kintner, Paul M. and Ledvina, Brent M., The Ionosphere, Radio Navigation, and Global Navigation Satellite Systems, School of Electrical and Computer Engineering, Cornell University, Ithaca, USA.

Sarin, S., Patel, Y.R. et al., Comparison and Analysis of Iterative and Non-Iterative Methods of GPS Solution from Pseudo Range Equations, *Proceedings of NAVCOM*, pp. 28–34, Hyderabad, December 20–21, 2012.

7

Integrated Inertial Navigation

Integrated inertial navigation is the process of estimating the navigation states such as position, velocity and attitude by using measurements from multiple navigation sensors one of which is an inertial system. Integrated stellar–inertial, integrated Doppler–inertial, barometric damped–inertial system, integrated stellar–inertial–Doppler and integrated SNS–INS, are some of the examples of integrated inertial navigation systems. Integrated GPS–INS system, which is currently operational, is being further extended for operationalisation of integrated INS–Global Navigation Satellite System (GNSS), as integration will be possible with emerging satellite receivers such as GLONASS or Galileo or even with a combined GPS–GLONASS receiver. Inertial attitude reference system for spacecrafts routinely uses star sensor or sun sensor data in orbit for calibration of gyro drift. Such involvement of more than one sensor has given rise to another name for integrated inertial system and is called *multisensor system*. Quite often, the terminology used is 'aided inertial system' that normally refers to unidirectional aiding to limit inertial navigation system time dependent error growth.

While the underline requirement and the basic principle of integrated inertial navigation have been stated in Section 1.6 in Chapter 1, the mathematical formulation of the basic principle of aiding with Kalman filter, various integrated configurations and their working methodologies, which are currently in operation, are further elaborated in this chapter.

Kalman filter advantage

The integration algorithm or the estimation methodology must consider the deterministic and statistical properties of each of the sensors being integrated to achieve optimal performance. Kalman filter framework, with many advantages over other methods, is considered to be the most optimal approach for multisensor navigation systems. Kalman filter estimator is the optimal observer **[Lewis, 1986]** of a linear dynamic system for both time invariant and time varying in which the statistical disturbances are assumed to be zero mean white Gaussian, either stationary or non-stationary. From the estimation theoretic point of view, it is a linear Bayesian Minimum Mean Squared Error (MMSE) estimator [**Kay, 1993**], which is different from classical estimators such as Best Linear Unbiased Estimator (BLUE) and Maximum Likelihood Estimator (MLE), wherein the parameters to be estimated are treated as deterministic but unknown. In classical estimators, there is no way of feeding in the priori information into the estimation process. Bayesian estimation, on the other hand, treats the parameters to be estimated as random variables with conditional posteriori Probability Density Function (PDF) wherein available prior information can be fed as Random Variables (RV) with priori PDFs. Hence, Bayesian estimators are considered superior to classical estimators in general.

Kalman filter is a formulation in state space form for time domain approach to Multi Input Multi Output (MIMO) systems and it naturally accommodates linear time varying systems and non-stationary noise statistics which frequency domain approach cannot handle. In fact, this is one of the key differences between a Wiener filter and a Kalman filter where Wiener filter is an optimal filter design using frequency domain approach, which demands linear time invariance and stationary noise statistics.

7.1 Processing of Measurements

One primary objective of processing of measurement is to obtain a best estimate of the measurements in the presence of noise. In the simplest case, we deal with single sensor and obtain a series of measurements which are spread over time. Then there is the more complex case where more than one sensor is involved and we try to get measurements from all the sensors at the same time and then compute the best estimate in the presence of identical noise in each. Section 7.1.1 explains the basic method followed for each.

7.1.1 Single and Multiple Sensor Cases on a Stationary Vehicle

Assume a single sensor makes n measurements spread in time of a stationary state X. Let X_i, $i = 1, 2, ..., n$ be the different measurements spread in time of a true state X_μ.

X_i, $i = 1, 2, ..., n$ are the different realisations of the random variable X with $N(X_\mu, \sigma)$ at different instants. The mean of these n measurements is an estimate of the true state X_μ at instant n. Thus the estimate is given by

$$\hat{X}_\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{7.1}$$

where $N(X_\mu, \sigma)$ is normal distribution with mean X_μ and variance σ^2.

The new random variable $\hat{X}_\mu$ has the distribution $N\left(X_\mu, \frac{\sigma}{\sqrt{n}}\right)$. Note that this sample mean estimator is another random variable, which reminds us that the performance of an estimator can only be determined statistically. Lower the variance of $\hat{X}_\mu$, the better is the estimator.

In this example, we have assumed a single sensor with n measurements spread in time and hence the above sample mean is actually the time average.

The same result could be obtained if we assume n identical sensors are used to get n measurements at the same instant and each is affected by identical noise source of zero mean and can be expressed as: $N(0, \sigma)$. In this case also, the mean of these n measurements, which are available simultaneously (commonly referred as n-ensembles), is same as:

$$\hat{X}_{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This ensemble average [**Taub et al., 1986**] has the same statistical property as the time average. The latter case can be thought of n-sensors on a stationary vehicle, but each sensor is affected by identical noise statistics.

7.1.2 Multiple Sensors on a Stationary Vehicle with Different Noise Statistics

For simplicity, we assume two sensors X_1 and X_2 on a stationary vehicle, but each one affected by different noise statistics $N(0, \sigma_1)$ and $N(0, \sigma_2)$ respectively. Then the optimal way to combine these two measurements at instant i is given by

$$\hat{X}_{\mu i} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} X_{1i} + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} X_{2i} \tag{7.2}$$

$$= \lambda_1 X_{1i} + (1 - \lambda_1) X_{2i}; \text{ where } \lambda_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \tag{7.3}$$

In addition, since the state being estimated is stationary, which means $\hat{X}_{\mu}$ is constant, each sensor can be time averaged and then combined to get better estimate. For the above two sensors, this result is as follows:

$$\hat{X}_{\mu_{n1}} = \frac{1}{n}\sum_{i=1}^{n} X_{1i}; \quad \hat{X}_{\mu_1} \text{with } N\left(X_{\mu}, \frac{\sigma_1}{\sqrt{n}}\right)$$

$$\hat{X}_{\mu_{n2}} = \frac{1}{n}\sum_{i=1}^{n} X_{2i}; \quad \hat{X}_{\mu_2} \text{with } N\left(X_{\mu}, \frac{\sigma_2}{\sqrt{n}}\right) \tag{7.4}$$

Combined estimate is given by

$$\hat{X}_{\mu_n} = \frac{\left(\frac{\sigma_2}{\sqrt{n}}\right)^2}{\left(\frac{\sigma_1}{\sqrt{n}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n}}\right)^2} \hat{X}_{\mu_{1n}} + \frac{\left(\frac{\sigma_1}{\sqrt{n}}\right)^2}{\left(\frac{\sigma_1}{\sqrt{n}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n}}\right)^2} \hat{X}_{\mu_{2n}} \tag{7.5}$$

$$\hat{X}_{\mu_n} = \lambda_{1n} \hat{X}_{\mu_{1n}} + (1 - \lambda_{1n}) \hat{X}_{\mu_{2n}}; \quad \lambda_{1n} = \frac{\left(\frac{\sigma_2}{\sqrt{n}}\right)^2}{\left(\frac{\sigma_1}{\sqrt{n}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n}}\right)^2}$$

λ_{1n}, $(1 - \lambda_{1n})$ are the weightages for individual estimates in the combining process. The weightages are 'Convex Coefficients' which can be defined as:

$$\lambda_{1n} + (1 - \lambda_{1n}) = 1$$

It can easily be observed that higher the variance, lesser the weightage for the same and vice-versa.

7.1.3 Combining Measurements and Prediction in Kalman Filter

In the case of dynamic system, where the system states are changing with time, the optimal estimate at every update time interval is provided by Kalman filter estimator using the available measurements, and which also predicts the best estimate when the direct measurements are not available. This mode of working of Kalman filter is described further.

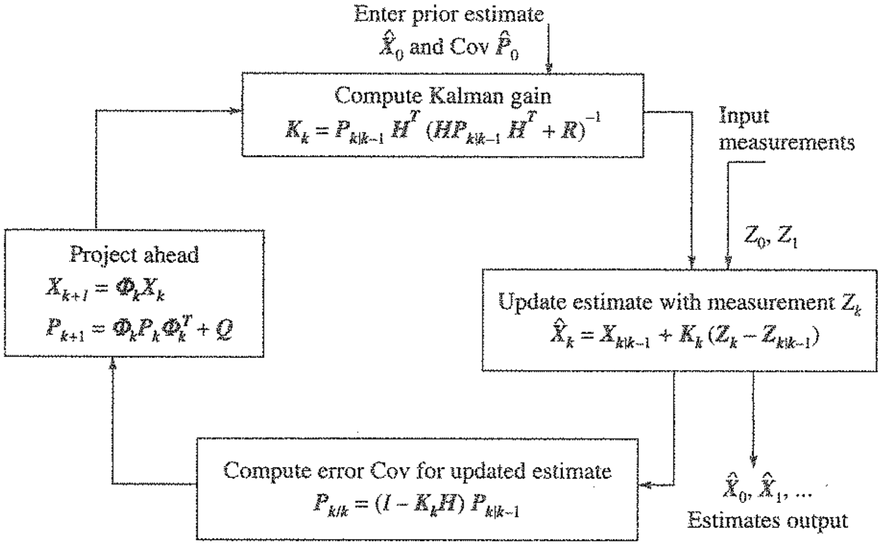

Enter prior estimate
$\hat{X}_0$ and Cov $\hat{P}_0$
Compute Kalman gain
$K_k = P_{k|k-1} H^T (HP_{k|k-1} H^T + R)^{-1}$
Input
measurements
Z_0, Z_1
Project ahead
$X_{k+1} = \Phi_k X_k$
$P_{k+1} = \Phi_k P_k \Phi_k^T + Q$
Update estimate with measurement Z_k
$\hat{X}_k = X_{k|k-1} + K_k (Z_k - Z_{k|k-1})$
Compute error Cov for updated estimate
$P_{k/k} = (I - K_k H) P_{k|k-1}$
$\hat{X}_0, \hat{X}_1, \ldots$
Estimates output

Figure 7.1 Discrete Kalman filter implementation steps.

The block schematic, shown in Figure 7.1, depicts the discrete Kalman filter implementation steps in a concise way. Here it is important to note the step 'update estimate with measurement Z_k' (shown in the right box). This step is repeated here for convenience.

$$\hat{X}_k = X_{k|k-1} + K_k(Z_k - Z_{k|k-1}) \tag{7.6}$$

where

$\hat{X}_k$ = state vector estimate at instant k
$X_{k|k-1}$ = predicted state vector at k, with given information at $k - 1$
Z_k = actual measurement matrix at k with dimension $m \times 1$
$Z_{k|k-1}$ = predicted measurement at k, with given information at $k - 1$
K_k = Kalman gain at instant k with dimension $n \times m$

Φ_k = state transition matrix at instant k with dimension $n \times n$

$\hat{X}_0$ = initial state matrix of dimension $n \times 1$

$\hat{P}_0$ = initial covariance matrix of dimension $n \times n$ of $\hat{X}_0$

Further

$$Z_{k|k-1} = HX_{k|k-1}$$

where

H = measurement matrix with dimension $m \times n$

$P_{k|k-1}$ = covariance matrix of $X_{k|k-1}$

The scalar form of the above equation is useful to gain insight about the measurement update equation and it can be expressed as:

$$\begin{aligned}\hat{X}_k &= X_{k|k-1} + K_k(Z_k - Z_{k|k-1}) \\ &= K_k Z_k + (1 - K_k H)X_{k|k-1}\end{aligned} \tag{7.7}$$

The weightage K_k applied to Z_k, the current measurement and the weightage $(1 - K_k H)$ is applied to the predicted state $X_{k|k-1}$.

If $H = 1$, (unit matrix), Eq. (7.7) simplifies further:

$$\hat{X}_k = K_k Z_k + (1 - K_k)X_{k|k-1}$$

The weightages are K_k and $(1 - K_k)$ for measurement and prediction respectively. We have two information about the state to be estimated. One of these is the measurement Z_k at instant k. The other is the predicted state at instant k using the previous estimate and the model, which represents the state propagation with respect to time. Then to combine these two information with optimal weightages is the true optimal estimation problem.

7.2 Complementary Nature of Sensors

In aided inertial navigation systems, the sensor error behaviours are complementary in nature. For example, in a GNSS–INS system, the INS errors grow with time, and hence its long-term accuracy is poor, but short-term accuracy is good. On the contrary, for GNSS, the long-term accuracy is good and the short-term accuracy is poor. A GNSS receiver stationary at a point will give its position over several days or months with an error bound determined by its short-term error behaviour primarily due to ionospheric and tropospheric fluctuations. The position error does not grow with time. This nature is complementary to that of an INS. Similarly, star sensor output noise characteristic and low frequency gyro drift characteristic are defined as complementary sensor characteristics.

In such cases, this priori information of the complementary error behaviour can be used systematically to obtain an optimal estimate in all multisensor systems provided they have such behaviour. The Kalman filter framework facilitates the use of such priori information and is ideally suited for the fusion of these complementary sensors. The following simple example illustrates the idea of using prior information.

Let the problem be recovering a DC signal corrupted by additive zero mean white Gaussian noise. The above problem has the inherent prior information that the signal is DC. This can be written as a signal model by the following state equation:

$$X_{k+1} = X_k \tag{7.8}$$

where k is discrete time index with sampling time T.

The signal is corrupted by additive measurement noise v_k, and this can be expressed as the following measurement equation:

$$Z_k = X_k + v_k, v_k : N(0, \sigma) \tag{7.9}$$

All that is required for Kalman filter implementation is available in the above two equations. In fact, the Kalman filter performance in this case can be compared with a classical Butterworth Low Pass Filter (LPF) with different cut off frequencies, and we can easily find that the Kalman filter performs always better.

7.3 The Basic GNSS–INS Fusion Using Kalman Filter

We know that the INS errors grow with time primarily due to integrating effect of errors in accelerometers and gyros that eventually contribute to continuous change in velocity and position errors. This motivates us to use the above prior information, and to model the INS error dynamics by a vector state equation of the following form:

$$X_{k+1} = \Phi_k X_k + w_k \tag{7.10}$$

where

X_k = vector of INS error states

Φ_k = linear time varying state transition matrix, independent of X_k

w_k = model uncertainty (process noise) with statistics of zero mean white Gaussian with covariance Q

Further using the priori knowledge of the complementary nature of GNSS and INS, the measurement equation can be formed as follows:

$$\text{INS--GNSS} = Z = HX_k + V_k \tag{7.11}$$

where

V_k = GNSS noise + Any short term noise of INS

The basic GNSS–INS fusion scheme using Kalman filter as an estimator, is shown in Figure 7.2.

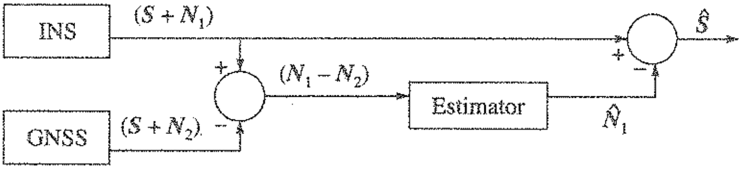


Figure 7.2 A basic GNSS–INS fusion using Kalman filter.

Here

S = true state vector to be estimated

N_1 = slow varying, time growing INS error vector

N_2 = relatively fast varying GNSS noise vector

$(N_1 - N_2)$ = measurement input to the estimator

$\hat{N}_1$ = estimated INS error vector

$\hat{S}$ = estimate of the true state vector S

This is the classical complementary filter configuration [**Brown et al., 1997**]. It can easily be identified that a classical choice for the estimator above is a proper Low Pass Filter (LPF) in the frequency domain approach. If the measurements are position and velocity what we require is 6 inputs, 6 outputs LPF, and hence the estimator will be 6 × 6 transfer function matrix assuming coupling of all inputs to all outputs.

But as we discussed earlier, Kalman filter is the ideal choice for optimal performance and also to easily deal with MIMO systems and linear time varying INS error dynamics $\boldsymbol{\Phi}_k$. Now, the basic equations for Kalman filter for the GNSS–INS fusion can be written as:

$$X_{k+1} = \Phi_k X_k + w_k \tag{7.12}$$

$$Z_k = H_k X_k + v_k \tag{7.13}$$

where Eq. (7.12) represents INS error dynamics, while Eq. (7.13) represents the measurement equation.

A typical example of INS error states, the formulation of $\boldsymbol{\Phi}_k$, $\boldsymbol{H}_k$ for chosen error states and basic assumptions on w_k and v_k, will be elaborated in the following sections. The error state formulation, instead of total state, has the important advantage that the dynamics is linear to a greater extent, which is a major requirement for the Kalman filter.

A typical set of error states in an INS can be written as:

$$H = [\Delta x\ \Delta y\ \Delta z\ \Delta\dot{x}\ \Delta\dot{y}\ \Delta\dot{z}\ b_x\ b_y\ b_z\ m_x\ m_y\ m_z]^T$$

where

Δx, Δy, Δz = position error components of INS in ECI frame

$\Delta\dot{x}$, $\Delta\dot{y}$, $\Delta\dot{z}$ = velocity error components of INS in ECI frame

b_x, b_y, b_z = accelerometer bias in body frame

m_x, m_y, m_z = computational frame misalignments

Measurements

1. $(\Delta x)_{INS} - (\Delta x)_{GNSS}$
2. $(\Delta y)_{INS} - (\Delta y)_{GNSS}$
3. $(\Delta z)_{INS} - (\Delta z)_{GNSS}$
4. $(\Delta\dot{x})_{INS} - (\Delta\dot{x})_{GNSS}$
5. $(\Delta\dot{y})_{INS} - (\Delta\dot{y})_{GNSS}$
6. $(\Delta\dot{z})_{INS} - (\Delta\dot{z})_{GNSS}$

From the definition of states and measurements, the measurement matrix H is:

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

And the error dynamic is defined in continuous time at instant k as follows:

$$A_k = \begin{bmatrix} 0 & I & 0 & 0 \\ J_G & 0 & C_B^I & f \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{7.14}$$

$$\Phi_k = I_{12} + A_k T + (A_k T)^2/2 + (A_k T)^3/6 + \cdots \tag{7.15}$$

where

A_k = sub matrix

T = sampling time interval of the discrete system

C_B^I = transformation matrix from body to inertial frame

J_G = Jacobian of gravitation defined as below for spherical earth assumption

$$J_{G/11} = \frac{\mu}{r^3}\left[-1+\frac{3x^2}{r^2}\right]; \quad J_{G/12} = \frac{\mu}{r^3}\left[\frac{3xy}{r^2}\right]; \quad J_{G/13} = \frac{\mu}{r^3}\left[\frac{3xz}{r^2}\right]$$

$$J_{G/21} = \frac{\mu}{r^3}\left[\frac{3xy}{r^2}\right]; \quad J_{G/22} = \frac{\mu}{r^3}\left[-1+\frac{3y^2}{r^2}\right]; \quad J_{G/23} = \frac{\mu}{r^3}\left[\frac{3yz}{r^2}\right]$$

$$J_{G/31} = \frac{\mu}{r^3}\left[\frac{3xz}{r^2}\right]; \quad J_{G/32} = \frac{\mu}{r^3}\left[\frac{3yz}{r^2}\right]; \quad J_{G/33} = \frac{\mu}{r^3}\left[-1+\frac{3z^2}{r^2}\right]$$

where

x, y, z = position components in ECI frame

$r = \sqrt{x^2+y^2+z^2}$

μ = gravitational constant (refer Section 2.3 in Chapter 2)

$$f = \begin{bmatrix} 0 & f_Z & -f_Y \\ -f_Z & 0 & f_X \\ f_Y & -f_X & 0 \end{bmatrix}$$

f_X, f_Y, f_Z are the specific force components expressed in ECI frame.

7.4 Kalman Filter Equations (Discrete Form)

Kalman filter theory is covered in various publications such as the theoretical aspects can be found in [**Lewis, 1986**], while an application oriented approach can be found in [**Brown et al., 1997**]. The Kalman filter formulation in continuous time domain and in discrete time domain has been covered in these references. The standard Kalman filter equations in discrete time are given below. The covariance update equation $P_{k|k}$ has alternate form such as Joseph stabilised form that has better numerical property but is not described here. Other forms of implementation such as square root filtering and UD factorisation methods can be found in [**Bierman, 1977**].

The divergence issues, innovation sequence (defined as actual measurement–predicted measurement) analysis, whitening compensation for non-white measurement noise, process noise and cross correlation between the measurement and process noise are important that are discussed in [**Kay, 1993; Bierman, 1977**]. The observability and reachability grammians and their relation to the uniqueness of the solution of the matrix Ricatti equation ($P_{k|k-1}$), which has direct relation to the Kalman gain, are covered in these references.

The backward form of Kalman filter is known as information filter, which are used in smoothing applications relevant to offline processing.

In the standard form given below, the process noise Q, measurement noise R and measurement matrix H are assumed to be constant, but they can be time varying as well.

Plant model:

$$X_{k|k-1} = \Phi_{k-1}X_{k-1|k-1} + w_{k-1}$$

Measurement equation:

$$Z_k = HX_k + v_k$$

Initialisation: X_0 and P_0: covariance of X_0

Time update:

$$X_{k|k-1} = \Phi_{k-1}X_{k-1|k-1}$$

$k|k$–1: prediction at k, given $k - 1$

$$P_{k|k-1} = \Phi_{k-1}P_{k-1|k-1}\Phi_{k-1}^T + Q$$

$k - 1|k - 1$: estimate at $k - 1$, given $k - 1$

Predicted measurement:

$$X_{k|k-1} = HX_{k|k-1}$$

Computation of Kalman gain:

$$K_k = P_{k|k-1}H^T[HP_{k|k-1}H^T + R]^{-1}$$

Measurement update:

$$X_{k|k} = X_{k|k-1} + K_k[Z_k - Z_{k|k-1}]$$

$$P_{k|k} = [I - K_kH]P_{k|k-1}$$

The cycle continues. Estimator or the observer as a dynamic system can be shown as:

$$\hat{X}_k = [\Phi_k - K_k H\Phi_k]\hat{X}_k + K_kZ_k \tag{7.16}$$

It is to be noted that the Kalman gain K_k behaves as input matrix and measurements Z_k act as input to the system. The system dynamics $[\Phi_k - K_kH\Phi_k]$ is time varying.

It is clear from the above that since Φ_k and K_k are time varying, the above estimator is a time varying dynamic system, which means a time varying filter in classical sense. Hence this optimal filter cannot be designed by conventional time invariant filter by frequency domain approach.

7.5 Classification and Description of GNSS–INS Integration Schemes

There are several GNSS–INS integration schemes in existence. Each has its merits and demerits based on ease of implementation, accuracy requirement, computational load and user dynamics. The major schemes are:

(a) Uncoupled, (b) Loosely coupled, (c) Tightly coupled, (d) Ultra tightly coupled/deep integration.

These coupling schemes have been categorised based on the level of coupling between the two sensors, which are described in the subsequent sections.

In most of the GNSS–INS integration applications, a GNSS receiver with position, velocity and time outputs are used. While using this type of receiver, aiding for attitude, in general, is not possible because:

1. GNSS receiver does not provide attitude output
2. Attitude may not be observable in many cases from position and velocity measurements.

However, GPS receivers are now available with attitude measurement capability using multiple antennae separated by known distance and baselines on the user vehicle. Attitude information is derived from phase difference measurements of the signal from satellites as received at different antennae using interferometric principle [**Parkinson et al., 1996**]. There are many difficulties still exist to get reliable attitude information mainly due to difficulties in ambiguity resolving uncertainties. In some cases, attitude may become observable from position and velocity due to the nature of the trajectory itself. The discussion to follow assumes that the GNSS receiver provides position, velocity and time output only.

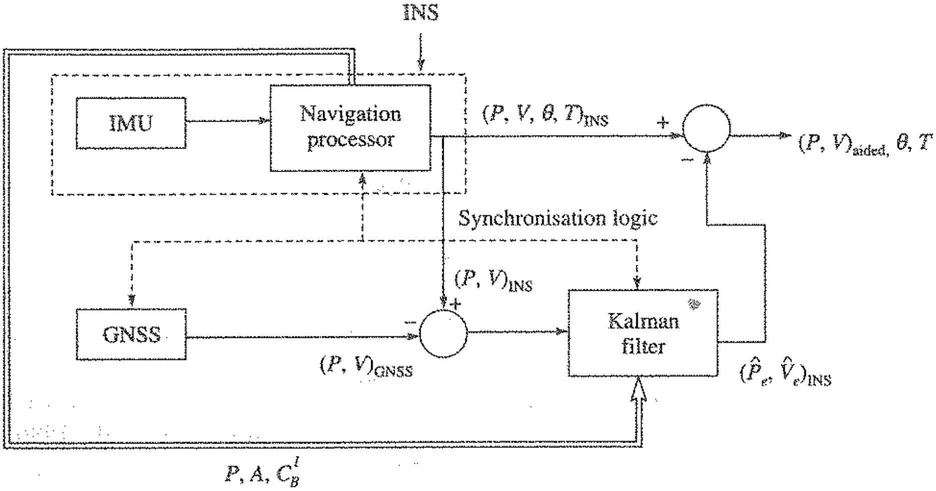
Currently, many receivers compute the navigation solution by using an Extended Kalman Filter (EKF) in the receiver itself. In such receivers, the solution error gets coloured, which is a non-white noise and that is not desirable for GNSS–INS integration using Kalman filter. Hence iterative least square solution of the nonlinear range and range rate equations [**Volk et al., 2007**] is desired for the receivers intended to be used for such INS–GNSS fusion.

Various data paths between INS, GNSS and integration processor can be used in varying degree to get more and more accurate and robust performance under higher user dynamic conditions and for long duration of operation.

7.5.1 Uncoupled Scheme

In this configuration, INS consists of IMU and navigation processor and the INS along with GNSS, produce independent navigation solutions with no influence of one on the other. An external integration processor mechanises the integrated navigation solution with Kalman filter. It is the easiest, fastest and cheapest approach when GNSS and INS are both available, but cannot communicate with each other due to restrictions imposed in such communication. The uncoupled scheme is shown in Figure 7.3.

The characteristic feature of this scheme shows that INS and GNSS are independent systems where the Kalman filter computes INS position and velocity error estimates $(\hat{P}_e, \hat{V}_e)_{INS}$, which in turn corrects the time growing INS navigation parameters and the resultant output is called $(P,V)_{\text{aided}}$. Data synchronisation needs to be carried out in the Kalman filter implementation processor prior to mixing INS and GNSS data in order that the error estimates are proper. If INS processor, GNSS receiver processor and the integration processor are all equipped with MIL–1553B interface, then synchronisation can be achieved with a typical residual synchronisation uncertainty of < 100 μs. Further, inertial acceleration (A_i), which means specific force in inertial frame with gravity correction, is required for this data synchronisation.



P: Position, *V*: Velocity, *θ*: Attitude, *T*: Time, *A*: Inertial acceleration; C_B^I: Body to inertial frame transformation matrix, (P_e, V_e): INS position and velocity error estimates respectively.

Figure 7.3 Uncoupled integration scheme.

The Kalman filter algorithm, data synchronisation logic and formation of measurements for Kalman filter are executed in the integration processor. It is important to note that the computed position, specific force transformed to inertial frame and the transformation matrix C_B^I are passed on to the Kalman filter to compute error dynamic matrix A_k and state transition matrix Φ_k in each cycle.

7.5.2 Loosely Coupled Scheme

In this scheme, as shown in Figure 7.4, the estimated INS error states can be fed back to INS for correcting its position and velocity solution. Hence INS error in position and velocity are

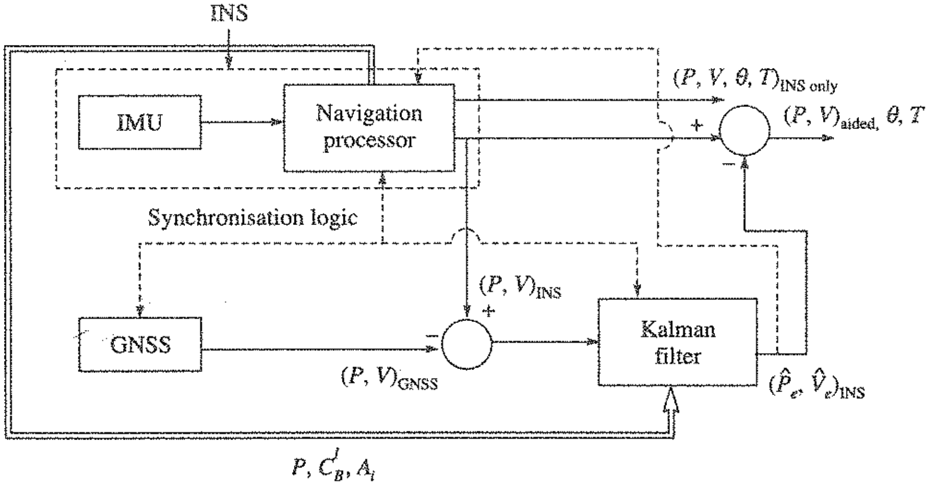


Figure 7.4 Loosely coupled integration block schematic.

reset in a continuous manner, which makes the INS error dynamics in the integration filter more and more linear due to the fact that the INS error is always kept in a small range. Thus the accuracy of integration is enhanced for long duration of operation.

Sensor bias, drift and misalignments, which may be the outputs of the integration Kalman filter, can also be fed back to INS for sensor error and misalignment correction. This has to be done with great care since proper convergence on these parameters is required, which depends on convergence time and observability issues. Quite often, requirement stipulates to retain the 'INS only' solution from INS even when the above said feedback is effected. In this case, INS processor has to keep two parallel paths one for INS only solution and the other aided solution after effecting the feedback from the integration Kalman filter.

7.5.3 Tightly Coupled Scheme

In a tightly coupled architecture (Figure 7.5), the raw measurements from the GNSS namely the pseudorange vector $\tilde{\rho}$ and pseudorange rate vector $\dot{\rho}$ and the raw measurements from the INS, namely, incremental velocity vector Δv and the incremental angle vector $\Delta\theta$ are combined in one integration processor that mechanises an appropriate single high-order integration filter. In this scheme, a common clock is preferred for INS and GNSS hardware, which implicitly avoids the problem of time synchronisation of the two systems.

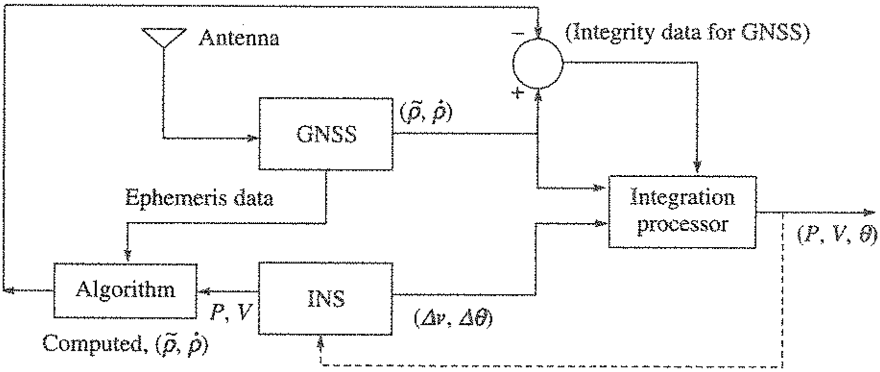

Antenna
(Integrity data for GNSS)
−
+
GNSS
$(\tilde{\rho}, \dot{\rho})$
Ephemeris data
Integration processor
(P, V, θ)
Algorithm
INS
P, V
$(\Delta v, \Delta\theta)$
Computed, $(\tilde{\rho}, \dot{\rho})$

Figure 7.5 Block schematic of tight coupling.

The higher order integration filter is usually formulated in total states rather than in error states [**Parkinson et al., 1996**], which results in an Extended Kalman Filter (EKF) implementation. Acceleration measurements (Δv) from INS, are also used as states in the integration EKF and hence the user dynamics modelling required for the EKF becomes more accurate particularly for high dynamic users. Usually receivers, such as GPS, output navigation solution nominally at 1s interval. However, internally most of the receivers have the raw measurements [$\tilde{\rho}$, ($\dot{\rho}$)] typically every 20 ms (data bit period) and GPS navigation output is usually computed using smoothed ($\tilde{\rho}$), ($\dot{\rho}$) over several future and past samples centred at the navigation solution instant. Due to the above said smoothing involved in a GPS receiver output, the standalone GPS solution becomes less accurate under fast varying acceleration (high jerk conditions) profiles of the user dynamics.

In the case of tight integration, the ($\tilde{\rho}$), ($\dot{\rho}$) data at 20 ms intervals can be directly processed with outputs from INS typically at 50 Hz rate. Since the smoothing can be avoided in the tight coupling along with the usage of acceleration data from INS at high rate, the overall integrated solution will be very accurate under large user dynamics even under jerk conditions.

Since the tightly coupled solution uses raw measurements, which are available as long as one or more satellites are being tracked, it is more robust under any GNSS outages, where less than four satellites visibility makes the receiver solution not available in the loosely coupled and uncoupled schemes, and hence affects the integrated solution considerably in those situations. It is important to include receiver clock bias and drift as states of the integration filter which facilitates user clock correction with the estimated bias and drift. This is particularly very important for long duration missions. This clock correction is important for accurate time tagging of range and range rate measurements from the GNSS also. Integrity monitoring of GNSS raw measurements can also be effectively implemented.

The ephemeris data of the GNSS satellites, being tracked, along with position, velocity data from INS can be used to check the integrity of the range and range rate of the satellites. The range and range rate of the satellites can be computed using position and velocity outputs of INS. This computed range and range rate could be compared with the measured range and range rate for integrity monitoring. The position, velocity output of the integration filter shall also be fed back to the INS to keep its P, V estimate more accurate so that its use for integrity monitoring would hold for long duration usage.

The advantages of the tightly coupled scheme are evident from the foregoing discussion. However, the system becomes more vulnerable to single point failure in either of the two systems. For critical missions, tightly coupled system is advisable with modular redundancy to address this reliability issue.

A formulation of state equations and measurement equations, required in this scheme for the Extended Kalman Filter, is described below. It is assumed that the navigation frame is ECI.

Let X^i, Y^i, Z^i = position components of ith satellite in ECI.

This can be obtained by converting the satellite position components, usually available in ECEF co-ordinate, to ECI co-ordinate using the ECEF to ECI transformation.

$\dot{X}^i$, $\dot{Y}^i$, $\dot{Z}^i$ = velocity components of ith satellite in ECI

b_u, $\dot{b}_u$ = bias and drift of the receiver clock respectively

$\tilde{\rho}^i$, $\dot{\rho}^i$ = pseudo range and pseudo range rate to the ith satellite

The state vector for the integration filter is defined as:

$$[x \; y \; z \; b \; \dot{x} \; \dot{y} \; \dot{z} \; \dot{b}_u \; a_x \; a_y \; a_z]^T$$

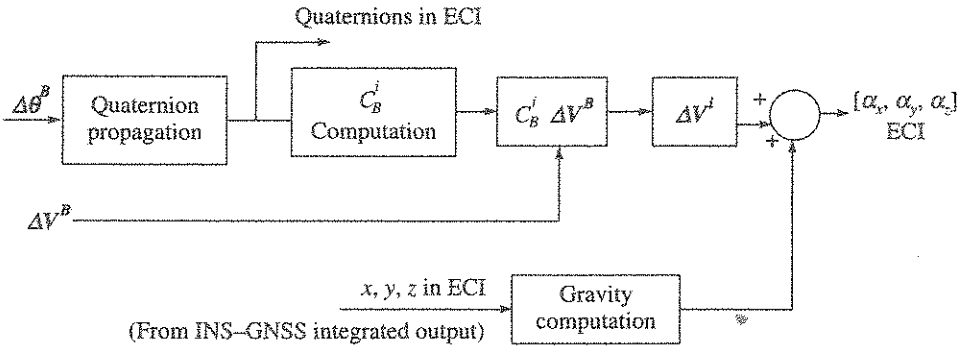
where

x, y, z = position components of user in ECI

$\dot{x}$, $\dot{y}$, $\dot{z}$ = velocity components of user in ECI

a_x, a_y, a_z = acceleration components of user in ECI

The first step, refer Figure 7.6, is to compute inertial acceleration using body measured ΔV^B, $\Delta\theta^B$ and gravity model.



i: ECI frame of reference

Figure 7.6 Computation flow diagram for a_x, a_y, a_z in ECI frame.

The process dynamics equation for the integration Kalman filter is as follows:

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \\ \dot{b}_u \\ \ddot{x} \\ \ddot{y} \\ \ddot{z} \\ \ddot{b}_u \\ \dddot{x} \\ \dddot{y} \\ \dddot{z} \end{bmatrix} = \begin{bmatrix} O_{4\times4} & I_{4\times4} & O_{4x3} \\ O_{4\times4} & O_{4\times4} & \begin{matrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{matrix} \\ O_{3\times3} & O_{3\times3} & O_{3\times3} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ b_u \\ \dot{x} \\ \dot{y} \\ \dot{z} \\ \dot{b}_u \\ \ddot{x} \\ \ddot{y} \\ \ddot{z} \end{bmatrix} \tag{7.17}$$

where

$O = n \times n$ null matrix

$I = n \times n$ Identity matrix

Measurement equations for satellite measurements, refer Chapter 6, are as follows:

$$\tilde{\rho}^{(i)} = \sqrt{\left[(X^{(i)} - x)^2 + (Y^{(i)} - y)^2 + (Z^{(i)} - z)^2\right]} + cb_u$$

$$\dot{\rho}^{(i)} = \sqrt{\left[(\dot{X}^{(i)} - \dot{x})^2 + (\dot{Y}^{(i)} - \dot{y})^2 + (\dot{Z}^{(i)} - \dot{z})^2\right]} + c\dot{b}_u$$

where $i = 1, 2, \ldots, N$; N being the maximum number of satellites visible to a receiver with atleast N or more number of channels.

Measurements from INS (ECI):

$$\ddot{x} = a_x$$
$$\ddot{y} = a_y$$
$$\ddot{z} = a_z$$

We have to combine the satellite measurements and INS measurements into one measurement matrix. For this purpose it becomes necessary to repeat the measurements from satellites, as nonlinear functions of the state.

$$\rho^i = f^i(x,\, y,\, z,\, b_u) \qquad \text{for } i = 1, 2, \ldots., N$$

$$\dot{\rho}^i = f^i(\dot{x},\, \dot{y},\, \dot{z},\, \dot{b}_u)$$

$$\ddot{x} = a_x$$

$$\ddot{y} = a_y$$

$$\ddot{z} = a_z$$

To design an EKF in the total state for the tight integration of GNSS and INS, it is required to linearise the satellite measurement equations at each of the iterations. The linearised measurement equation in compact matrix form is as follows:

$$Z = HX + v$$

where, H contains partial derivatives of the nonlinear measurement equations with respect to the states and v is measurement noise. If there are 4 visible satellites, then H is of dimension (11 × 11) and if 5 visible satellites are there, then H will be of dimension (13 × 11) matrix. Since the number of visible satellites can vary with time, the numbers of available measurements vary. Hence it is convenient to implement sequential measurement update [**Bierman, 1977**] for this case. In sequential measurement update, measurements are updated one after another until all available measurements are exhausted.

Tightly coupled integration Kalman filter can be formulated in error states also [**Gautier, 2003**].

In this case, the predicted range and Doppler using INS position, velocity and ephemeris of satellites will have to be subtracted from measured range and Doppler of GNSS. These delta range and delta Doppler will be the measurement input to the Kalman filter. Since Kalman filter, in this scheme, is formulated in error states, the error dynamics is linear, and hence EKF implementation is not required. Even though, this scheme has certain advantages, INS acceleration cannot be directly used as a state, which is an important factor in the formulation for high dynamic cases. Such aspects are needed to be considered while choosing a particular scheme.

7.5.4 Ultra Tight Coupling and Deep Integration

In addition to the tight coupling discussed in the previous section, the INS velocity solution may be fed back to the GNSS code and carrier tracking loops (carrier loop aiding) to reduce the apparent dynamic range of these loops. This has following effects:

(a) A fixed bandwidth aided tracking loop can maintain lock on GNSS signals in the presence of high user dynamics that would cause the unaided receiver to break lock.

(b) The tracking loop bandwidths can be reduced to the minimum amount required to track the errors in the INS aiding signals.

(c) The net result of the above is that the INS aided GNSS receiver can maintain lock and provide GNSS measurements over a much wider range of vehicle dynamics and radio frequency interference than the unaided standalone receiver.

It is important that the time delay between sensing of inertial velocity and its receipt at the GNSS carrier loop should be less than few tens of microseconds for the loops to be robust, and this would require dedicated hardware data paths between INS and GNSS which is realisable in an embedded design implementation of the tight coupling.

Some authors **[Parkinson et al., 1996]** have represented this ultra tight coupling scheme, which is the coupling of INS data directly with tracking loops of GNSS, as part of tight coupling. This interpretation is desirable to be highlighted under ultra tight coupling or deep integration.

7.5.5 Integrated Performance

Over the last two decades, large numbers of systems have been realised exploiting the benefits of SNS–INS integration. These integrations with SNS involved high performance INS to low performance INS and further extending to micro inertial systems as well. An illustration is provided on the features of a micro inertial system and substantial improved performance realised when it is integrated with SNS.

SNS integrated with MEMS based systems

Considering the current low performance of such micro inertial sensors, standalone navigation configuration is not usually attempted, instead, such inertial systems are mechanised as an IMU that interfaces with SNS receiver, Kalman filter and navigation software to compute integrated navigation output. The features and realised performance of one such system **[Ford et al., 2007]** is provided.

IMU micro inertial sensor performance (1 sigma value):

Gyro rate scale factor error	: 150 ppm
Gyro rate bias (in-run stability)	: 10°/h
Gyro angle random walk	: 0.1°/$\sqrt{\text{h}}$
Accelerometer scale factor error	: 300 ppm
Accelerometer linearity	: 500 ppm
Accelerometer bias	: 1000 μg

The GPS receiver is a dual frequency 24-channel version with provision for DGPS data reception and fusion. Mechanising geographic frame navigation, the test data of the integrated system shows the following performance:

Horizontal position error (with 20 s GPS outage)	: 2.5 m
Vertical position error (with 20 s GPS outage)	: 0.3 m
Steady state error in pitch, roll and heading	: 0.05, 0.05 and 0.2°

It is thus seen the significant performance to cost benefit of such integration. It also demonstrates the advantage of MEMS-SNS systems for large number of applications. There are other designs of micro IMU whose micro sensor performance may be even inferior (so, quite likely to be cheaper) compared to the above mentioned performance, yet the integrated IMU–SNS has demonstrated encouraging navigation performance **[Brown and Yan, 2004]**.

SUMMARY

The chapter has brought out some important aspects of modern integrated inertial navigation highlighting the merits of Kalman filter as an estimator. Use of Kalman filter for updating the error states and finally the method of correcting the navigation outputs, have been brought out. Such methodology of updating the error states will have numerous applications including calibration of sensors and systems. Modern integrated schemes involving SNS and INS exhibit diverse features, which are application dependent. Salient aspects of these schemes have been brought out explaining some of their advantages and disadvantages. Tightly coupled or ultra tightly coupled scheme has features which individual systems do not possess and is becoming the choice for aerospace high dynamics area where performance and mission reliability are of prime importance. Some design aspects involving extended Kalman filter have been discussed, which will be suitable for high dynamic application, in general. Satellite navigation aiding is becoming more important in the context of low cost navigation where low cost strapdown micro inertial systems are integrated with the SNS receivers whose cost has also come down over the years. The chapter also provides examples on the substantial enhancement of integrated performance against standalone performance of the inertial system.

REFERENCES

Bierman, Gerald J., *Factorization Methods for Discrete Sequential Estimation*, Academic Press, USA, 1977.

Brown, Alison K. and Lu, Yan., Performance Test Results of an Integrated GPS/MEMS Inertial Navigation Package, *Proceedings of ION GNSS*, Long Beach, California, September 2004.

Brown, Robert Grover and Patrick, Y.C. Hwang, *Introduction to Random Signals and Applied Kalman Filtering*, John Wiley and Sons, USA, 1997.

Ford, Tom., Hamilton, Jason., Morrision, John R. et al., MEMS Inertial on an RTK GPS Receiver: Integration Options and Test Results, Joint paper from *NovAtel* and *Honeywell Commercial Aviation Products*, USA, 2007.

Gautier, J.D., GPS/INS Generalised Evaluation Tool (GIGET) for the Design and Testing of Integrated Navigation Systems, Department of Geomatics Engineering, University of Calgary, Canada, 2003.

Kay, Steven M., *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall Signal Processing Series, New Jersey, USA, 1993.

Lewis, Frank L., *Optimal Estimation with an Introduction to Stochastic Control Theory*, John Wiley & Sons, USA, 1986.

Parkinson, Bradford W. and James, J. Spilker Jr. (Eds.), Global Positioning System: Theory and Applications, Vol I and II, Progress in Astronautics and Aeronautics, AIAA, **164**, USA, 1996.

Taub, Herbert. and Schilling, Donald L., *Principles of Communication Systems*, 2nd ed., McGraw-Hill, USA, 1986.

Volk, Charles, Lincoln, Jonathan, and Tazartest, Daniel, Northrop Grumman's Family of Fiber-Optic Based Inertial Navigation, 2007.

8

Signal Processing of Inertial Sensors

Signal is a stream of information. Normally these signals originate as the output of a sensor or a system. These signals are processed to suit the various applications for which they are to be used. These operations include, filtering, enhancement, digitisation, compression, and many other methods. Although enhancement and filtering can be carried out in analog domain; the need for digitising the signal becomes necessary because it provides scope to use more complex algorithm to improve signal-to-noise ratio (S/N), provides compression, provides emphasis and de-emphasis, all of which can be carried out easily using a digital signal processor.

Digital Signal Processing (DSP) is the mathematical operation that uses different algorithms and various other techniques which are necessary to manipulate the signals after they have been converted to a digital form. Statistics and probability play an important role in digital signal processing applications. Low level signals received from sensors are normally submerged in noise or interference. These corrupted signals can be processed, using statistical methods and probability, to remove the noise components without losing the intelligible information.

For information received from inertial sensors, such signal processing plays an important role. Inertial sensor output often needs to be interfaced to onboard computer for further processing. When the sensors provide analog output, the analog to digital conversion process employed must be very accurate to meet the overall requirements of accuracy and the dynamic range which is normally of the order of 10^7. To avoid the error in conversion as well as to provide simplification in electronics, digital rebalance sensor loop is also used, where the sensor output can be interfaced to a computer directly. For ring laser gyro, mechanical dithering about its sensitive axis is often employed. In order to eliminate the gyro signal corruption due to this dithering, digital signal processing is used, similarly in capacitive micro accelerometers, electromechanical sigma delta modulator is extensively used to improve performance through closed loop operation and also to provide direct digital output.

Thus digital signal processing is essential to process the inertial sensor data to achieve the required accuracy and stability.

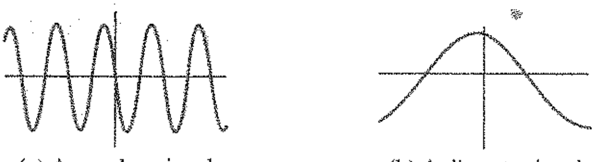
8.1 Digitisation

An analog signal, shown in Figure 8.1(a) is represented by

$$s = S(t) \tag{8.1}$$

where the argument t is continuous and S is given by $A \sin \omega t$.

If this signal is sampled and passed through an Analog to Digital Converter (ADC), the analog signal gets digitised depending upon its amplitude. This signal is referred to as discrete signal or digitised signal as shown in Figure 8.1(b).



(a) An analog signal (b) A discrete signal

Figure 8.1 Digitisation.

8.2 Sampled Data Systems

A block diagram of a typical sampled data system is shown in Figure 8.2. The signal before being digitised is passed through a signal conditioning circuit which does amplification and/or attenuation and filtering. The combination of low pass and band pass filter, removes unwanted signals which are outside the bandwidth and prevent aliasing. The signal to the ADC is sampled at a rate f_s and the digital value is presented to the DSP. The DSP carries out all the computations within the period t_s. There may be requirements that the final output signal be in digital domain or in analog domain.

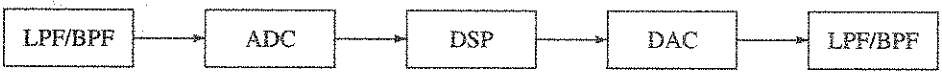


Figure 8.2 A sampled data system.

Two important parameters to be looked into for analog to digital conversion and digital to analog conversion are the effect of discrete time sampling and quantisation errors due to finite amplitude resolution. The discrete time sampling is shown in Figure 8.3. If more samples are taken the digitisation will be more accurate; but if less samples are taken, beyond a certain point,

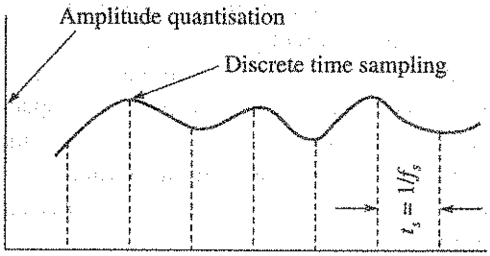


Figure 8.3 Discrete time sampling.

critical informations are lost. This limiting frequency is called the *Nyquist limit*. Nyquist's criteria can be stated as *for digitising a signal of bandwidth f_B without loss of information, the signal should be sampled at a rate $f_s > 2f_B$.*

If the sampling is carried out at a frequency less than twice the analog signal bandwidth, aliasing will occur. In Figure 8.4, the sampling frequency is slightly more than the signal frequency and does not satisfy the Nyquist's criteria and produces an aliased signal at a frequency $f_s - f_B$ and $f_s + f_B$. The signal $f_s + f_B$ seldom creates any problem. It is the component $f_s - f_B$ which creates problem when the input signal f_B exceeds the sampling frequency $f_s/2$.

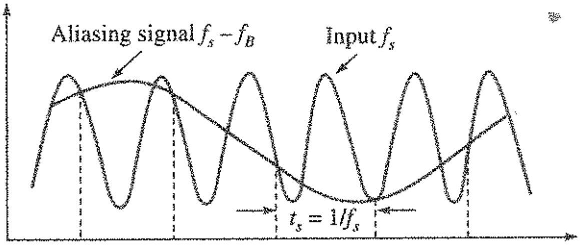


Figure 8.4 Aliasing.

The frequency domain output of the sampled signal aliases around every multiple of f_s at frequencie $\pm Nf_s \pm f_B$ for $N = 1, 2, 3, \ldots$. Example 8.1 explains the effect of aliasing.

EXAMPLE 8.1 (Nyquist frequency and Nyquist rate): If $x(t)$ is a band limited signal such that $x(\omega) = 0$ for $\omega > \omega_0$ as shown in Figure E8.1 where $\omega = 2\pi f$

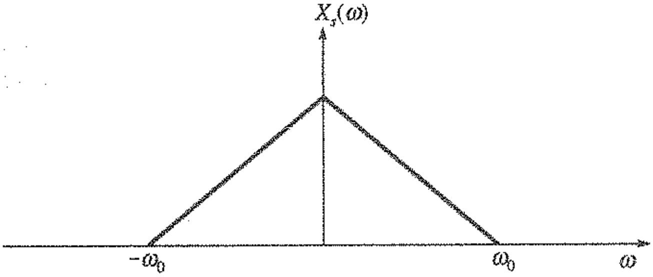


Figure E8.1 Band limited signal.

If $x(t)$ is sampled at a frequency $\omega_s \geq 2\omega_0$, the Fourier transform of $x(t)$ has the shifted spectra $X(\omega - k\omega_s)$ periodically repeating $x(\omega)$ as illustrated in Figure E8.2.

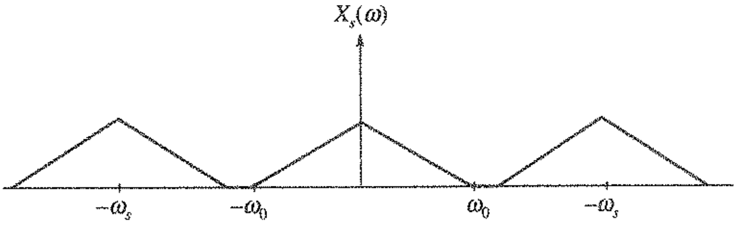


Figure E8.2 Frequency shifted spectrum.

If $\omega_s \leq 2\omega_0$, the shifted spectra overlap as shown in Figure E8.3.

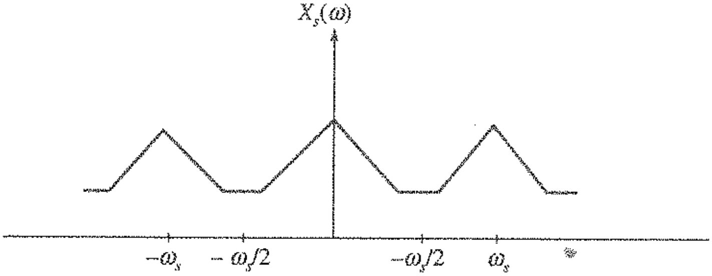


Figure E8.3 Overlapping spectra.

The overlapping of the spectral components is called aliasing. Aliasing corrupts the frequency components of $x_s(\omega)$ and dynamic range and $x(t)$ cannot be recovered from the sampled signal $x_s(\omega)$.

If $x(t)$ is a band limited signal by ensuring that the highest frequency component of $x(t)$ is less than ω_0 and by choosing a sampling frequency greater than ω_0 aliasing can be avoided.

Thus for a band limited signal such that

$x(\omega) = 0$ for $|\omega| > \omega_0$ then $x(t)$ can be recovered from the sampled signals $x(nT_s)$ if

$$\omega_s = \frac{2\pi}{T_s} \geq 2\omega_0$$

The frequency ω_0 is called the Nyquist frequency and ω_s is the Nyquist rate.

The Nyquist's bandwidth is defined as frequency spectrum from DC to $f_s/2$.The frequency spectrum is divided into an infinite number of Nyquist zones each having a bandwidth equal to $0.5f_s$ as shown in Figure E8.4. If a signal is considered, which is just outside the first frequency zone, it aliases into the frequency zone as shown in Figure E8.5. Any noise signal, appearing at any of the harmonics of f_B, will also fall within the first Nyquist zone. This necessitates the requirement of a pre-sampling filter before the ADC so that the frequencies which are outside the Nyquist band are prevented from aliasing into the bandwidth.

The performance of this filter will decide on how much attenuation the filter will provide for outside the pass band. The effect of aliasing on the dynamic range of a sampled data system is shown in Figure E8.6 and Figure E8.7.

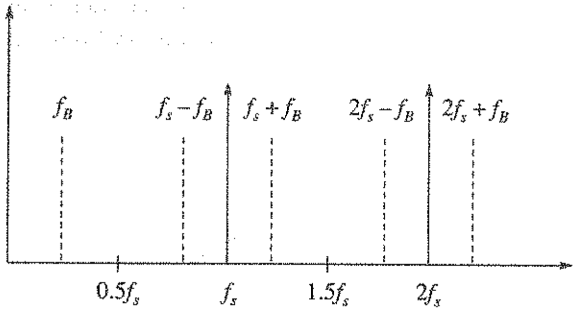


Figure E8.4 Nyquist zones.

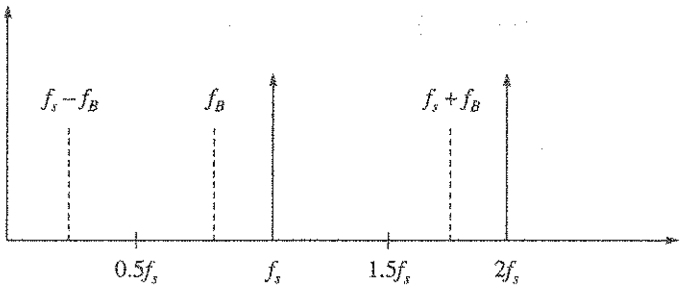


Figure E8.5 Aliasing into Nyquist zone.

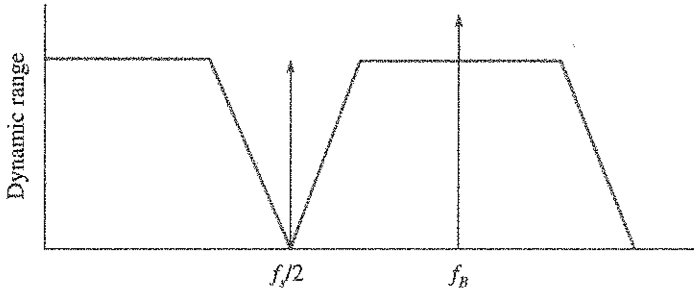


Figure E8.6 Aliasing effect on dynamic range.

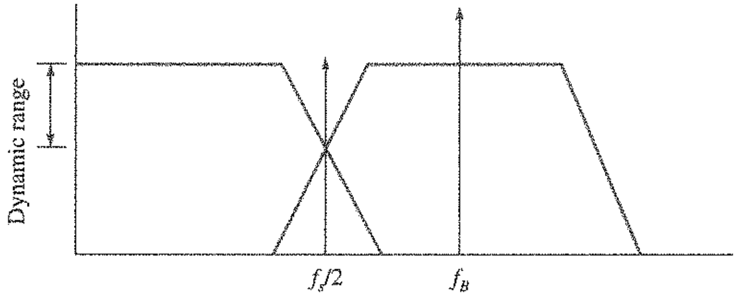


Figure E8.7 Effect of aliased signal on dynamic range.

8.3 Baseband Antialiasing Filters

Baseband antialiasing filter ensures that the signal to be sampled lies within the first Nyquist zone. Without this filter, any noise or signal, which lies outside the Nyquist band, will be aliased into the first Nyquist zone. Depending on the characteristics of the signal to be sampled, an antialiasing filter is used in all sampling ADC to remove this noise. If the signal is f_B, the antialiasing filter should pass frequencies from DC to f_B, at the same time attenuating signals above f_B to retain the desired dynamic range. The dynamic range is the ratio of the amplitude of largest signal to the amplitude of smallest signal to be resolved and is expressed in decibels (dB).

8.4 Antialiasing Filter Requirements

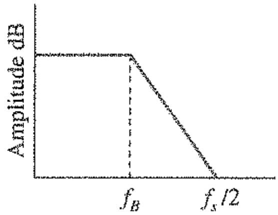


Figure 8.5 Antialiasing filter.

In order to design an antialiasing filter, first set the corner frequency equal to the bandwidth f_{BW}. Hence the pass band of the filter becomes $f_{\text{pass}} = f_{BW}$. The stop band of the filter is then set to start from $f_s/2$. Then $f_{\text{stop}} = f_s/2$, which is shown in Figure 8.5.

The attenuation of the filter is set as the dynamic range in dB. The filter should, therefore, achieve a stop band attenuation equal to the dynamic range over $\log_2 (f_{\text{stop}}/f_{\text{pass}})$ octaves. The approximate order of the filter N can be determined as the filter roll off rate is approximately $6N$ dB/octave. The above procedure does not consider the requirements of phase and ripple characteristics in pass band and stop band. In addition to this, there is natural attenuation in the pass band at higher input frequencies. This can further reduce the order of filter required.

The above discussion clearly indicates that the requirements on the antialiasing filter can be relaxed if higher sampling rates are used. The procedure to accomplish this is called oversampling.

EXAMPLE 8.2 When two different continuous time signals are sampled with a sampling frequency of 10 kHz, it produce the same sequence $x(n) = \cos(0.25n\pi)$. Find the two signals.

Solution: A continuous time signal can be represented as $x(t) = \cos(2\pi f_0 t)$ where f_0 is the signal frequency. When this signal is sampled with a frequency f_s it produces a discrete time sequence

$$x(n) = x(nT_s) = \cos\left(2\pi n \frac{f_0}{f_s}\right)$$

where T_s is the sampling time corresponding to the sampling frequency f_s.

For any integer k, the expression can be re-written as:

$$x(n) = \cos\left(2\pi n \frac{f_0}{f_s}\right) = \cos\left(2\pi n \frac{f_0 + kf_s}{f_s}\right)$$

So any sinusoidal signal with a frequency $f_0 + kf_s$ will produce the same sequence.

With $x(n) = \cos(0.25n\pi)$

$$2\pi n \frac{f_0}{f_s} = 0.25n\pi$$

The two input signals that give identical sequence are:

$$f_{01} = 0.125 f_s$$

$$f_{01} = 1250 \text{ Hz}$$

$$f_{02} = f_{01} + kf_s = 21250 \text{ Hz}$$

8.5 Oversampling and Decimation

The major advantage of oversampling is the resulting simplification of the antialiasing filter design. By oversampling, the data rate increases and the requirements on the processing circuits in real time operations becomes more demanding. An alternative way of handling these constraints is to use oversampling and decimation. If the signal is oversampled with an oversampling ratio of K, the roll of requirements of the analog filter gets relaxed by that factor since the Nyquist frequency increases to $Kf_s/2$. After sampling and digitising, a digital filter with sufficient stop band attenuation at $f_s/2$ and dynamic range is introduced as shown in Figure 8.6.

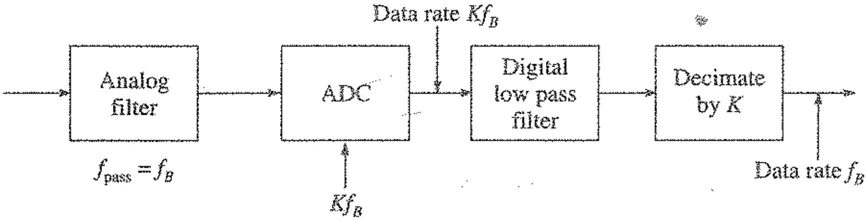


Figure 8.6 Oversampling and decimation for sampled data system.

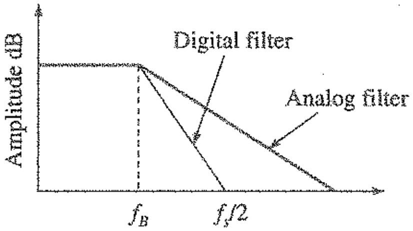


Figure 8.7 Effect of decimation.

Since the bandwidth has been reduced to $f_s/2$ by the digital antialiasing filter, the data coming out of the digital filter contains redundant information and is not necessary to consider every sample but instead it is sufficient to include only the Kth sample as shown in Figure 8.7. This procedure is called decimation. The actual decimation is carried out in the filter by computing a single output from every K input samples.

8.6 Digital Filters

Digital filters play a significant role in digital signal processing. The need for powerful computers and specialised design techniques often limited their use in various applications previously. The introduction of more powerful digital signal processing devices and fast multiplier–accumulators caused a paradigm shift from analog filters to digital even in simple applications. Compared to the analog counterpart, digital filters give a more stable characteristics, sharper roll off and repeatable performance. The digital filter works with digitised data often received from an A/D converter.

8.7 Digital Filter Basics

The digital filters are broadly classified into two categories: Finite Impulse Response (FIR) and Infinite Impulse Response (IIR) or more generally termed as nonrecursive and recursive filters. Although, filters are required to modify the response in the time domain, it is easier to understand the behaviour of a filter in frequency domain. A filter alters the spectrum of the output signal by multiplying the spectrum of the input signal by the frequency response of the filter. This multiplication, in frequency domain is convolution of the input waveform and the time domain response of the filter. In Figure 8.8, the function *H*(*f*) represents a function with unity gain for sinusoidal signal frequency ranging from 0 Hz to f_1 Hz.

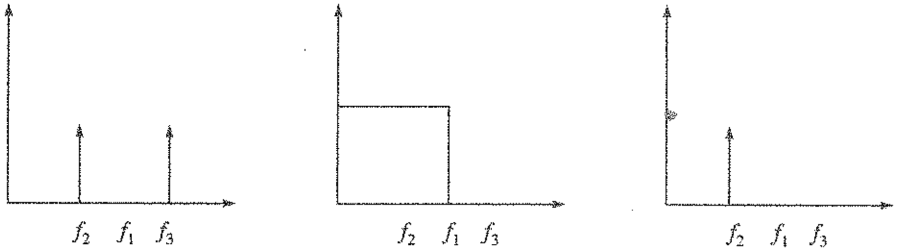


Figure 8.8 Convolution.

X(*f*) represents an input signal with spectrum consisting of sinusoidal signals with frequency f_2 and f_3 with equal amplitudes. By passing this signal through *H*(*f*), the components of frequency greater than f_1 are removed, whereas frequency less than f_1 are passed with unity gain. In other words, it is the same as the product of *H*(*f*) and *X*(*f*), and output *Y*(*f*) contains only signals with frequency components f_2 with unity gain.

For signals which are sampled data, where a function of time consists of a finite number *k* of discrete values *x*(*n*) per second and *k* is the sampling rate and *n*/*k* is the discrete variable corresponding to time. Thus a sine signal in discrete time can be represented as sin $(2\pi fn/k)$.

Fourier transform of a continuous signal maps a signal from time domain to the complex frequency domain. Similarly, inverse transfer function maps from frequency domain to time domain. The discrete Fourier transform maps signals from discrete functions of time into discrete frequency domain and the reverse.

Fourier theorem, which equates multiplication in frequency domain to convolution in time domain, provides a way to get the time response of product of two discrete time functions directly by convolution:

$$y(n) = [h * x](n) \tag{8.2}$$

is equal to the sum of the products of the signal and the frequency response

$$y(n) = \sum_{m=1}^{N} h(m) * x(n-m) \tag{8.3}$$

for all values of *n*. Equation (8.3) represents a series of multiplication and addition. The equation assumes that $h(m) = 1$ for $1 \le m \le N$ and 0 for all other values of *m*. To perform the calculation for equation 1 the values of *h*(*n*) and *x*(*n*) are required which is the inverse Fourier transform of *x*(*f*) and *h*(*f*) in Figure 8.8. The transform of *x*(*f*) will be:

$$x(n) = \sin\left(2\pi f_2 \frac{n}{k}\right) + \sin\left(2\pi f_3 \frac{n}{k}\right) \tag{8.4}$$

If f_2 and f_3 are known, $x(n)$ can be calculated for various values of n/k. The $h(n)$ value are nothing but the filter coefficients for each value of $x(n)$.

EXAMPLE 8.3 A filter is defined by the difference equation $y(n) = ay(n - 1) + bx(n) + x(n - 1)$ where a and b are real and $|a| < 1$. Find the relationship between a and b that must exist if the frequency response is to have a constant magnitude for all ω so that

$$|H(e^{j\omega})| = 1$$

Solution: The frequency response of the above system is given by the difference equation

$$H(e^{j\omega}) = \frac{b + e^{-j\omega}}{1 - ae^{-j\omega}}$$

The squared magnitude is given by

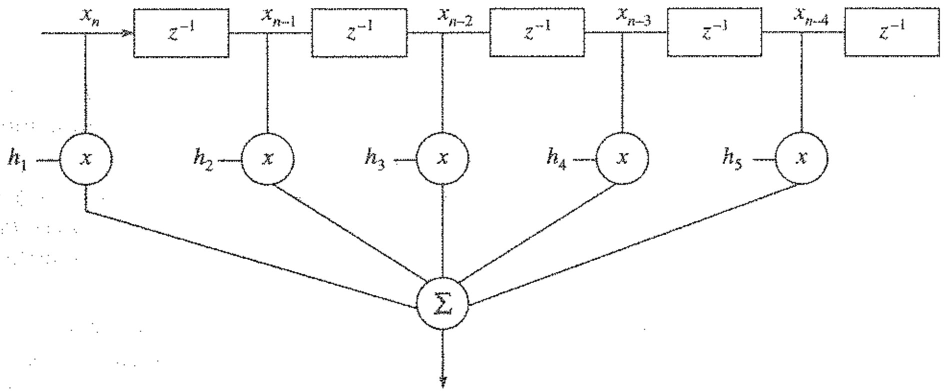
$$H(e^{j\omega}) = \frac{(b + e^{-j\omega})(b + e^{j\omega})}{(1 - ae^{-j\omega})(1 - e^{-j\omega})} = \frac{1 + b^2 + 2b\cos\omega}{1 + a^2 - 2a\cos\omega}$$

So $|H(e^{j\omega})| = 1$ when $b = -a$.

8.8 Types of Digital Filters

The FIR filters have no feedback terms. The output is a function of the finite number of the previous input values and is nonrecursive.

The FIR filters, shown in Figure 8.9, do not have any poles in their z plane transfer function. Thus their output is finite and always stable. They have linear phase response which means the phase delay between output and input increases linearly with frequency. Realisation of this type of filter is easy. The IIR filters, shown in Figure 8.10 and Figure 8.11, have recursive terms which means that the value of the output is affected by the previous values of the output $y(n)$ as well as the input values. The filter shown in Figure 8.11, is the lattice type.



$$y_n = h_1x_n + h_2x_{n-1} + h_3x_{n-2} + h_4x_{n-3} + h_5x_{n-4}$$

Figure 8.9 A FIR filter.

$$y_n = a_0 x_n + a_1 x_{n-1} + a_2 x_{n-2} - b_1 y_{n-1} - b_2 y_{n-2}$$

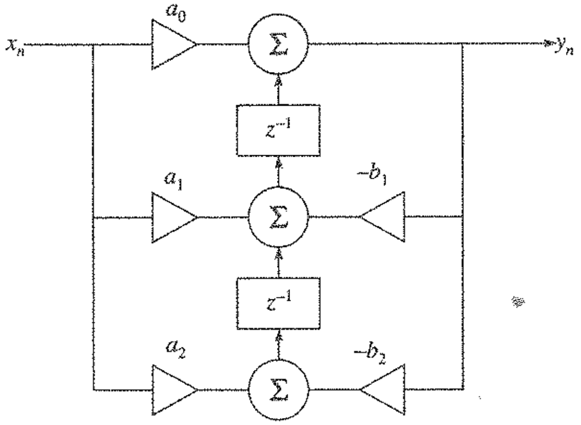


Figure 8.10 IIR filter.

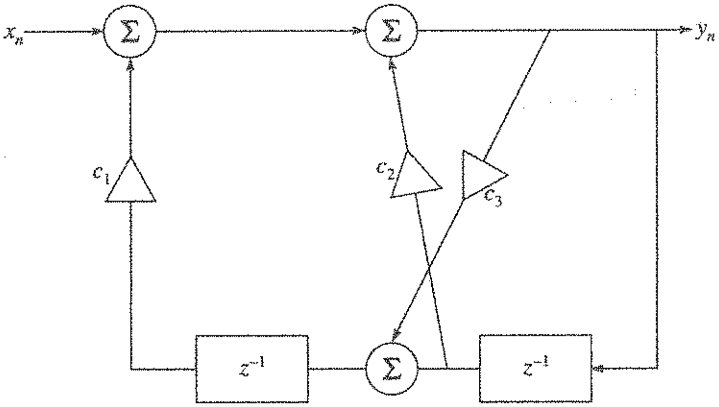


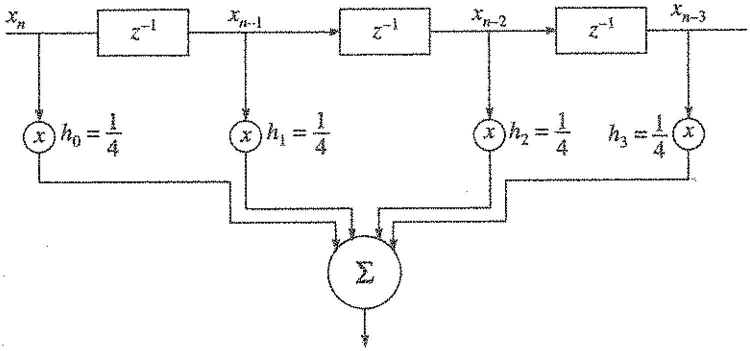
Figure 8.11 A lattice type filter.

Since FIR filters have no poles in their z plane transfer function, the output is always finite and stable. IIR filters need careful design to ensure stability. Since the FIR filters are based on discrete time delays of input variable and have no poles in their transfer function, it is possible to realise filters which do not have analog equivalents. With FIR filters it is possible to design filters with phase delay linearly increasing with frequency. This feature is very useful in applications relating to speech processing, sonar and radar. In contrast, IIR filters provide nonlinear phase characteristics with frequency. Even for continuous filters, it is difficult to achieve linear phase characteristics.

The FIR filters have low sensitivity to coefficient value change and hence less susceptible to finite word length effects. Because of their stable performance they can be easily designed as adaptive filters where coefficients can be changed in real time. IIR filters need lesser coefficients less computationally intensive and less memory space for storage. Lattice filters are more stable and requires lesser coefficients compared to equivalent FIR.

8.9 Designing of Filters

The most elementary form of FIR filter is the moving average filter. They are used in smoothing data. The input samples $x(n)$ are passed through a series of delay elements. Each sample is multiplied by a coefficient decided by the number of delay elements and finally added to get the output $y(n)$ as shown in Figure 8.12. The step response of this filter does not show any overshoot. Hence this type of filter is useful for applications where the shape of the input needs to be preserved. The rise time and fall time of the pulse is equal to the total number of taps multiplied by the sampling time. The response of a moving average FIR filter can be significantly improved by choosing coefficients with different values. The roll off characteristics can be altered by increasing the number of tap points. This is shown in Figure 8.13. The basis



$$y_n = h_0 x_n + h_1 x_{n-1} + h_2 x_{n-2} + h_3 x_{n-3}$$

$$= \frac{1}{4}(x_n + x_{n-1} + x_{n-2} + x_{n-3})$$

$$y_n = \frac{1}{N} \sum_{k=0}^{N-1} x(n-k) \quad \text{(For } N \text{ point averaging)}$$

Figure 8.12 A four-point moving average filter.

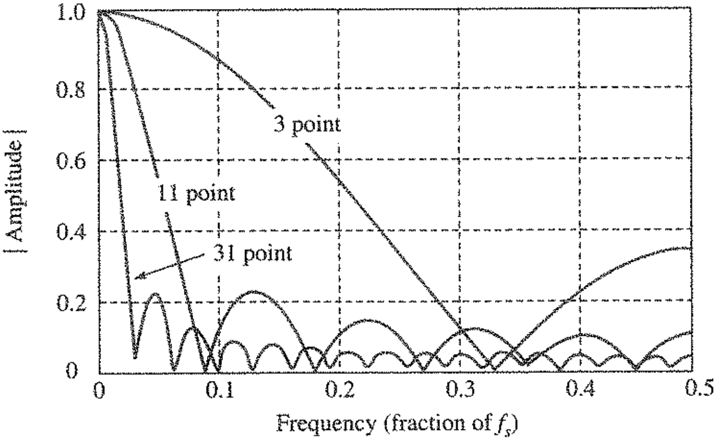


Figure 8.13 Frequency response of moving average filter.

for the design of an FIR filter is that the coefficients $h(n)$ of the FIR filter are the same as the quantised values of the impulse response of the frequency transfer function $H(f)$. For FIR filters, the requirements to be specified for design are similar to that of the analog filters–like defining the maximum amount of ripple allowed in the stop band and pass band, maximum amount of attenuation in the stop band and roll off characteristics.

The parameters for design are the number of taps N, the pass band cut off frequency f_p, the stop band cut off frequency f_s and the ratio of ripple in the pass band to the ripple in stop band (ρ_1/ρ_2).

8.9.1 Design of an FIR Filter

For any filter, the frequency response is determined by the impulse response. The quantised impulse response and the FIR filter coefficients are same. If the input to an FIR filter is an impulse as the impulse propagates through the filter, the output of each delay line is the same as the coefficient of the filter. So the procedure for designing the FIR filter is to set the desired frequency response and find the impulse response of the filter. The quantised impulse response gives the filter coefficients. This is shown in Figure 8.14. The procedure for finding convolution in a sampled data system is by series of multiplication and accumulating. The convolution operation in time domain is point by point multiplication in frequency domain and conversely convolution in frequency domain is point by point multiplication in time domain as shown in Figure 8.15. Filtering operation in frequency domain can be achieved in time domain by multiplying each frequency component by a 1 in pass band and by a 0 in stop band. The transfer function in frequency domain either 1 or 0 can be transformed into impulse response in the time domain by Fast Fourier transform. Since multiplication in frequency domain is convolution in time domain, the filtering can be carried out by convolving the sampled signal with the impulse response. Generally, various types of transforms are used to design filters. For continuous signals, the Laplace transform is used, whereas for the sampled data systems the Z-transform is used.

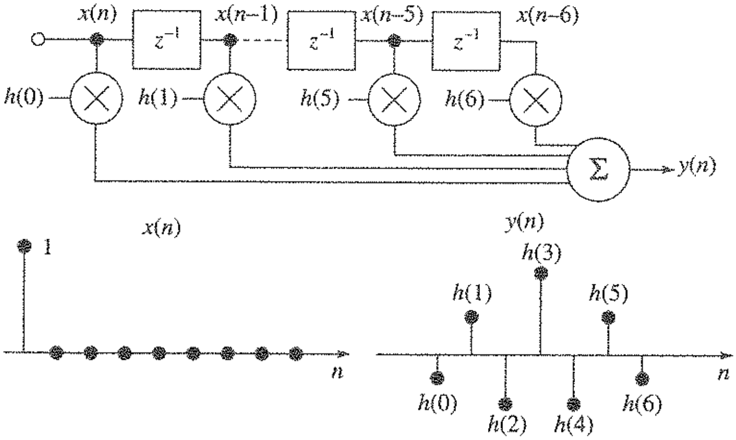


Figure 8.14 Coefficients of FIR filter from impulse response.

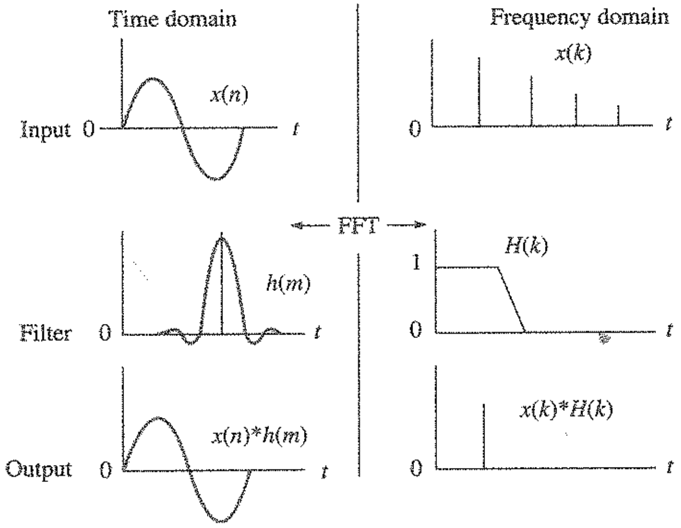
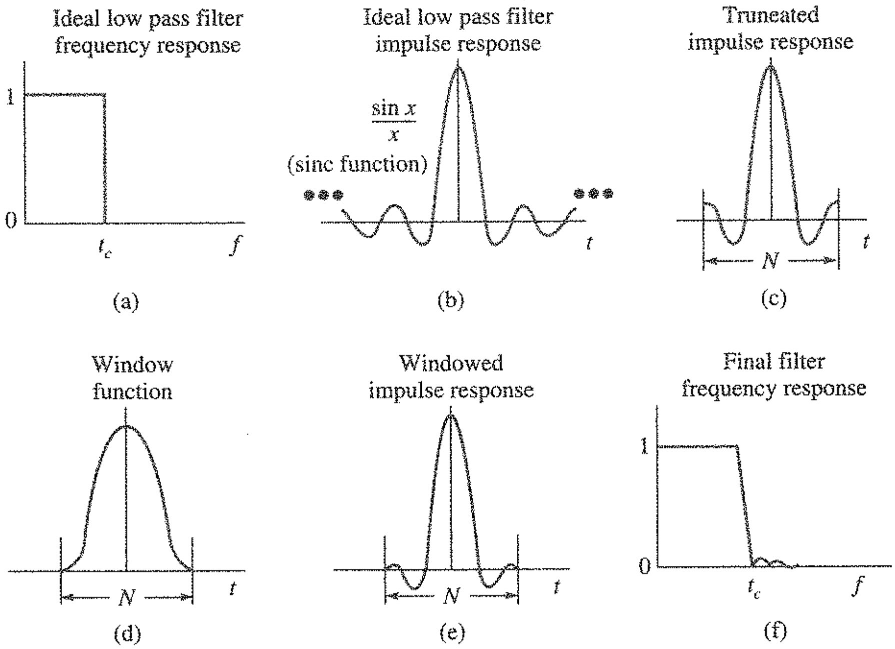


Figure 8.15 Duality of frequency domain and time domain.

8.9.2 Windowing Method for FIR Filter Design

An idealised low pass filter is shown in Figure 8.16(a). The impulse response of this filter in time domain is shown in Figure 8.16(b) and is the sin $(x)/x$ or the sinc function. If this filter is to be realised, an infinite number of taps are required.



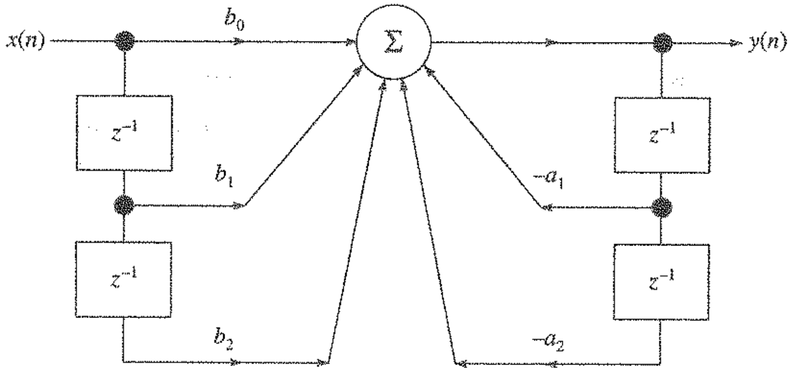
Ideal low pass filter frequency response
1
0
t_c
f
(a)
Ideal low pass filter impulse response
$\frac{\sin x}{x}$
(sinc function)
t
(b)
Truncated impulse response
N
t
(c)
Window function
N
t
(d)
Windowed impulse response
N
t
(e)
Final filter frequency response
1
0
t_c
f
(f)

Figure 8.16 FIR filter design using windowing method.

If the number of taps are limited to a finite number, the truncated response in Figure 8.16(c) is realised. This has truncated side lobes as seen in Figure 8.16(d). By choosing a proper window function, it is possible to smoothen the side lobes so that the end points of the side lobes are from zero as shown in Figure 8.16(e). The choice of the window function decides the roll off characteristics and side lobe performance of the filter. The overall response of the filter is shown in Figure 8.16(f).

8.9.3 IIR Filters

The FIR filters have transfer function without any poles. They do not have an analog equivalent. For IIR filters, they have the analog counterparts like Butterworth filter, Chebychev, elliptic and Bessel filters. They are recursive filters, and they have impulse response extending to infinite time. They do not have linear phase with frequency like FIR. They are realised as cascaded biquad structures consisting of two poles, using the quadratic equation in z domain roll off characteristics. The zeros of the transfer function are formed by the coefficients b_k and poles by a_k. The structure of an IIR filter is as shown in Figure 8.17.



- $y(n) = b_0 x(n) + b_1 x(n-1) + b_2 x(n-2) - a_1 y(n-1) - a_2 y(n-2)$
- $y(n) = \sum_{k=0}^{M} b_k x(n-k) - \sum_{k=1}^{N} a_k y(n-k)$
- $H(z) = \dfrac{\sum_{k=0}^{M} b_k z^{-k} \quad \text{(Zeros)}}{1 + \sum_{k=1}^{N} a_k z^{-k} \quad \text{(Zeros)}}$

Figure 8.17 IIR filter.

The most popular method of designing an IIR filter is to first design an analog filter meeting the pass band and stop band design requirements. The transfer function $H(s)$ is then transformed into z domain. Multiple poles design is implemented as cascaded biquad. The IIR filters are more efficient than the FIR. They can be unstable.

8.9.4 Design of IIR Filters from Analog Filters

Since IIR filters have their equivalent in analog filters, it is easier to realize IIR filters by suitable transformation from analog domain. The mapping of the transfer function from s domain to z domain can be written as:

$$H(z) = H(s)|\ s = D(z) \tag{8.5}$$

where $D(z)$ is a mapping function from s to z domain. This mapping should satisfy the following conditions to ensure acceptable performance and stability.

(a) The mapping from the $j\omega$ axis to the $z = 1$ unit circle should be one-to-one to maintain the frequency response characteristics.

(b) Points in the left half plane of the $j\omega$ axis should map inside the unit circle $z = 1$ to maintain stability.

(c) The mapping function $D(z)$ should be rational to maintain the rational nature of $H(s)$ and $H(z)$.

8.9.5 Impulse Invariance Method

From the sampling theorem, the frequency response of a digital filter is related to the analog filter as:

$$H(e^{j\omega}) = \frac{1}{T_s}\sum_{k=-\infty}^{\infty} H_a\left(\frac{j\omega}{T_s} + \frac{j2\pi k}{T_s}\right) \tag{8.6}$$

The mapping between the z plane and s plane is shown in Figure 8.18.

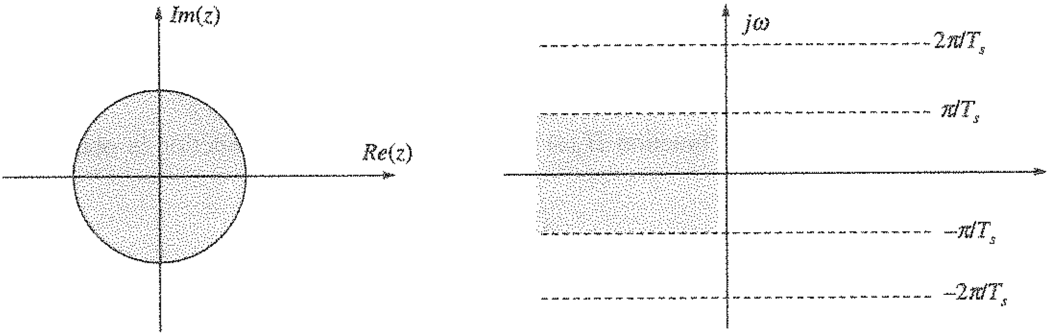


Figure 8.18 Mapping between z and s plane.

Although the $j\omega$ axis is mapped into the unit circle, it is not one to one. Each interval of frequency $2\pi/T_s$ on the $j\omega$ axis is also mapped into the unit circle. This is due to the aliasing effect. Each strip of width $2\pi/T_s$ of the left half of the s plane is mapped inside the unit circle.

If the frequency response of the analog filter is band limited, then Eq. (8.6) can be written as:

$$H(e^{j\omega}) = \frac{1}{T_s} H_a\left(\frac{j\omega}{T_s}\right) \tag{8.7}$$

EXAMPLE 8.4 The following example shows how the poles of an analog filter gets mapped to the poles of a digital filter by impulse invariance method.

Solution: Let the analog filter is represented as

$$H_a(s) = \sum_{p=1}^{k} \frac{H_p}{(s - s_p)}$$

The impulse response of this is given by

$$H_a(t) = \sum_{p=1}^{k} H_p e^{s_p t} u(t)$$

The digital filter that is formed by impulse invariance is

$$H_d(n) = H_a(nT_s) = \sum_{p=1}^{k} H_p e^{s_p\, nT_s} u(nT_s) = \sum_{p=1}^{k} H_p (e^{s_p T_s})^n u(n)$$

So the equivalent transfer function of the digital filter is

$$H(z) = H_a(s) = \sum_{p=1}^{k} \frac{H_p}{(1 - e^{s_p T_s} z^{-1})}$$

Thus a pole at $s = s_p$ of the analog filter is mapped to $z = e^{s_p T_s}$ in the digital filter.

8.9.6 Bilinear Transformation

Bilinear transformation is used to map the s domain to z domain. This is given by

$$s = \frac{2(1 - z^{-1})}{T_s(1 + z^{1})} \tag{8.8}$$

Given an analog filter transfer function $H_a(s)$ the digital filter transfer function is obtained by substituting as shown below:

$$H(z) = H_a \left[\frac{2(1 - z^{-1})}{T_s(1 + z^{1})} \right] \tag{8.9}$$

The bilinear transformation maps the left half of the s plane inside the unit circle and the $j\omega$ axis on to the unit circle. The mapping of the $j\omega$ axis to the unit circle is highly nonlinear and this effect is called frequency warping.

$$\omega = 2 \arctan\left(\frac{\omega_a T_s}{2} \right) \quad \text{where } \omega_a \text{ is the frequency in the } s \text{ domain} \tag{8.10}$$

EXAMPLE 8.5 A low pass digital filter with a –3 dB cut off frequency of $\omega_c = 0.5\pi$ is to be designed using a first order Butterworth analog filter with transfer function

$$H_a(s) = \frac{1}{1 + s/\omega_{ca}}$$

where ω_{ca} is the cut off frequency of the analog filter.

Solution: Using the inverse of the warping function, equivalent ω_{ca} for the desired cut off frequency of the digital filter is given by

For $T_s = 4$

$$\omega_{ca} = \frac{1}{T_s}\tan\left(\frac{0.5\pi}{2}\right) = \frac{1}{T_s\sqrt{2}}$$

For this revised value of ω_{ca} the Butterworth filter transfer function is estimated, and using bilinear transform, the z domain transfer function is calculated.

$$H_a(s) = \frac{1}{1 + T_s s/\sqrt{2}}$$

The z domain equivalent can be got by substituting for

$$s = \left(\frac{2(1 - z^{-1})}{T_s(1 + z^{1})}\right)$$

and

$$H(z) = \frac{1}{1 + \sqrt{2}\,\dfrac{(1 - z^{-1})}{(1 + z^{1})}} = \frac{0.414(1 + z^{-1})}{1 - 0.1716\,z^{-1}}$$

8.10 Analog to Digital Converters

The most popular ADCs are based on four generally used architectures. They are the successive approximation, sigma delta, flash and voltage to frequency conversion.

8.10.1 Successive Approximation ADC

The successive approximation ADC have been extensively used in signal processing application for the last few decades. It has been the mainstay of signal processing for many years. Newer device technology and design methods have extended both resolutions as well as conversion speed of these devices. Figure 8.19 gives the schematic of a successive approximation ADC.

The analog input is sampled and held. The successive approximation register sets the MSB to 1 and the analog voltage output of the D/A which will be half the full range is compared with the S/H output. Now, the comparator can have two values either high or low. If the comparator output is high, the value set by the SAR is retained and the next lower MSB value is set. If the comparator output is low, the MSB value is set to zero and the next lower MSB is set and the process continues till the conversion is completed. Thus, for an eight-bit A/D converter, eight comparisons are sufficient to get the digital value. The necessary timings and control signals are generated by the timing and control circuits.

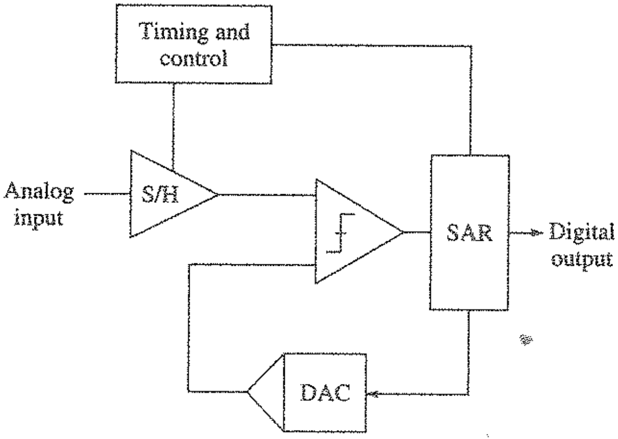


Figure 8.19 Successive approximation ADC.

The overall accuracy of the ADC is determined by the accuracy of the D/A converter. For better accuracy, laser trimming the individual resistors of R–2R ladder network is required and the process cost becomes prohibitive. For this reason, the switched capacitor or charge redistribution DAC have become popular. The advantage of this type of DAC is that the accuracy of the capacitors is determined by the plate area of the capacitors and does not require precise trimming of values. Moreover, the capacitors can be easily trimmed if required by adding parallel capacitors. The schematic of a switched capacitor DAC is shown in Figure 8.20.

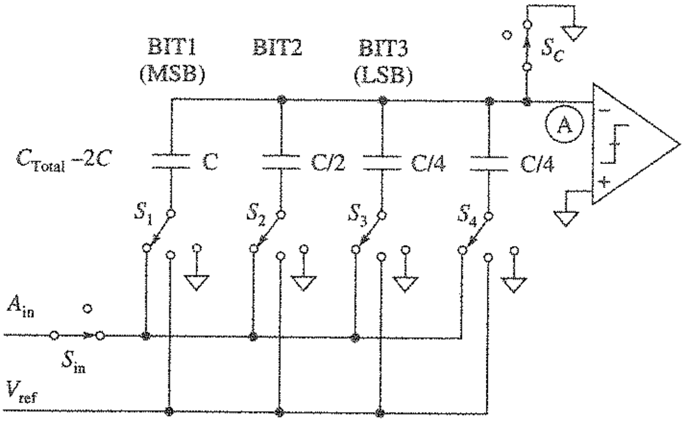


Figure 8.20 A switched capacitor DAC.

The timing details of a successive approximation ADC is given in Figure 8.21. The ADC conversion is initiated by a conversion start signal. The busy signal or end of conversion indicates whether the conversion is in progress or completed. The end of conversion which is a low going signal can be used to read the digital data.

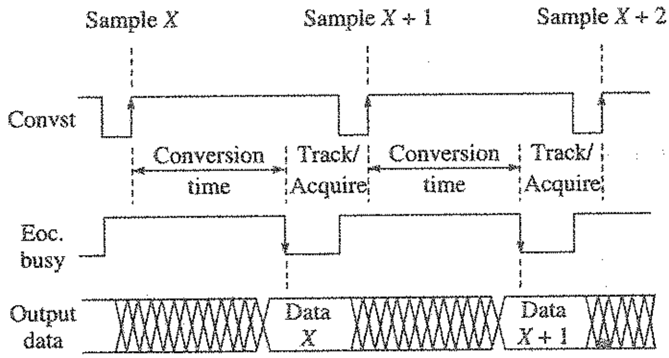


Figure 8.21 Timing details of successive approximation ADC.

8.10.2 Sigma Delta ADC

The sigma delta, symbolically shown as ΣΔ ADC is used in applications where low cost, low power, low bandwidth, high resolution ADC is required. This ADC contains a comparator, reference voltage, a switch, integrators and analog summing amplifiers and computational modules which does the function of low pass filters. The important functions, relating to the functioning of a ΣΔ ADC, are the oversampling, quantisation noise shaping digital filtering and decimation. A normal sampling ADC has a quantisation noise given by $q/\sqrt{12}$ uniformly distributed over the Nyquist band DC to $f_s/2$ where q is the LSB value and f_s is the sampling frequency as shown in Figure 8.22(a). If we use a much higher sampling rate Kf_s, the rms quantization noise $q/\sqrt{12}$ is distributed over a larger bandwidth DC to $Kf_s/2$ as shown in Figure 8.22(b).

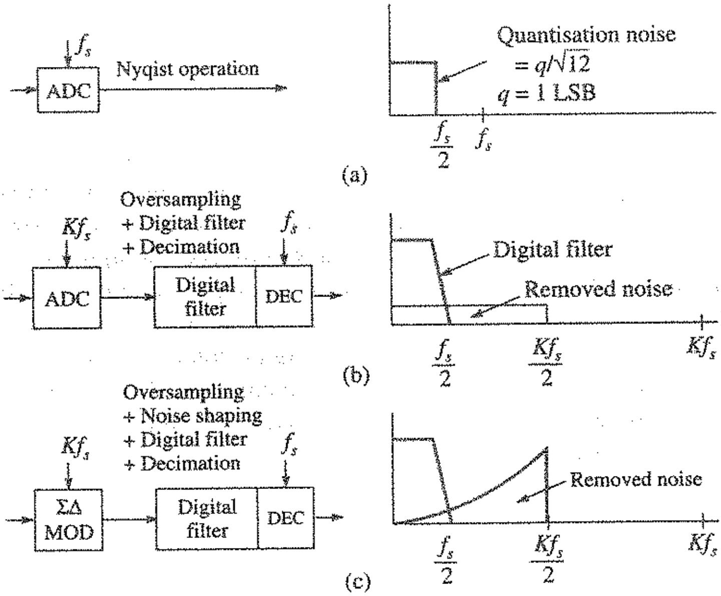


Figure 8.22 Oversampling, digital filtering, noise shaping and decimation.

If we use a digital low pass filter without affecting the wanted signal, much of the noise is removed and the effective number of bits is improved. Thus by oversampling and filtering, we can improve the resolution of a low resolution ADC. The factor K is called the oversampling factor, and it also enables in simplifying the analog anti-aliasing filter. Since the low pass filter reduces the output bandwidth, the output data rate will be less than the sampling rate Kf_s. Thus it is sufficient to take the Mth value at the output and discard the remaining. This process is known as decimation. This is shown in Figure 8.13. If the oversampling is used to improve resolution for an N bit ADC, the oversampling should be increased by a factor 2^{2N} to obtain N bit increase in resolution. But, in reality, it is not necessary to increase the sampling that much because oversampling not only limits the signal pass band but also provides noise shaping as shown in Figure 8.22(c), so that most of the noise falls outside the pass band. A first order ΣΔADC is realised by connecting 1 bit comparator, an integrator, a DAC, a digital low pass filter and decimator as shown in Figure 8.23.

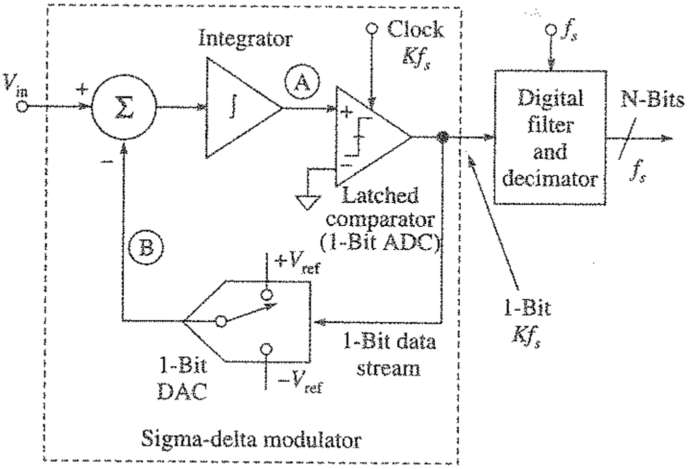


Figure 8.23 A first order sigma delta modulator.

The functioning of a sigma delta ADC is as follows. For an input voltage V_{in} applied depending on the polarity of error voltage between V_{in} and 1 bit DAC output, the integrator ramps up and down at node A. The comparator output switches high or low and is fed as a bit stream to the DAC input. The negative feedback provided through the DAC ensures that the average value of the DAC output B is equal to the input voltage V_{in}. The average DAC output is controlled by the number of ones in the bit stream. So the bit stream at the output of the comparator is the digital equivalent of V_{in}. The digital filter and the decimator process the bit stream and gives the final output. A frequency domain equivalent Circuit of the ΣΔADC is shown in Figure 8.24. Here the integrator is represented by the transfer function $H = 1/f$ where the gain rolls of at 20 dB/decade against frequency.

The output Y can be written as

$$Y = \frac{x}{f+1} + Q\frac{f}{f+1} \tag{8.11}$$

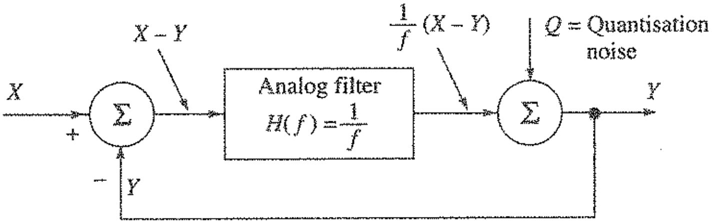


Figure 8.24 Frequency domain equivalent circuit of sigma delta ADC.

As the frequency f approaches zero, the output will be mostly contributed by the signal X and as frequency increases, the output will be contributed by noise. In effect the integrator provides a low pass characteristics for signal and high pass characteristics for quantisation noise. Thus it provides a noise shaping as shown in Figure 8.22(c). By using higher order filters, a better noise shaping and higher Effective Number Of Bits (ENOB) can be achieved.

8.10.3 Flash A/D Converters

The flash A/D converters are also referred to as the parallel ADC and use 2^{N-1} comparators where N is the number of bits. Each comparator has a reference voltage which is 1 LSB above that of the comparator below it. The individual bits identified are encoded to an N bit word by a digital encoder. The schematic of a flash ADC is shown in Figure 8.25.

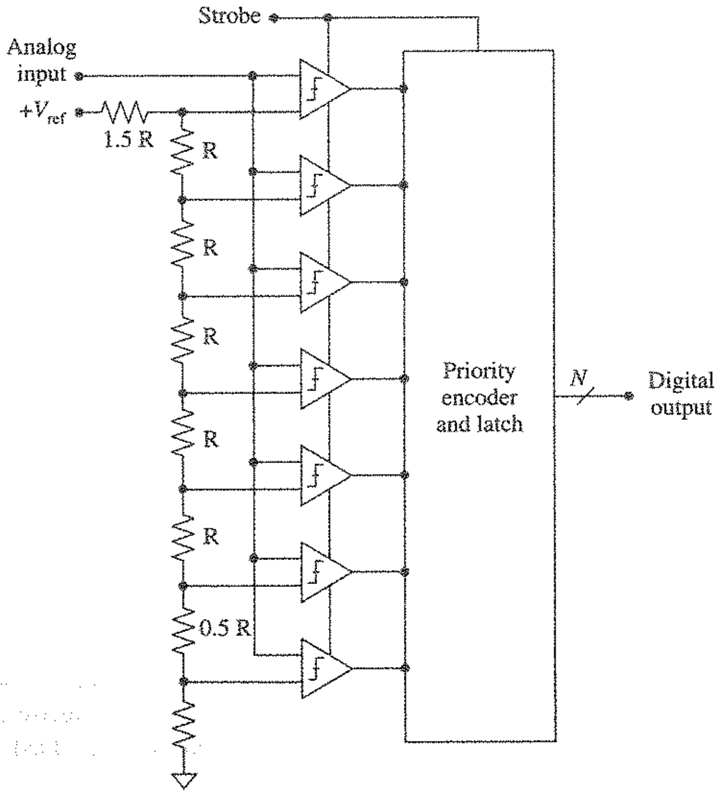


Figure 8.25 A flash ADC schematic.

The overall delay in the A/D conversion is the settling time of the comparator and the delay contributed by the digital encoder. Generally a strobe signal is applied to all the comparators simultaneously. Thus these ADC are also sampling ADC and because of the large number of comparators and the inherent delay variations between them, for the flash ADC, there will be degradation in ENOB with higher frequencies. Due to the large number of comparators, the flash ADC have higher dissipation.

8.10.4 Voltage to Frequency Converter

Another analog to digital conversion scheme which offers very high accuracy and very large dynamic range of the order of 10^7 is the Voltage to Frequency Converter (VFC). The term voltage to frequency implies that the frequency of some periodic signal is varied proportional to the analog input voltage. The output can be any periodic wave sine, square, triangle or pulse. A square or a pulse train is preferred if the output has to finally drive a counter. Figure 8.26 gives the block diagram of a typical voltage to frequency converter. The analog voltage V_{in} is applied at the input of the integrator. The output voltage V_o of the integrator is given by

$$V_o = \frac{1}{R_1 C} \int V_{in} dt \tag{8.12}$$

where $R_1 C$ is the integration time constant. The reversal of polarity is due to the amplifier being connected in the inverting configuration.

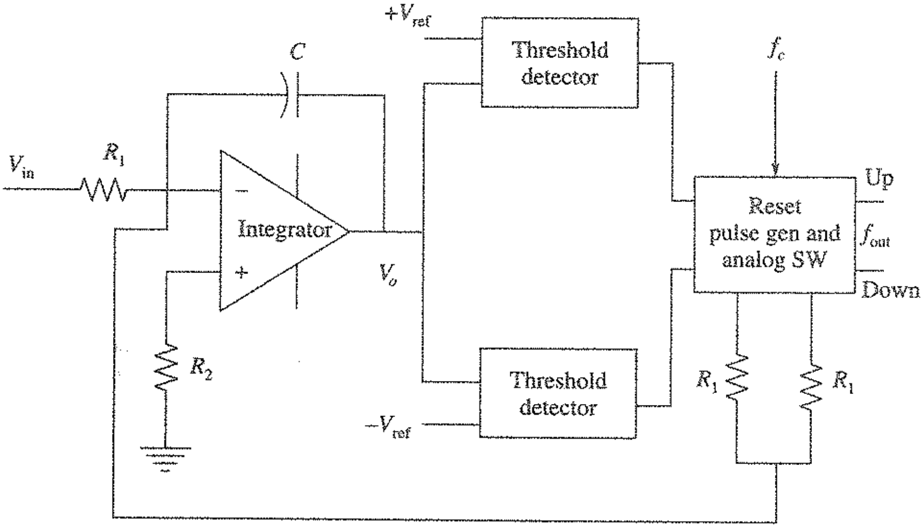


Figure 8.26 Voltage to frequency convertor.

The output voltage V_o is compared with either a positive or a negative threshold depending on the output. The output of the threshold detector triggers the reset pulse generator whenever V_o exceeds the set threshold. The pulse generator provides a pulse of precise amplitude to the input of the integrator so that a known amount of charge is removed or added to the capacitor.

A precise and stable reference source is used to generate the reset pulse with a fixed amplitude h. During reset period t_2, as shown in Figure 8.27, the integrator integrates the difference of V_{in} and the reset pulse and once the reset pulse duration is over only the input voltage is integrated. The UP/DN clock pulses are generated by the reset counter during the reset period. The overall input/output relation is given by

$$f_o = V_{\text{in}} \left(\frac{R_f * f_c}{R_1 * V_{\text{ref}}} \right) \tag{8.13}$$

where

R_f = feedback resistance

f_c = clock frequency

R_1 = input resistance

V_{ref} = reference voltage

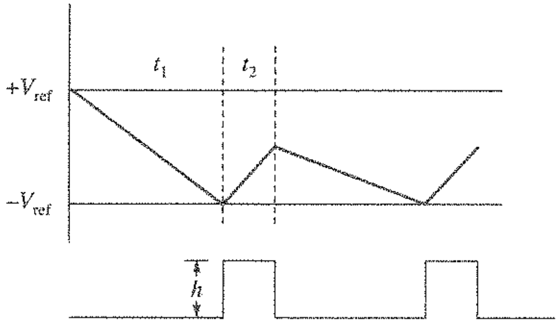


Figure 8.27 Operation of a voltage to frequency converter.

8.11 Digital Pulse Rebalance

This method provides a direct digital output by providing digital control loop for the inertial sensor. So the requirement of an A/D converter is avoided. The gyroscopically generated torque due to the input rate is balanced by a current pulsed mode control of the torque. In the case of an accelerometer, this is done by applying the current pulses to the force coil to rebalance the force generated on the proof mass due to acceleration. There are different schemes used to achieve this, which are known as binary, ternary and pulse width modulation.

In the binary scheme, pulse for required weightage is generated by suitably selecting the number of positive pulses and negative pulses with equal amplitude and duration in a sampling period. W en there are equal number of pulses, the weightage is zero, and the weightage will be positive for more positive pulses and negative for more negative pulses as shown in Figure 8.28.

In the ternary pulse rebalance loop, the current pulse are allowed to flow according to the requirement.

The current is maintained at zero for zero correction and positive and negative pulses for the positive and negative current requirements respectively. Thus the ternary loop has three states unlike the binary. This is shown in Figure 8.29.

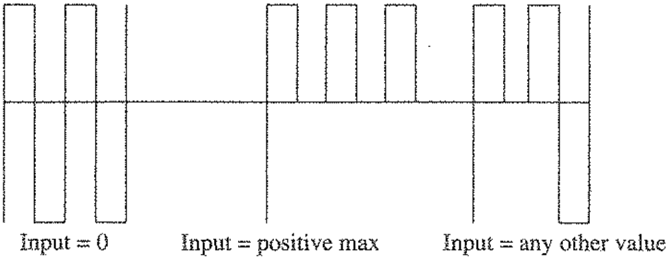


Figure 8.28 Binary pulse rebalance.

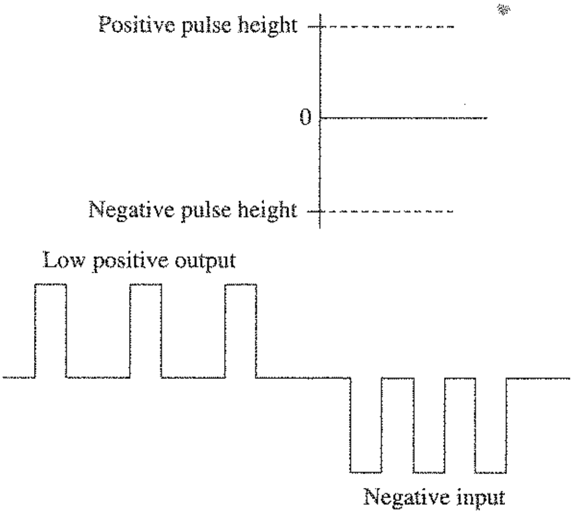


Figure 8.29 Ternary pulse rebalance.

In both cases, the pulse height and duration is constant, but their repetition rate changes with the amplitude of the correcting signal.

In the pulse width modulated binary system, the amplitude of the pulse is constant as in the binary system with fixed values for positive pulse height and negative pulse heights. The ratio of the ON/OFF period of the pulse changes so that for positive output ON period is more and for negative output the OFF period is more and 50% for zero output. This scheme is shown in Figure 8.30.

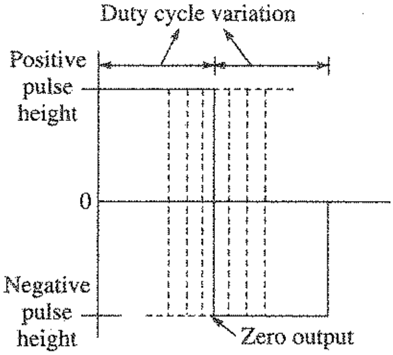


Figure 8.30 Pulse width modulated scheme.

SUMMARY

The basics of signal processing methods which are very essential for inertial systems are discussed in this chapter. This includes sampling, Nyquist rate, aliasing effects which are introduced in the beginning of the chapter. This is followed by antialiasing filter requirements and basic concepts of over sampling and decimation are discussed. The basics of digital filtering, types of digital filters and their realisation methods are also given. The concept of windowing and various windowing methods are also addressed. A brief introduction to realisation of digital filters using impulse invariance method and bilinear transformation are also discussed. Various types of analog to digital converters, voltage to frequency converters and digital rebalancing techniques are also given. A few examples as well as problems and solutions are also given.

EXERCISES

8.1 Given a continuous time signal $x_a(t)$ what is the minimum sampling rate required for

(a) $x_a(2t)$ (b) $x_a(t)\cos 3\pi Bt$ (c) $x_a^3(t)$

[**Ans:** (a) $2f_s$, (b) $2B$, (c) $3f_s$]

8.2 Determine the characteristics of the s plane to z plane mapping for the mapping function given by

$$s=\frac{1-z^{-2}}{1+z^{-2}}$$

8.3 Using bilinear transformation convert the analog filter with system function

$$H(s)=\frac{s+1}{s^2+s+25}$$

into an IIR filter using $T = 0.01$.

8.4 A digital filter has transfer function

$$H(z)=\frac{2}{1-0.35z^{-1}}+\frac{1}{1-0.25z^{-1}}$$

If this filter is designed using impulse invariant method with $T_s = 0.02$ find the equivalent transfer function of the analog filter. If this filter is designed using the bilinear transformation with $T_s = 0.02$ find the equivalent analog filter.

8.5 Design an FIR filter approximating the ideal response

$$H_d(\Omega)=\begin{Bmatrix} 1 & \text{for} & |\Omega|\le\frac{\pi}{6} \\ 0 & \text{for} & \frac{\pi}{6}<|\Omega|<\frac{\pi}{3} \\ 1 & \text{for} & \frac{\pi}{3}<|\Omega|\le\pi \end{Bmatrix}$$

(a) Determine the coefficient of a 20 tap filter using the rectangular window.

(b) Determine and plot the magnitude and phase characteristics.

Answer to Problems

8.1 (a) If $x_a(t)$ is a band limited signal and the minimum sampling frequency is f_s then the spectrum of $x_a(t)$ can be arbitrarily represented as given in Figure A.1. The signal $x_a(2t)$ represents decimation of the signal in time. This causes the spectrum to expand in frequency domain

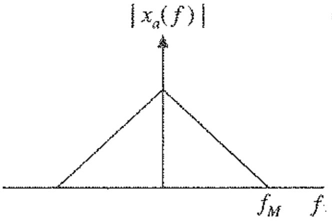


Figure A.1

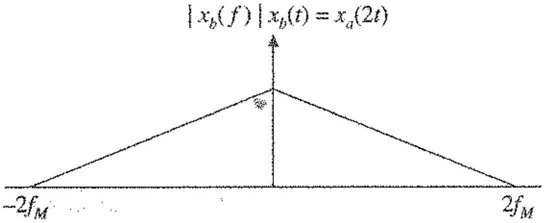


Figure A.2

The minimum sampling frequency required is

$$f_s' = 2f_s$$

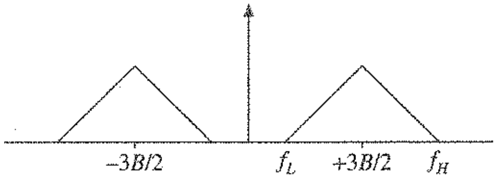
(b) Let $y(t) = x_a(t) \cos 3\pi Bt$

$$y(t) = \frac{x_a(t)}{2}(e^{j3\pi Bt} + e^{-j3\pi Bt})$$

In frequency domain this can be written as

$$Y(f) = \frac{1}{2}\left[x_a\left(\frac{f-3B}{2}\right) + x_a\left(\frac{f+3B}{2}\right)\right]$$

This can be represented as



$f_H = 3B/2 + f_M$ $f_L = 3B/2 - f_M$

since y(t) is a band pass signal the minimum sampling rate is given by

$$f_s' = \frac{2f_H}{k_{\max}} \quad \text{where } k \le \left(\frac{f_H}{f_H - f_L}\right)$$

If $f_H - f_L = B$ then $f_M = \dfrac{B}{2}$

and $$k \le \frac{\left(\frac{3B}{2} + f_M\right)}{B} = \frac{\left(\frac{3B}{2} + \frac{B}{2}\right)}{B} = 2$$

or $$k_{\max} = 2$$

$$f_s' = \frac{2f_H}{2} = f_H = \frac{3B}{2} + \frac{B}{2} = 2B$$

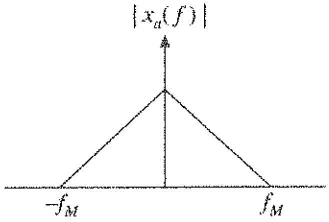
(c) $x_a^3(t)$

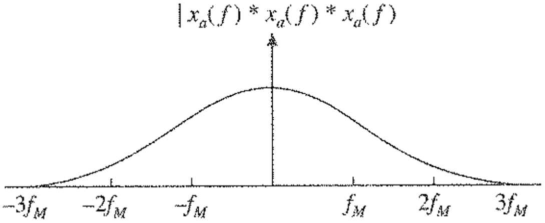
Let $$y(t) = x_a^3(t)$$

In frequency domain

$$y(f) = x_a(f) * x_a(f) * x_a(f)$$

where '*' denotes convolution operation.





Thus the signal is bandlimited to a frequency $3f_M$. Hence the minimum sampling frequency is

$$f_s' = 2(3f_M) = 3(2f_M) = 3f_s$$

where f_s is the sampling rate required for $x(t)$.

8.2 For the mapping function given by $s = \dfrac{1 - z^{-2}}{1 + z^{-2}}$

$$s + sz^{-2} = 1 - z^{-2}$$

$$z^{-2} = \frac{1 - s}{1 + s}$$

$$z^2 = \frac{1 + s}{1 - s}$$

$$z = \sqrt{\frac{1 + s}{1 - s}}$$

substituting $s = j\omega$

$$z = \sqrt{\frac{1 + j\omega}{1 - j\omega}}$$

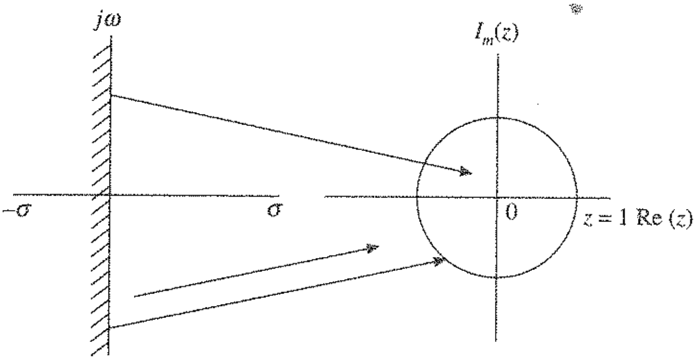
Multiplying by $1 + j\omega$ both numerator and denominator

$$z = \frac{1 + j\omega}{\sqrt{1 + \omega^2}} = \frac{1}{\sqrt{1 + \omega^2}} + \frac{j\omega}{\sqrt{1 + \omega^2}}$$

Expressing as $x + jy$

$$x = \frac{1}{\sqrt{1 + \omega^2}} \qquad y = \frac{\omega}{\sqrt{1 + \omega^2}}$$

$$x^2 + y^2 = 1$$



- The $j\omega$ axis maps on to the perimeter of the circle of unit radius in z plane.
- Left half of s plane maps into the unit circle with centre at $z = 0$.
- Right half of the s plane maps into region outside the unit circle.

8.3 $H(s) = \dfrac{s + 1}{s^2 + s + 2s}$

Using bilinear transformation for $T_s = 0.01$

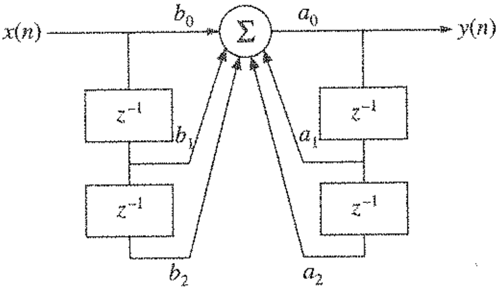
$$H(z) = H(s) \mid s = \frac{2}{T_s}\left(\frac{1 - z^{-1}}{1 + z^{-1}}\right) = 200\left(\frac{1 - z^{-1}}{1 + z^{-1}}\right)$$

$$H(z) = \frac{200\left(\dfrac{1 - z^{-1}}{1 + z^{-1}}\right) + 1}{200\left(\dfrac{1 - z^{-1}}{1 + z^{-1}}\right)^2 + 200\left(\dfrac{1 - z^{-1}}{1 + z^{-1}}\right) + 25} = \frac{(z + 1)[201z - 199]}{425z^2 - 350z + 25}$$

$$= \frac{201z^2 + 2z - 199}{425z^2 - 350z + 25}$$

The zeros are at $z_1 = 0.99$, $z_2 = -1$, Poles are at $z_3 = -0.745$, $z_4 = 0.079$

$$\frac{y(n)}{x(n)} = \frac{b_0 + b_1 z^{-1} + b_2 z^{-2}}{a_0 + a_1 z^{-1} + a_2 z^{-2}} = \frac{1 + 0.0099z^{-1} + 0.99z^{-2}}{1 - 0.823z^{-1} + 0.058z^{-2}}$$



8.4 $H(z) = \dfrac{2}{1 - 0.35z^{-1}} + \dfrac{1}{1 - 0.25z^{-1}}$, $T_d = 0.02$

From Impulse Invariance Technique

$$H(z) = \sum_{k=1}^{k=n} \frac{T_d A_k}{1 - e^{s_k T_d} z^{-1}}$$

Comparing with given $H(z)$

$$H(z) = \frac{T_d A_1}{1 - e^{s_1 T_d} z^{-1}} + \frac{T_d A_2}{1 - e^{s_2 T_d} z^{-1}}$$

$$T_d A_1 = 2 \quad \text{or} \quad A_1 = 100, \qquad T_d A_2 = 1 \quad \text{or} \quad A_2 = 50$$

$$e^{s_1 T_d} = 0.35, \quad e^{s_2 T_d} = 0.25$$

$$s_1 = \ln 0.35/T_d = -52.5, \quad s_2 = \ln 0.25/T_d = -69.3$$

The time domain response is given by

$$\begin{aligned} h(t) &= A_k \sum_{k=1}^{k=n} e^{s_k t} \\ &= A_1 e^{s_1 t} + A_2 e^{s_1 t} \\ &= 100e^{-52.5t} + 50e^{-69.3t} \end{aligned}$$

Taking Laplace transform

$$H(s) = \frac{100}{s + 52.5} + \frac{50}{s + 69.3}$$

Bilinear transformation is given by

$$s = \frac{2}{T_d}\left(\frac{1 - z^{-1}}{1 + z^{-1}}\right)$$

$$\frac{T_d s}{2} = \frac{1 - z^{-1}}{1 + z^{-1}}$$

$$sT_d + z^{-1}\, sT_d = 2 - 2z^{-1}$$

$$z^{-1} = \frac{2 - sT_d}{2 + sT_d}$$

$$z = \frac{2 + sT_d}{2 - sT_d}$$

$$H(z) = \frac{2}{1 - 0.35z^{-1}} + \frac{1}{1 - 0.25z^{-1}}$$

$$H(s) = H(z)\Bigg|\, z^{-1} = \frac{2 - sT_d}{2 + sT_d} = H(z)\Bigg|\, z^{-1} = \frac{1 - 0.01s}{1 + 0.01s}$$

$$= \frac{2}{1 - 0.35\left(\dfrac{1 - 0.01s}{1 + 0.01s}\right)} + \frac{1}{1 - 0.25\left(\dfrac{1 - 0.01s}{1 + 0.01s}\right)}$$

$$= \frac{2(1 + 0.01s)}{(0.65 + 0.135s)} + \frac{1 + 0.01s}{(0.75 + 0.0125s)}$$

8.5 Design an FIR filter approximating the ideal response.

$$H_d(\omega) = \begin{Bmatrix} 1 \text{ for } |\omega| \le \dfrac{\pi}{6} \\ 0 \text{ for } \dfrac{\pi}{6} < |\omega| < \dfrac{\pi}{3} \\ 1 \text{ for } \dfrac{\pi}{3} < |\omega| \le \pi \end{Bmatrix}$$

(a) Determine the coefficient of a 20 tap filter using the rectangular window.

(b) Determine and plot the magnitude and phase characteristics.

[**Hint:** First the equivalent impulse response is calculated using

$$h_d(n) = \int_{-\pi}^{\pi} H_d(e^{j\omega})e^{j\omega n}\, d\omega$$

which gives $\dfrac{2}{n}\left[\sin\left(\dfrac{n\pi}{6}\right) - \sin\left(\dfrac{n\pi}{3}\right)\right]$ for $n \neq 0$

The 20 tap filter is given by

$$h(n) = h_d(n) \quad n = 0, 1, 2, 3, \ldots, 19$$

$$h_d(0) = \int_{-\pi}^{\frac{\pi}{3}} d\omega + \int_{-\frac{\pi}{6}}^{\frac{\pi}{6}} d\omega + \int_{\frac{\pi}{3}}^{\pi} d\omega$$

$$h(n) = \frac{5\pi}{3} \quad \text{for } n = 0$$

$$h(n) = \frac{2}{n}\left[\sin\left(\frac{n\pi}{6}\right) - \sin\left(\frac{n\pi}{3}\right)\right] \text{ for } n = 1, 2, 3, \ldots, 19$$

The frequency response of the FIR filter is given by $H(e^{j\omega}) = \sum_{n=0}^{19} h(n)e^{-j\omega n}$.

9

Application of Navigation and Inertial Sensors

Traditionally, military market covering different types of aircrafts, missiles of different types, aircraft carriers, submarines and mobile armoured vehicles have been the dominant user of navigation systems and inertial sensors, and they still continue to do so. From military aircrafts, navigation systems had a natural penetration into civilian aircrafts with its worldwide utility. Besides such large-scale traditional users, there were some specific applications witnessed over the previous few decades in satellite launch vehicles, spacecrafts, interplanetary missions, airborne remote sensing and precision pointing of payloads.

Similarly, MEMS based inertial systems have found applications in certain aerospace sectors where usage of traditional inertial systems would have been difficult due to their size and environmental capability limitations. Application of MEMS is emerging in space in the field of relative navigation. These applications cover autonomous inertial navigation as well as integrated inertial navigation involving satellite navigation. Use of MEMS based inertial sensors in automobile is fast emerging and has been covered as it shows some interesting feature. A new type of inertial system is emerging, although extremely limited in use, covers the use of accelerometer only, no gyros, in configuring a navigation system. A few of these applications will be discussed in the subsequent sections.

9.1 Inertial Navigation System for Satellite Launch Vehicles

Inertial navigation system is in use in satellite launch vehicles for nearly four decades. A launch vehicle normally follows closed loop guided trajectory to inject the satellite into the desired orbit, such as an inclined orbit or a Geostationary Transfer Orbit (GTO) or a Sun Synchronous Polar Orbit (SSPO) or a navigation orbit. Table 9.1 shows some such spacecraft orbits.

Table 9.1 Typical Orbits for Spacecraft

Orbit type	Apogee (in km)	Perigee (in km)	Inclination (in degree)
GSO	~36,000	~36,000	0.0
GTO	~36,000	~200	~19*
SSPO	817	817	98.7

*Depends on launch latitude. GSO: Geostationary orbit.

These orbits are shown in Figures 9.1 and 9.2.

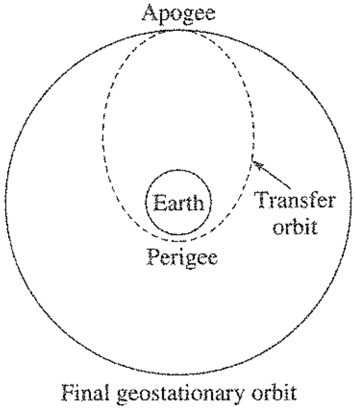


Figure 9.1 Transfer orbit and the final geostationary orbit.

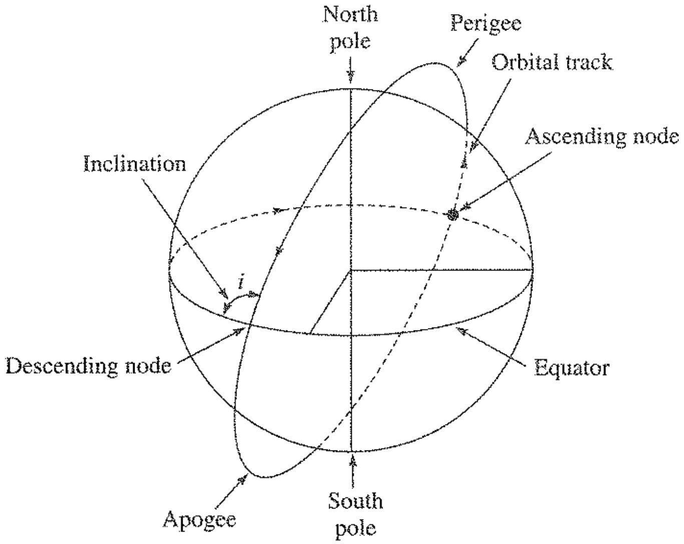


Figure 9.2 A sun synchronous polar orbit.

The closed loop guidance scheme needs information on vehicle position, velocity and attitude in a continuous manner from lift-off till satellite injection so that the guidance scheme can take appropriate action to correct the deviation of the vehicle trajectory from the desired path and finally realise the intended orbit through appropriate action to cut-off vehicle thrust.

The actual orbit deviates from the desired one where the primary error contribution is from INS in an otherwise nominal mission, which implies that the vehicle deviation is within the correctable capacity of the guidance. The INS accuracy for launch vehicle is specified in terms of errors at spacecraft injection comprising injection parameters such as altitude, velocity and flight path angle, which comprise three-dimensional specification. The errors at spacecraft injection can be further propagated to provide the corresponding error in orbital parameters, such as apogee, perigee and inclination. Currently, operational mission requirements specify these orbital errors to lie typically within 150 km in apogee, 3 km in perigee and 0.05° in inclination for a GTO mission. Mission requirements stipulate that the INS should be accurate, should have low mass and perform reliably. The need for these requirements will be explained further along with some typical procedures followed to realise these objectives. Location of INS in a satellite launch vehicle is shown in Figure 9.3.

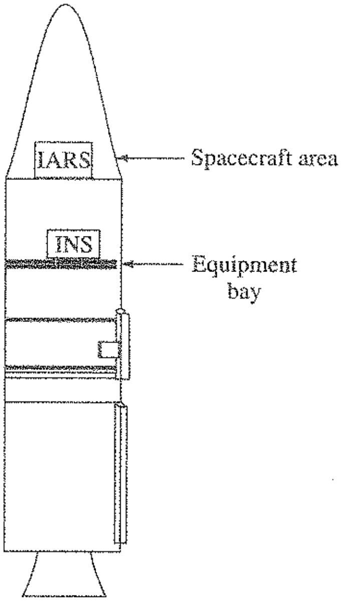


Figure 9.3 Location of INS in the satellite launch vehicle.

Spacecraft carry propulsion fuel for initial correction of the orbital errors and subsequently for maintaining the orbit. For a geosynchronous spacecraft, the spacecraft carries additional propulsion fuel for taking the spacecraft to geosynchronous orbit from geosynchronous transfer orbit; refer Figure 9.4, where the spacecraft is initially injected.

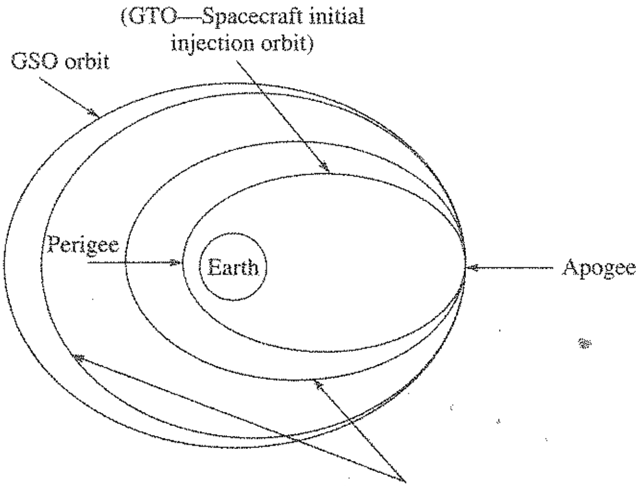

(GTO—Spacecraft initial injection orbit)
GSO orbit
Perigee
Earth
Apogee
Raised intermediate orbits

Figure 9.4 Spacecraft transfer from GTO to GSO.

With a better specified injection parameters, implying an accurate INS, correction fuel needed in spacecraft, to correct the initial orbital error, is less and the consequent saving in fuel mass can be utilised for increasing the revenue earning payload mass or for increasing the life of the spacecraft. A low weight INS similarly increases the payload mass or can allow the spacecraft to carry extra fuel for increased life where orbital life of modern spacecraft is primarily governed by the fuel mass it carries. It may be noted that the INS is normally located on the uppermost stage of the vehicle, often called equipment bay, which along with the spacecraft attains orbital velocity and gets injected into the orbit. At a point of time, the spacecraft is separated from the equipment bay by adding some velocity to the spacecraft. Thus the INS mass adds to the spacecraft mass till this separation from equipment bay. In a multi-stage vehicle, the lower stages are jettisoned as the vehicle climbs up, so the requirement arises to locate the INS at the uppermost stage. Besides the benefits to spacecraft life and payload mass, accuracy in INS, especially in initial level alignment, also provides safety to the vehicle to clear the launch tower during vertical take-off keeping sufficient clearance from the umbilical tower that is quite close. Ideally, the inertial sensors used in such missions should have zero or negligible sensitivity to acceleration and vibration as these are quite significant in these vehicles. Towards fulfilling these requirements, modern solid state gyros, like ring laser gyro or hemispherical resonator gyro or fiber optic gyro, are found increasingly suitable. When dynamically tuned gyro is used, extensive modelling, calibration and compensation are necessary to meet similar accuracy requirement, as this gyro has significant sensitivity to such environments. Accelerometers are normally macro-sized torque to balance type but use of vibrating beam accelerometers is also emerging.

INS outputs are used by guidance, control, vehicle sequencing and telemetry as shown in Figure 9.5. These are typical outputs and their use, but can change from mission to mission.

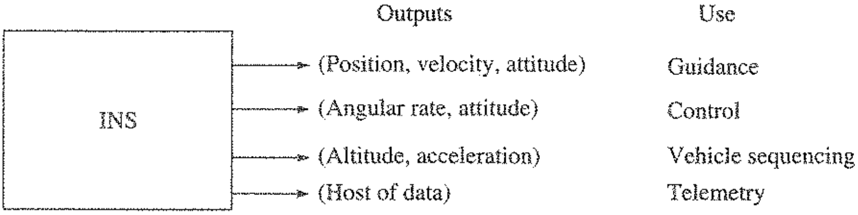


Figure 9.5 INS outputs and their use.

High reliability is needed as the cost of the mission is significantly high and that the performance is to be demonstrated under vacuum, variable temperature, high vibration and shock environment. Considering the simultaneous requirements of accuracy and reliability, a typical configuration for the INS quite often invokes in-flight sensor redundancy. This is in addition to the use of military grade electronic components for reliability enhancement. Reliability is also ensured by long hours and large number of days of pre flight-testing under simulated launch environments to take care of infant mortality failure regime. Additionally, such long duration test provides data to be loaded for sensor error compensation and statistical data set for the sensor performance dispersions, which are then further used for predicting the likely orbital errors. Such preflight prediction of orbital error is done either using Monte Carlo simulation or by covariance simulation or sometimes by both the methods. INS meeting the orbital accuracy and reliability criteria is only allowed to fly.

Initially, INS orthogonal body frame is aligned to vehicle orthogonal body frame during mounting of the INS on the equipment bay. Thereafter, the angular misalignment between the INS orthogonal body frame and the navigation frame is established. While self alignment is a normal feature in such navigation systems for establishing misalignment angles, the mission permits the use of optical azimuth alignment at launch pad in case the azimuth alignment specification is too stringent for the gyros to ensure alignment specification through self alignment process. Orbital errors show significant sensitivity to the alignment error and hence the need for tight specification on alignment accuracy. Navigation frame typically uses Launch Point Inertial (LPI) frame or the more versatile Earth Centred Inertial (ECI) frame. Applicable spacecraft orbit and the launch vehicle trajectory constraints, determine the azimuth orientation in the navigation frame.

The resident navigation software in strapdown INS goes through detailed evaluation known as hardware-in-loop simulation. The test bed permits simulation of vehicle trajectories, nominal as well as dispersions in them. Navigation software is normally configured to provide the predicted spacecraft orbit using the trajectory information available to it at satellite injection. In the absence of any other information, this information becomes extremely important as the injection information is required by the tracking ground stations to lock onto the spacecraft after injection into orbit. Later on, when the precise orbit is available from ground tracking of the spacecraft, determination of INS error becomes a possibility and this is then further used for comparison with the pre flight prediction. Calibration of sensor parameters at launch pad and update of selected coefficients, if needed, are the features built into the system.

9.2 Inertial Systems for Spacecraft

Inertial systems for spacecraft are normally configured as Inertial Attitude Reference System (IARS). It should be noted that spacecrafts are normally equipped with other types of attitude reference systems such as sun sensor, star sensor and earth sensor systems. Since orbiting spacecrafts are in freefall condition, INS cannot be used to compute its position. Spacecraft position is obtained by ground tracking and for lower earth orbits below the navigation satellite orbit, a satellite navigation receiver onboard the spacecraft can compute and transmit the spacecraft position. Requirements on IARS vary to some extent depending on the type of spacecraft orbit. Two dominant orbits in use are the Geostationary Orbit (GSO) and the Sun Synchronous Polar Orbit (SSPO), where the former orbit is circular over the equator at an altitude of nearly 36,000 km from earth, while the later orbit is also circular and nearly over the poles at typical altitudes from 500 km to 1000 km. Typical functional requirements of IARS, at different phases of orbit, are described in the subsequent paragraphs.

In a typical GTO injection scenario, the vehicle initially injects a GSO spacecraft into a GTO orbit, refer Figure 9.4, with a typical perigee altitude of around 200 km along with the required apogee altitude of 36,000 km and at an inclination, which primarily depends on launch site latitude and secondarily, on a few more trajectory related parameters.

Once the spacecraft is separated from the vehicle, the propulsion system in the spacecraft raises the perigee altitude to nearly 36,000 km and also corrects the inclination so that it is brought to near zero. This is realised by a series of spacecraft thruster firings at the apogee point with precise spacecraft orientation needed to add the incremental velocity correctly. Inaccurate pointing of spacecraft attitude, during this phase of operation, leads to increased consumption of fuel, which in turn reduces spacecraft life. Determination of the initial spacecraft attitude and subsequent holding of the spacecraft in the desired orientation during the thruster firing, are executed with the low drift inertial reference provided by the IARS. Typical requirement on these two parameters can be as follows:

Initial attitude determination error	:	< 0.05°
Attitude drift	:	< 0.05°/h

Additionally, the attitude data update rate in IARS is much faster and smoother compared to the other attitude reference sensor packages like earth sensor and sun sensor which are also available in the spacecraft. Use of IARS facilitates better closed loop control of spacecraft orientation during the thruster firing operation or during any subsequent payload operation phase.

To ensure accurate determination of the initial spacecraft attitude, the gyro drifts are calibrated in orbit using true attitude informations available from earth sensor and /or sun sensor packages. Since nature of signals from the gyro and the sun sensor are complementary, Kalman filter is used onboard to accurately estimate gyro drift and subsequently the drift compensation is updated prior to the orbit raising thruster firing operations. This indicates that the IARS should be configured to provide for these periodic drift update operations.

Velocity increment measurement is often made real time during the thruster firing operation to know the current status of the orbit being raised. This measurement requires an accelerometer package that directly outputs incremental velocity at a prefixed update interval. An accurate accelerometer with digital output often reduces the need for raised orbit determination using ground tracking which involves considerable observation time.

Once the spacecraft is parked at its intended GSO orbital slot, the subsequent correction is infrequent and primarily for north–south station keeping. In between, these operations, the inertial system can be switched OFF to conserve spacecraft power, and if the gyro is rotating wheel based, the switch OFF increases the gyro life.

Modern communication satellites, which are the dominant user of geostationary orbit, are expected to have operational life from 10 years to 15 years, and this requirement needs that the IARS is configured with redundant gyros along with electronics to avoid single point failure. This is in addition to the use of military grade radiation hardened electronic components for the gyro electronics. Space environment is not conducive to electronic components, and they need to be qualified with total radiation doze before they can be used. Redundancy management for spacecraft inertial system is not always configured like in-flight active mode as in launch vehicle. Typical macro-sized gyros in use are HRG, DTG and IFOG.

Spacecrafts in *sun synchronous polar orbits* are directly injected at the intended orbit by the launch vehicle. These orbits are primarily useful for remote sensing spacecrafts which are to take clear and high-resolution pictures of the targets on earth using cameras. The picture quality requirement imposes accurate performance specification on the IARS for pointing and for providing spacecraft attitude control with very low jitter during the imaging operation. A typical performance can be expressed as:

Attitude determination	:	$< 0.05^\circ$
Noise	:	$< 0.04^\circ$/h (rms)
Attitude drift	:	$< 0.2^\circ$/h

Periodically and prior to imaging operation, the gyro drifts are calibrated with attitude sensor packages such as earth sensor or digital sun sensor or star sensor and the drifts are updated, if found necessary. Kalman filters are used during gyro drift calibration process. Since the imaging operations are frequent, gyros are always kept ON. Use of gyro data has demonstrated extremely low jitter spacecraft control for imaging operation in comparison with other types of attitude sensors. The spacecraft life is expected to be greater than 7 years, which means that gyro redundancy is needed as in a GSO spacecraft.

An extension to such gyro based accurate payload pointing for spacecraft is seen in the case of Hubble telescope which is required to be pointing to sky with very high accuracy and very low jitter. Class of gyros which are typically used for the spacecraft operations discussed earlier, is shown in Table 9.2.

Table 9.2 Typical Gyro Accuracy Specification for Spacecrafts

Precision telescope/payload pointing	High accuracy ($< 0.001^\circ/\sqrt{h}$)
Orbit control and mission operations	Medium accuracy ($< 0.01^\circ/\sqrt{h}$)
Orbit transfer	Low to medium ($\sim 0.05^\circ/\sqrt{h}$)

9.3 Accelerometer based INS

Conventionally, both gyros and accelerometers are needed in an inertial navigation system where gyros provide inertial rotation information, while the accelerometers provide measurement of specific force. In an all accelerometer based INS, inertial rotation as well as specific force measurements are carried out using accelerometers only.

Earlier, a feasibility analysis of all accelerometer based Inertial Measurement Unit (IMU) in the navigation of spin stabilised artillery projectiles was reported by [**Harkins, 1994**]. The requirements of high spin rate and compact IMU size, were not found suitable to the gyro technology available at that time, nearly two decades ago. Recently [**Kasevich, 2002**] reported the development of all accelerometer based INS using cold atom accelerometer. The principle and features of such INS is discussed further.

9.3.1 Measurement of Angular Rate with Accelerometer

There are two approaches to measuring angular rate with accelerometers: The first approach measures the centrifugal force, whereas the second approach measures the tangential acceleration when the proof mass is acted on by angular acceleration.

In the first approach, Figure 9.6 shows the scheme of measuring angular rate ω_z using two single axis accelerometers A_1 and A_2 with their input axes 180° from each other and along Y-axis. The distance from the axis of rotation to the centre of accelerometer proof mass is assumed to be r for both the accelerometers. The outputs of the two accelerometers with identical bias can be written as follows:

$$A_1(y) = \omega_z^2 r + \text{Bias}$$

$$A_2(y) = \omega_z^2(-r) + \text{Bias} \tag{9.1}$$

Taking a difference, we get:

$$\omega_z = \sqrt{\frac{A_1(y) - A_2(y)}{2r}} \tag{9.2}$$

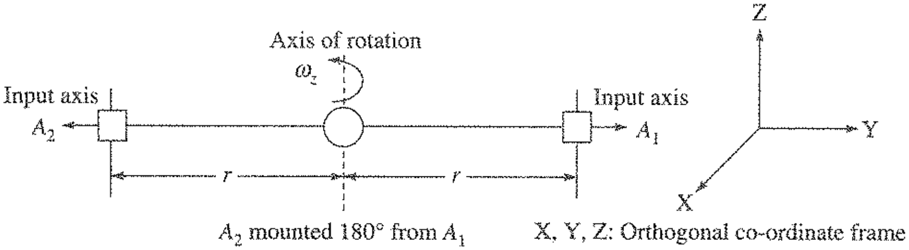

Axis of rotation
ω_z
Input axis
A_2
Input axis
A_1
r
r
Z
Y
X
A_2 mounted 180° from A_1
X, Y, Z: Orthogonal co-ordinate frame

Figure 9.6 Angular rate measurement using two accelerometers.

The difference operation improves measurement sensitivity and eliminates some common mode errors like bias in the accelerometers. However, it will be noted that the scheme fails to provide the polarity of applied rate, and this has led to the formulation based on angular acceleration, which is the second scheme. This scheme is shown in Figure 9.7 with one accelerometer A_3 whose output is given by

$$A_3(x) = \dot{\omega}_z r \tag{9.3}$$

where $\dot{\omega}_z$ is the angular acceleration as measured by the linear accelerometer $A_3(x)$.

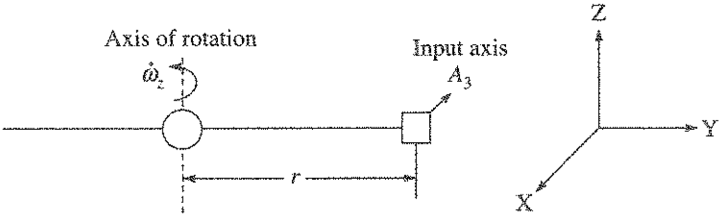


Figure 9.7 Angular acceleration measurement with one accelerometer.

The integration of accelerometer output then provides polarity sensitive angular rate.

9.3.2 All Accelerometer based Navigation Configuration

Combination of both the schemes, just described earlier, provides considerable algorithmic flexibility in the extraction of angular rate and linear acceleration. Utilising this concept and out of several schemes for an all accelerometer INS, Figure 9.8 shows a typical scheme, where the configuration with nine single axis accelerometers distributed over six locations where each box is a location, can provide simultaneous determination of specific force, angular acceleration and angular rates.

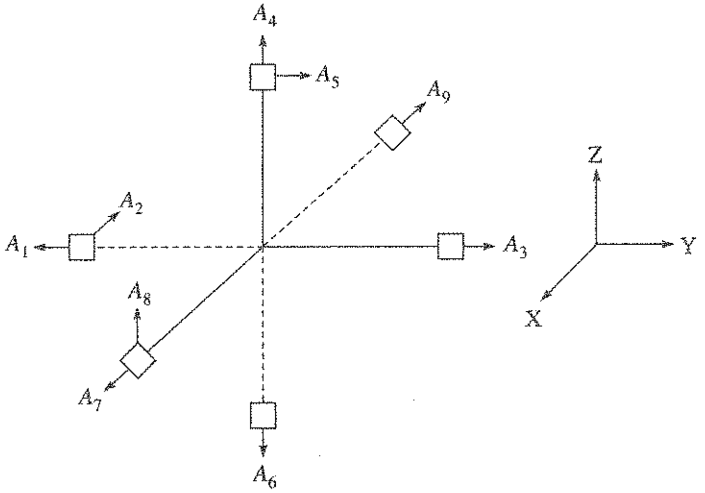
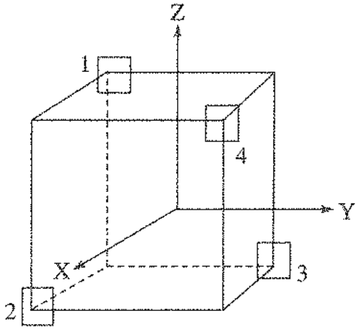


Figure 9.8 A nine-accelerometer configuration for an all accelerometer INS.

Angular rate accuracy as obtained from integration of angular acceleration is found to be dependent on accelerometer resolution and accuracy. These errors in turn will contribute to integrated angular rate to have time dependent drift as against extracted rate from Eq. (9.2). Combining both the information provide improved rate estimation.

The scheme has gone through further development and a more recent version uses three orthogonally mounted single-axis accelerometers at each of the four box type locations. The twelve accelerometers based autonomous INS scheme is shown in Figure 9.9. Additional sensors in the scheme provide redundancy for autonomous FDI also improvement in accuracy.



X, Y, Z : Orthogonal body frame
Three-axis accelerometer package located at 1, 2, 3, 4

Figure 9.9 Twelve accelerometers based autonomous and redundant INS configuration.

However, replacing gyros with accelerometers is not an easy proposition unless compelling reasons for such choice exist. Some such pros and cons are listed below.

(a) Angular rate needed is too high and which cannot be met by the available gyros, e.g. experienced in a roll stabilised vehicle configuration. This situation than provides a basis for choosing accelerometer for rate measurement.

(b) Since lever arm distance is important for improvement in rate extraction accuracy, such a system is feasible where adequate space is available such as in locomotive.

(c) The presence of lever arm means that the system be compensated for size effect.

(d) Cost of accelerometers, macro or micro, is considerably less than gyros. In emerging commercial applications, especially with MEMS, cost is a driving factor where such schemes may be considered.

9.4 MEMS based Inertial Systems

Currently, MEMS based inertial systems are able to fulfill the requirements in some aerospace systems where similar inertial systems, configured with macro-sized sensors, would have been totally unsuitable. One typical application, where MEMS technology has been found to be suitable due to its low mass, micro size and lower cost but high reliability, is in the area of *Gun launch and guide advanced projectile weapon.*

In this scheme, a projectile is launched from the barrel of a gun to hit a target. The projectile is designed with the shape of a miniaturised rocket carrying munition with its own propulsion, steering, navigation and guidance. It is required to guide the munition using the INS after the launch of the projectile. Typical requirements for execution of such a mission are:

1. The INS to be electrically power OFF condition at the time of firing (launching) the projectile. This means that the pre launch alignment of INS is of no use.
2. INS has to withstand the launch shock of more than 16,000 g.
3. Subsequent to launch, the flight time is typically less than 60 s.
4. During this short time, INS is to be powered ON, to go through in-flight alignment, put into flight mode and provide accurate navigation information for the steering program.

5. Space available for the INS is extremely small.
6. The cost is low.

[Kuhn, 2005] has done analysis of such demanding requirements and provides some interesting aspect of inertial system design requirements.

- Space is too small and launch shock too high for use of macro inertial sensors leading to choice of MEMS, which has a very high shock withstandability due to its method of fabrication.
- Since current MEMS based sensor cannot provide autonomous navigation performance, integration with GPS becomes a necessity.
- Inertial sensors cannot be calibrated prior to launch as the launch shock will most likely alter the parameter. The solution lies in in-flight calibration in a short time of less than twenty seconds during a phase called ballistic ascent.
- Initial alignment prior to launch is not possible due to power shut down during launch requiring in-flight alignment within a similar short time.
- Roll rates can become very high requiring gyros with high rate capability.
- Short flight duration along with high projectile dynamics require a tight coupling with GPS and use of extended Kalman filter scheme.

A view of the location of MEMS based IMU with integrated GPS for gun launched and guided projectile is shown in Figure 9.10. This application points to extreme environmental capability of MEMS based inertial sensors due to highly successful bulk fabrication process employed in their manufacture. The application is also an indication of successful use of integrated inertial navigation under tight coupling mode of operation, and a host of such aerospace usages are expected to emerge with MEMS based inertial systems.

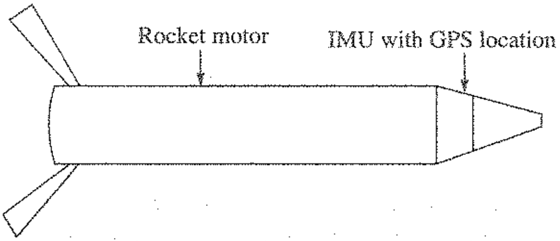


Figure 9.10 Location of MEMS based INS for gun launched and guided projectile.

9.5 Spaceborne Relative Navigation

Relative navigation is used in schemes where two vehicles are in relative motion with each other and where the motion and the position of one vehicle relative to the other are required to be known. Landing of aircraft on an aircraft carrier that is moving on sea or the mid air refueling of one aircraft with the help of another aircraft or the docking of spacecraft with the International Space Station (ISS), are some of the examples of relative navigation. In general terms, the bodies in relative motion with respect to each other, are identified as the *Target* and the *Chaser*. In the example of docking of spacecraft with ISS, the space station is the *Target*, while the approaching spacecraft is the *Chaser*. There are schemes where the instrumented relative navigation is assisted by a man in the loop, either in the *Target* or in the *Chaser* or in both. Docking with ISS is an example of such assisted scheme where there are men in the loop in both.

In spaceborne application, it can be said that inertial relative navigation is the process of determining position, velocity and direction of a target spacecraft relative to those of an active, chaser spacecraft. Taking the chaser spacecraft towards the target is the job of guidance and control.

There are new and emerging applications in space where such relative navigation schemes are needed where man in the loop is not possible. A typical scheme is described in the subsequent section.

9.5.1 Inertial Relative Navigation for Unmanned Spacecrafts

A relative navigation scheme is described where one unmanned spacecraft, the chaser, intends to move closer to another unmanned spacecraft, the target. The configuration of the chaser that is required to propel itself towards the target spacecraft, is shown in Figure 9.11.

The chaser spacecraft is equipped with a MEMS inertial sensor based INS along with a host of other sensors, whose functions are described further.

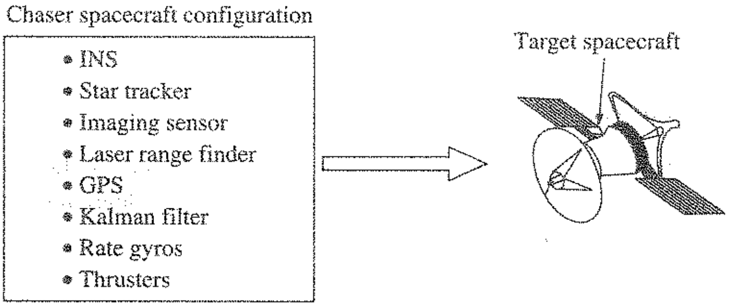


Figure 9.11 Chaser spacecraft configuration to approach a target.

The imaging sensor and the laser range finder provide angle and relative range measurements between the chaser and the target. The state vector of the Kalman filter consists of inertial position and velocity of the target spacecraft governed by a high-fidelity nonlinear orbital dynamics model, and a model of the navigation errors of the chaser spacecraft arising from the INS. The estimation error covariance matrix, on the other hand, is formulated in terms of the estimation errors in the relative position and velocity of the target spacecraft, consistent with the relative measurements from the imaging sensor and the range finder. Inertial attitude for pointing the imaging sensor at the target spacecraft and inertial rate commands for tracking it, are determined using the inertial relative position and velocity estimates of the target spacecraft. Inertial attitude of the chaser spacecraft is estimated by the INS. An auxiliary steady-state three-axis attitude estimator blends optimally the gyro and the star tracker attitude measurements.

Reaction wheels are used to rotate the spacecraft to realise the desired pointing to the target. Similarly, thrusters are used to move chaser spacecraft towards the target using appropriate glide slope guidance scheme.

Typical emerging space applications on relative navigation [**Bauer et al., 1998**] encompass unmanned rendezvous, repair of disabled spacecrafts, inspection of unfriendly satellites, protection of high valued spacecrafts and station keeping of a cluster of satellites in formation.

9.6 MEMS based Inertial Sensors in Automobiles

Use of inertial sensors in automobiles was not an economically viable proposition not long ago. But the advent of MEMS based inertial sensors is changing that scenario with ever increasing varieties of functions being witnessed in today's automobiles. The functions, which are witnessed till recent past, can be categorised as follows [**Weinberg, 2002**].

(i) Safety
(ii) Navigation and
(iii) Security.

Amongst the above three listed functions, safety systems have been the prime mover in spreading the use of micro inertial sensors in automobiles where, currently, three systems are seen as follows:

- Crash sensing
- Vehicle dynamic control
- Roll over detection

9.6.1 Crash Sensing

Crash sensing has been the earliest system to get deployed and is the most widely used safety system in automobiles. Figure 9.12 shows a typical crash detection system where two accelerometers A_1 and A_2, are used simultaneously to detect a crash by detecting abrupt change (retardation) of the velocity. Two numbers of accelerometers are shown from the redundancy point of view, but such configuration depends on the price of the automobile. A typical crash-sensing requirement stipulates that a confirmed detection of crash is to be made within 20 ms. Various signal processing schemes are employed to detect the crash but avoid false alarms which can lead to unintended deployment of airbags. In order to reduce the probability of unintended deployment, various processing schemes have emerged. In one such processing scheme, the signatures of typical crashes are also loaded in the computer for making a proper decision. The accelerometer range is around 100 g with bandwidth greater than 1 kHz. Inherent noise in high bandwidth micro accelerometer output needs to be tackled through signal processing. The output of the signal processing electronics is connected to the driver circuit, which in turn, fires the squib to deploy the airbag. From the crash detection and confirmation of decision, the airbag needs to be deployed within 30 ms.

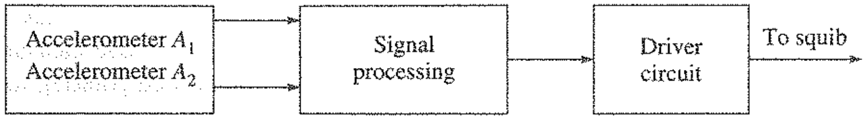


Figure 9.12 Automobile crash detection system.

All these indicate fast response along with high reliability, in the whole of crash sensing system even though the accelerometer accuracy is low. The number of such detection cum deployment system depends on the number of occupants in the car to be protected. Typical airbag locations for a car with four-occupant protection are shown in Figure 9.13. Mostly, these accelerometers are open loop but often these are configured with a forcer for self-test purpose.

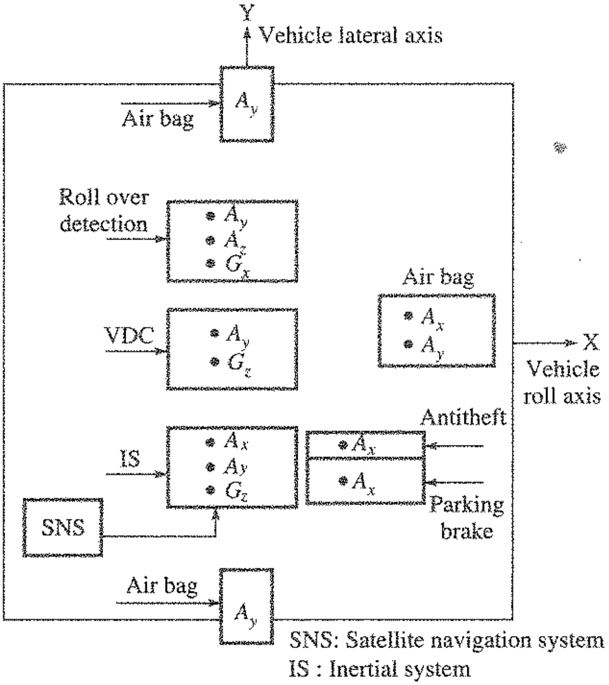


Figure 9.13 Distributed inertial sensors in automobiles.

9.6.2 Vehicle Dynamic Control

Vehicle Dynamic Control (VDC) system helps the driver to regain control of the vehicle when it starts to skid. VDC system consists of a single axis gyro, called yaw gyro, a low acceleration range lateral accelerometer (refer Figure 9.13) and wheel speed sensors at each wheel. The wheel speed data is used to compute and predict a safe yaw rate of turn for the vehicle. If the gyro (G_z) measured yaw rate exceeds from this predicted safe value or if the lateral accelerometer (A_y) detects sliding motion, single wheel braking or torque reduction is automatically activated to put the car back on the road line.

Currently, quartz vibrating tuning fork gyro or a silicon gyro is used for such application. These micro gyros, coming under the category of Coriolis Vibrating Gyro (CVG), are sensitive to road surface related vibration to appear as rate noise. Separating such noise from actual signal is a difficult design problem. As a result false alarm or missed detection can happen. Efforts are underway to design gyros with more noise immunity and improved signal processing.

9.6.3 Roll Over Detection

A roll over detection system, on detecting a possible roll over, passes information to the airbag deployment system to deploy the side airbags to protect the occupants. Vehicles, with higher centre of gravity, have greater roll over possibility. The roll over detection system uses a gyro along the vehicle roll axis (refer Figure 9.13) along with an accelerometer (A_z) along the vertical (yaw) vehicle axis. Depending on the type of system configuration, typically data from a low range lateral accelerometer (A_y) is also used to detect lateral sliding or hitting. The gyro detects the roll rate as well as the integrated roll angle, but these two may not be sufficient to predict a roll over when travelling over a banked curve where large roll angle can raise a false alarm. The vertical accelerometer data is used to assess the presence of such banked curve. The gyro rate range is usually higher than that needed in VDC system requiring excellent tolerance to impact shock, as quite often, roll over takes place after hitting another vehicle or a stationary object.

9.6.4 Navigation

Availability of two-dimensional navigation facility in a car is of recent origin and currently it is found in the higher priced segment of automobiles. A satellite navigation system, such as GPS, constitutes the core of an automobile navigation system, but such GPS, currently, does not provide direction of travel known as heading information. Also, GPS signal blackout or drop in strength is possible inside a tunnel or in the proximity of tall buildings. So, quite often GPS data is used along with an Inertial System (IS) consisting of two accelerometers, A_x and A_y, and one vertical gyro G_z. Initial vehicle direction is set up using a magnetic compass to which the gyro inertial reference is initialised. Subsequently, the gyro can provide direction of travel, also called vehicle attitude, information. A periodic update of gyro derived direction information is obtained through the use of map. In the absence of GPS data, accelerometer data is used for navigation in a dead reckoning mode where the velocity error can be periodically reset. Both the accelerometers are of low acceleration range.

Security is a typical emerging function, where a low range accelerometer (A_x) is used to assist an automated parking brake system, especially useful while parking on a slope, where the accelerometer data is used in calculating the brake force necessary.

Similarly, another system against vehicle theft has emerged which can raise alarm in the event a vehicle is stolen by towing it with another vehicle. Here, a low range accelerometer is used to sense the tilt during the towing operation. The parking accelerometer can do this function also.

In Figure 9.13, it is seen that the inertial sensors are distributed and catering for specific functions as described above. A more elegant and cost effective scheme may emerge in the shape of a three-axis IMU, which can do similar functions and expand the inertial system usage in automobiles to include integrated navigation in the near future.

SUMMARY

The applications, described in this chapter, have highlighted the current diversity of usage and some of their emerging scenario with particular relevance to aerospace. While standalone precision inertial navigation aerospace continues, gradual emergence and usage of cost effective satellite integrated inertial navigation are becoming quite noticeable. MEMS based inertial systems with their low cost and lower size yet offering high reliability, are extending the application horizon where traditional inertial sensors would have been technically unsuitable.

REFERENCES

Bauer, Frank H., Hartman, K., and Lightsey, E.G., *Spaceborne GPS—Current Status and Future Visions*, IEEE Aerospace Conference, 3, pp. 195–208, USA, 1998.

Harkins, Thomas E., *Assessing the Feasibility of Accelerometer—Only Inertial Measurement Units for Artillery Projectiles*, U.S. Army Research Laboratory, June 1994.

Kasevich, Mark, *Science and Technology Prospect for Ultra Cold Atom*, Stanford University, CAMOS, November 2002.

Kuhn, T., *In Flight Initialisation and Alignment of Launched Ammunitions*, Symposium Gyro Technology, Stuttgart, Germany, 2005.

Weinberg, H., *MEMs Sensors are Driving the Automotive Industry*, *Sensors*, February 2002.

Appendix A

Laser Principle and Basic Characteristics for Gyro

Laser Operation

The laser is an optical oscillator. The word laser is an acronym for 'light amplification by stimulated emission of radiation'. This indicates that to make a laser, it is necessary to generate stimulated radiation at the expense of spontaneous emission. An ordinary neon light emits photons that are categorised as spontaneous emission and the transformation of this wave to stimulated emission, originally attributed to Einstein, converts it to laser.

Stimulated emission takes place when a photon encounters an excited atom and forces it to emit another photon of the same frequency, in the same direction and in the same phase. The photons go-off as coherent radiation. Thus coherency is an essential feature of laser light beams and the process is briefly described in the next paragraph.

In the absence of external radiation, the ratio of the number of excited atoms n_E to the number of unexcited atoms n_0 is very low. Now, if a radiation of frequency f that corresponds with the energy difference E is sent through the cell, it increases the ratio n_E/n_0. By increasing the radiation if a condition can be created, known as *population inversion*, the rate of energy radiation by stimulated emission may exceed the rate of absorption. The system then acts as a source of radiation with photon energy E. Since the photons are the result of stimulated emission, they all have same frequency, phase, polarisation and direction. The resulting radiation is thus extremely coherent as compared to the light from neon where the emissions from the atoms are not co-ordinated.

In a helium–neon gas laser, the population inversion is achieved by sealing the appropriate gas mixture in a partially evacuated tube that is provided with electrodes. Application of sufficiently high voltage between the electrodes then creates the condition for stimulated emission. To create the population inversion, which means laser transition in neon, mixing of neon with helium is necessary because in this condition helium atoms get excited and pass energy to the neon atoms. A gas laser is thus a coherent light source. The degree of coherence, high or low, depends on the length of time for which the phase is continuous.

Coherence

Coherence is an important property in laser useful to ring laser gyro to provide optimum performance.

Coherence is a measure of the ability of a light source to produce high contrast interference fringes when the light is interfered with itself in an interferometer. On the other extreme, a non-coherent light such as ordinary light will not provide visible fringes. Normally, the interference pattern occurs when the light beam is divided and then recombined with phase difference between the two paths. One method of creating the phase shift is by introducing difference in the path length traversed by the two beams. However, to ensure that path length difference alone produces visible interference, it is quite important that the light beam frequency is stable and coherency is the property of the laser light that gives the stability.

There are two types of coherence—temporal and spatial. Temporal coherence describes the correlation or predictable relationship between waves observed at different moments in time. Spatial coherence describes the correlation between waves at different points in space. Temporal coherence is quantified by the expression:

$$L_c = c\tau_c$$

where

τ_c = coherence time
c = light velocity
L_c = coherence length

τ_c can be defined as the time duration for which the phase of the laser wave is stable.

Linear Resonator and Laser Modes

The optical waves within an optical resonant cavity are characterised by their resonant modes, which are discrete resonant conditions governed by the dimensions of the cavity. The laser beam radiated from the laser cavity is thus not arbitrary. Only the waves oscillating at modes that match the oscillation modes of the laser cavity can be produced. The occurrence of longitudinal modes is explained with a two-mirror Fabry–Perot resonator, also known as linear resonator.

When a laser gain medium is inserted in a Fabry–Perot cavity with mirrors located at the two ends, as shown in Figure A.1, modes in the form of standing wave patterns build up in the cavity which are equally spaced in frequency. Various standing waves, each of a different frequency, are known as longitudinal modes. These are called longitudinal as these modes are associated with longitudinal direction of the light waves within the cavity.

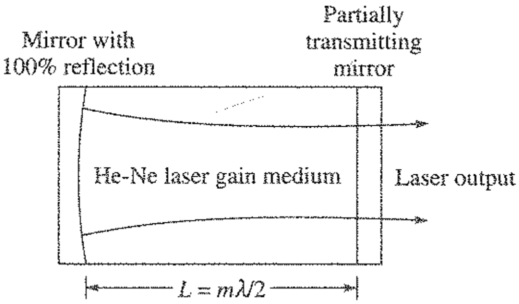


Figure A.1 Two-mirror Fabry–Perot resonator.

Two conditions are needed to be satisfied for the resonator longitudinal modes to develop.

(i) The gain provided by the gain medium exceeds the losses in the cavity at that frequency. The losses are due to scattering, absorption or wrong velocity.

(ii) There exists an integral value m such that the specific frequency is satisfied by the relation $L = m\lambda/2$, where λ is wavelength and L is the cavity length.

The longitudinal mode number will be very high for typical L and λ. But the actual number of longitudinal mode in the laser output can be visualised when the gain curve is superimposed as described further. The gain curve, resembling a Gaussian pattern, is shown in Figure A.2. It is a typical plot of output light intensity or power against frequency where cavity loss is shown as constant in the frequency band.

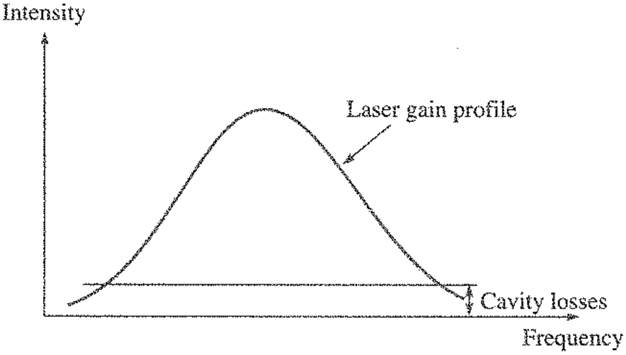


Figure A.2 Laser gain curve showing the loss.

Figure A.3 shows longitudinal modes in a cavity with frequency separation given by $c/2L$ where c is the light velocity.

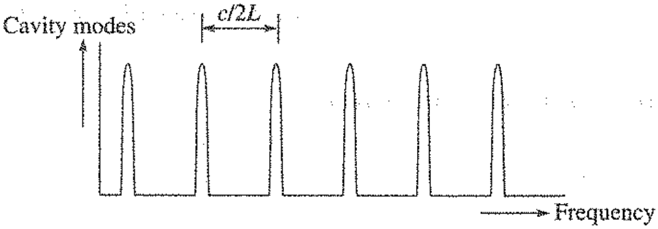


Figure A.3 Typical longitudinal modes in a cavity.

Figure A.4 shows the modes in the output power when the gain curve is superimposed to show number of modes will be only four that actually fit in the gain curve and the corresponding power associated with each mode.

For a 25-cm length Fabry–Perot resonator, frequency separation is computed as:

$$\frac{3\times10^8}{2\times0.25}=6\times10^8\ \text{Hz}=0.6\,\text{GHz}$$

With four number supporting longitudinal modes as shown in Figure A.4.
Possible gain bandwidth = 3 × 0.6 = 1.8 GHz

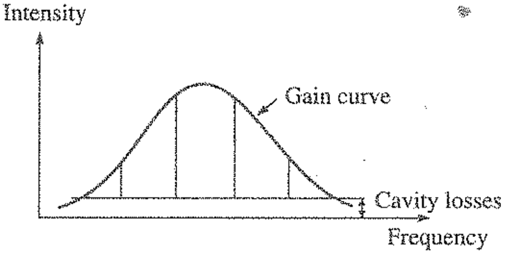


Figure A.4 Longitudinal modes in the output power.

Ring Resonator

In contrast with the linear resonator, in a ring laser resonator, the CCW wave and the CW wave circulate due to the reflection of the suitably positioned mirrors and can have different amplitude and frequency of oscillation and so no standing wave modes. To achieve the ring resonator, minimum of three suitably positioned mirrors are necessary. In a gyro, the ring cavity is tuned to have only one longitudinal mode to align with the maximum intensity point of the gain curve shown in Figure 3.32 in Chapter 3. Demonstration of Sagnac effect is possible in ring configuration and to optimise the performance as a Sagnac effect gyro, such resonator has to eliminate quite a few errors through accurate understanding of the resonator error mechanism.

Langmuir flow

In active laser media, neutral atoms along the centre of the discharge move toward the cathode, while the atoms near the walls move toward the anode. Since the lasing light is concentrated in the centre of the tube, so it passes through the gas mixture flowing towards the cathode. Two counter propagating beams see opposite motion for the lasing medium and hence different index of refraction. This gas flow has been characterised as Langmuir flow that leads to Fresnel drag, which is a source of bias. This drag effect is compensated by providing two anodes with opposite polarities.

Mode competition between the beams

Electrons within the gas are accelerated and excite helium atoms, which in turn populate metastable He energy levels. Some of this energy is transferred by atom collision to Ne atoms, which populate energy levels having excitation energy similar to that of metastable He energy levels. Excited Ne atoms radiatively decay to lower Ne energy states generating CW and CCW optical beams. To ensure that the resonant modes co-exist in the active cavity and have the same optical power level, the mode competition between them must be avoided.

A gain competition between the counter propagating beams is seen in He–Ne gas resonator when only one Ne isotope is present. In order to avoid the gain competition in a He–Ne transition with typical wavelength of 0.633 μm, a combination Ne_{20} and Ne_{22} isotopes in equal proportion is used.

SUGGESTED READING

Mario, N. Armenia, Caterina, Ciminelli, Francesco, Dello'lio, and Vittoria, M. Passaro, *Advances in Gyroscope Technologies*, Springer–Verlag, Berlin Heidelberg, 2010.

Sears, Francis W., Zemansky, Marc W., and Young, Hugh D., *University Physics*, Indian Student Edition, Narosa Publishing House, New Delhi, 1992.

Silvast, William T., *Laser Fundamentals*, 2nd ed., Cambridge University Press, New York, USA, 2003.

Appendix B

Fibre-Optics Features and Basic Characteristics

The fibre-optics features and characteristics are brought out keeping in mind the use in an interferometric gyro.

An optical fibre is a flexible, transparent fibre made of glass (silica), slightly thicker than a human hair. It functions as a waveguide or light pipe to transmit light between the two ends of the fibre. Optical fibres typically include a transparent core surrounded by a transparent cladding material with a lower index of refraction. Light is kept in the core by total internal reflection. This causes the fibre to act as a waveguide. Figure B.1 shows a plane interface formed between two media of refractive indices n_1 and n_2. The incident light travelling in a medium of refractive index n_1 and making an angle θ_1 to the normal is partly refracted also into the medium with refractive index n_2 and making angle θ_2 with the normal as shown. Snell's law defines the relationship between the angles and the indices given by:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

When the angle of incidence θ_1 is such that θ_2 is 90°, then, this incident angle is called *critical angle*, and any further increase in θ_1 will make the incident light totally reflected.

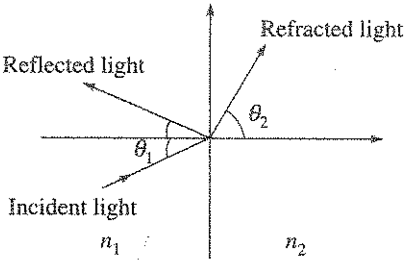


Figure B.1 Illustration of refraction of light.

This effect is used in optical fibres to confine light in the core. Light travels through the fibre core, bouncing back and forth off the boundary between the core and the cladding as shown in Figure B.2. The angle of incidence at the core–cladding interface remains same.

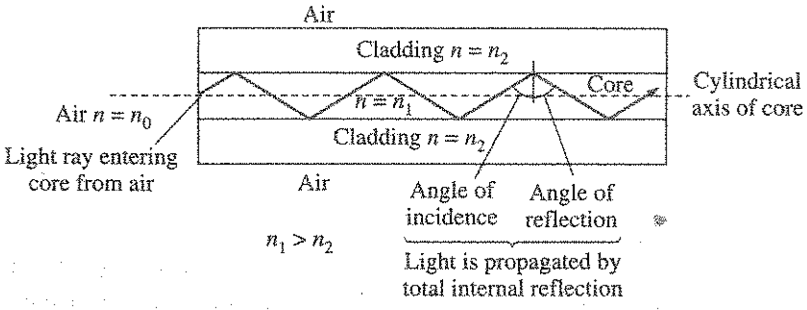


Figure B.2 Illustration of light propagation in fibre-optic cable.

In the case of bent fibre, as in the case of gyro coil, the same total internal reflection takes place. But in this case, the angle of incidence at the core–cladding interface changes, and this leads to a possibility of the incident angle falling below the critical angle which allows loss of light due to refraction. Sharp bends and kinks in the fibre lead to such loss. The typical refractive index value for the cladding material of an optical fibre is 1.52, and for the core, the value is typically 1.62. The most widely used material is silica.

Single Mode and Multimode Fibre

The feature of light propagation in the fibre core should be such that the phases of the reflected rays reinforce each other, so that the phase difference must be multiples of 2π. There are incident angles which allow the rays to propagate, but they may be having different modes. If the fibre size and index of refraction are appropriate, then it is possible to allow only one mode at a particular wavelength to propagate. Such a fibre is called single mode with typical core diameter between 8 μm–10 μm and clad diameter of around 125 μm. This type of optical fibre has been found suitable for gyro. Multimode mode fibre, on the other hand, support many spatial modes of light propagation each with its own modal velocity. Coupling between these modes through light scattering creates non-reciprocity, which is a source of error in the gyro. Multimode mode fibres are bigger, typically the core diameter is around 50 μm.

Polarisation-maintaining Fibre

Polarisation-maintaining fibre is a specialty optical fibre with strong built-in birefringence, preserving the properly oriented linear polarisation of an input beam. The word polarisation for laser light is dealt in many books [**Ghatak and Thyagaragan, 1999**] which describe linearly polarised light, circularly and elliptically polarised light waves and their characteristics and can be referred for further insight. The word birefringence is the phenomenon of double refraction, or the polarisation dependence of the refractive index in a medium.

Optical fibres always exhibit some degree of birefringence, even if they have a circularly symmetric design, because in practice there is always some amount of mechanical stress or other effect which breaks the symmetry. As a consequence, the polarisation of light propagating in the fibre gradually changes in an uncontrolled (and wavelength-dependent) way, which also depends on any bending of the fibre and on its temperature.

This problem can be fixed by using a *polarisation-maintaining fibre*, which is *not* a fibre without birefringence, but on the contrary a fibre with a strong built-in birefringence (*high-birefringence PM fibre*). Provided that the polarisation of light launched into the fibre is aligned with one of the birefringent axes, this polarisation state will be preserved even if the fibre is bent.

Attenuation in fibre-optics, also known as transmission loss, is the reduction in intensity of the light beam (or signal) with respect to distance traveled through a transmission medium. Attenuation coefficients in fibre-optics usually use units of dB/km through the medium due to the relatively high quality of transparency of modern optical transmission media. Research has shown that attenuation in optical fibre is caused primarily by both scattering and absorption. Typical loss figure is around 0.2 dB/km at 1.55 μm wavelength of laser.

REFERENCE

Ghatak, Ajay, and Thyagaragan, K., *Introduction to Fiber Optics*, Cambridge University Press, First South Asian Edition, New Delhi, 1999.

Appendix C

Quality Factor

The term Quality factor, normally abbreviated as Q-factor, is an important parameter in the design and operation of modern inertial sensors and has appeared in the description of gyros, accelerometers and MEMS based inertial sensors. The derivation and interpretation of Q-factor, especially when the Q is extremely high, is presented here.

In an RLC circuit, the characteristic equation for a damped linear oscillator can be written as:

$$L\frac{d^2q}{dt^2}+R\frac{dq}{dt}+\frac{q}{C}=0 \tag{C.1}$$

In the case of less than critical damping and assuming an initial charge A on the capacitor, the solution is given by

$$q=\Re Ae^{-(R/2L)t}e^{j\omega_1 t}\text{, where } \omega_1=\sqrt{\frac{1}{LC}-\left(\frac{R}{2L}\right)^2} \tag{C.2}$$

In the absence of the damping, the resonant frequency of the circuit is:

$$\omega_0^2=\frac{1}{LC} \tag{C.3}$$

Since Q is qualitatively defined as a measure of energy loss in the circuit, Q is given by the relation:

$$Q=\frac{\omega_0 L}{R} \tag{C.4}$$

Using this defined relation of Q and putting in the equation for ω_1, we have:

$$\omega_1=\omega_0\sqrt{1-\left(\frac{1}{2Q}\right)^2} \tag{C.5}$$

When $Q >> 1$, then ω_1 and ω_0 are practically similar in magnitude, then the solution for q can be expressed in the form as:

$$q = \Re A e^{-(1/2Q)\omega_0 t} e^{j\omega_0 t} \tag{C.6}$$

Defining, the time constant τ of the circuit as:

$$\tau = \frac{2Q}{\omega_0} \tag{C.7}$$

Equation (C.6) can be expressed as:

$$q = \Re A e^{-(t/\tau)} e^{j\omega_0 t} \tag{C.8}$$

Slater and Frank further showed that Q can also be expressed as:

$$Q = 2\pi \times \frac{\text{Total energy}}{\text{Decrease in energy in one period}} \tag{C.9}$$

Thus there are three forms of definitions for Q given by Eqs. (C.4), (C.7) and (C.9) which are normally used as well as used in this book. For very high Q resonators, Eq. (C.7) is extremely useful as it provides experimental measurement of Q. For electronically boosting the Q, Eq. (C.9) is useful as it provides a clue that if the energy losses can be compensated, it is possible to raise the Q further.

REFERENCE

Slater, John C. and Frank, Nathaniel H., *Mechanics*, McGraw-Hill, New York and London, 1947.

Appendix D

Inertial Sensor Noise

Introduction

Noise characterisation of modern macro sized strapdown sensors is of particular relevance to extract its performance. It has been explained earlier in Chapter 3 that in a ring laser gyro, randomised dither is applied to the gyro to get rid of the problem of lock-in dead band, but this process generates noise in the gyro output known as angle random walk. Spontaneous emission of photons in RLG resonator produces noise in the output or the detector current random fluctuation creates shot noise in a fibre-optic gyro that is characterised as random walk. Micro gyros and accelerometers are exhibiting considerable magnitude of noise in its output due to thermo-mechanical effect.

Noise Classification

Some important noise parameters, which are used to represent the inertial sensor noise, are discussed as follows:

Quantisation noise

It is due to finite resolution in the sensor digital output when the output data pulse frequency is not synchronised with the sampling frequency of the counter. The digital resolution is called pulse weightage. If, for example, q is the pulse weightage for a gyro expressed in rad/pulse, the number of pulses N generated over an arbitrary sample time τ at a constant angular rate Ω rad/s can be expressed as:

$$N = \frac{\Omega\tau}{q} \tag{D.1}$$

Within the sampling interval N will consist of a whole number W and a fraction part F. This will mean that a defined number of the sampling intervals will record the whole number pulses W, while an extra pulse will be recorded at a periodic interval of $1/F$. Mathematically, this is equivalent to subtracting a uniformly distributed random variable $n(t)$ from the true value of the signal. The probability distribution of n is given as:

$$p[n(t)] = \begin{cases} \frac{1}{q} & \text{if } 0 \leq n \leq q \\ 0 & \text{Otherwise} \end{cases} \tag{D.2}$$

where q is the sensor resolution, gyro or accelerometer. The quantisation noise can be assumed to be uncorrelated in time and is given by

$$\frac{q^2}{3}\delta(\tau) \tag{D.3}$$

where, $\delta(\tau)$ is the delta function. The noise has a slope of -1 in the Allan variance time domain noise characterisation plot as shown in Figure D.1. Unit of quantisation error is in radian or arc-s.

Random walk

It results in a sensor output due to integration of white noise process. A white noise process usually has a zero mean and when stationary, it has a constant Power Spectral Density (PSD). In the case of a gyro with white noise in the angular rate output, the integration gives rise to *Angle random walk.* When a white noise angular rate error is integrated over time, the effect it has on the integrated result is equal to:

$$\sigma_\theta(\tau) = \sigma\sqrt{\tau\delta\tau} \tag{D.4}$$

where $\sigma_\theta(\tau)$ = standard deviation of the integrated angular error, τ = integration time.

Angle Random Walk (ARW) coefficient is defined when the integration is done over a period of 1 second, and we can write:

$$\text{ARW} = \sigma_\theta(\tau) \tag{D.5}$$

Now we further notice that the unit of ARW coefficient is given as $°/\sqrt{h}$. So, the angular error due to gyro ARW can be expressed as:

$$\text{Angular error} = \text{ARW}\sqrt{\tau} \tag{D.6}$$

For example, if it is required to know the angular error due to gyro ARW of $0.01°/\sqrt{h}$ over a period of 100 s, then we can compute the angular error as $0.01 \times \sqrt{100} = 0.1°$.

Random bias

It is sometimes referred as instability in bias. It is an unpredictable random quantity with constant but random amplitude. The characteristic condition of the noise is Gaussian with zero mean and constant variance. If σ_B is standard deviation of bias error, then the variance of bias error can be defined as:

$$E[\sigma_B^2] = \text{constant} \tag{D.7}$$

where E is the statistical expected value operator. In Allan variance plot, it has a slope of zero. The unit of random bias in gyro is expressed as °/h.

Random walk in rate

Random walk in rate is described as an exponentially correlated noise with a very long correlation time. Rate biased RLG exhibits this noise term. If K is the rate random walk coefficient, then it can be represented in Allan variance form as:

$$\sigma^2(\tau) = \frac{K^2\tau}{3} \tag{D.8}$$

It has a slope of +1/2 in the characteristic shown in Figure D.1, and the unit is expressed as °/h/$\sqrt{h}$.

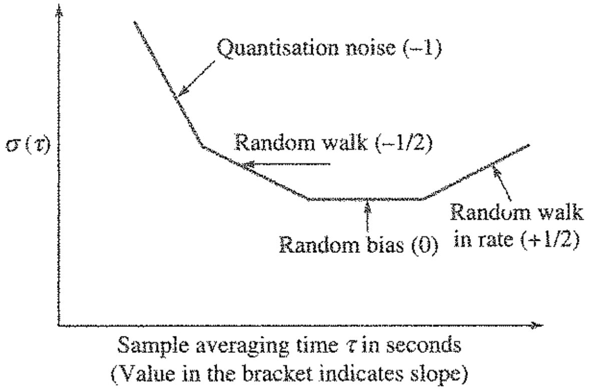


Figure D.1 Allan variance plot depicting sensor noise.

Total noise equivalent acceleration

In micro accelerometer, thermal energy causes random motion of mechanical components known as Brownian motion and its effect on the proof mass determines the fundamental noise limit. Total Noise Equivalent Acceleration (TNEA) with unit as $[\text{g}/\sqrt{\text{Hz}}]$, defines the magnitude of noise in the accelerometer output and is given by

$$\text{TNEA} = \sqrt{\frac{4K_B T\omega_n}{QM}} \tag{D.9}$$

where

K_B = Boltzman's constant
T = absolute temperature in Kelvin
M = proof mass
Q = quality factor
ω_n = natural frequency of accelerometer.

Analysis of Noise

There are two widely used methods for analysing the noise terms:

1. Allan variance time domain method.
2. Power Spectral Density (PSD) frequency domain method.

Over the years, Allan variance method has been successfully adopted in evaluating the sensor noise terms, and this method is described further.

Allan Variance

The Allan variance, named after David W. Allan, was originally developed for precision measurement of stability in clocks and oscillators. It is also known as the *two-sample variance*. It is defined as one-half of the time average of the squares of the differences between successive readings sampled over the sampling period, and so it is expressed as:

$$\sigma_y^2(\tau) = \frac{1}{2}(y_{n+1} - y_n)^2 \tag{D.10}$$

where

τ = averaging period

y_n = average value of the nth sample over the time period

The samples are taken with no dead time between them. The Allan variance depends on the time period used between samples, therefore, it is a function of the sample period, as well as the distribution being measured, and is displayed as a graph rather than a single number. This method has been adopted for estimating the inertial sensor noise terms where precision is important.

In practice, the sensor data will contain combination of such noise terms and when they are statistically independent, the total variance is sum of the individual process variance. Thus, to estimate the magnitude of the random noise at any sampling interval τ needs knowledge of other noise parameters in the same region. Cluster sampling technique is a method where finite number of clusters can be generated from any finite set of data. Allan variance of any noise term is estimated using the total number of clusters of a given length that can be created. In the plot form, the noise is defined as shown in Figure D.1. The plot is between the averaging time τ and the square root of Allan variance $\sigma^2(T)$ and normally plotted on a log–log scale.

The plot clearly shows the separation of noise terms using appropriate sample averaging time.

Generation of the Allan Variance Plot

Sensor data is accumulated at the desired sampling rate that should be high enough, typically 100–400 Hz. A prior kno ledge on the frequency component of the noise helps in arriving at the data sampling rate. The accumulated data is divided into a set of identical clusters based on the averaging time τ where each cluster will have sufficient data. Average the data in each cluster. The averaging time τ should start from the lowest and the number of data points in a cluster must be > 9. Higher the number, better is the estimation accuracy. It is recommended [**IEEE Std 647–1995**] that the data length should be around 3 orders of magnitude, which means that for $\tau = 1$ s, the data length should be for 1000 s. Thereafter, change τ to the next higher value and repeat the process. Allan variance is estimated from the following equation:

$$\sigma_y^2(\tau) = \frac{1}{2(n-1)} \sum_i (y(\tau)_{i+1} - y(\tau)_i)^2 \tag{D.11}$$

where

y_i = average value of the measurement in the ith cluster
n = total number of clusters

When the computed values of $\sigma_y(\tau)$, known as square root of Allan variance, are plotted against the corresponding values of τ, the pattern similar to that of Figure D.1 is obtained from where the slopes can be estimated. As described earlier, angle random walk is obtained from the data measured at 1 s averaging. Certain precautions such as environmental temperature control may be needed if the sensor shows high temperature sensitivity.

REFERENCE

IEEE STD 647–1995 on IEEE Standard Specification Format Guide and Test Procedure for Single-Axis Laser Gyros, Published by IEEE, New York, USA, May 15, 1996.

Glossary

Navigation

Aided Inertial Navigation: A scheme that corrects the time dependent growth of inertial navigation errors by using data from a complementary sensor whose errors do not propagate with time.

Attitude: Orientation of the body fixed co-ordinate system of a vehicle with respect to reference co-ordinate frame.

Attitude and Heading Reference System: An inertial system which measures the direction of motion of the vehicle (mostly aircraft) with respect to the instrumented geographic north and local level reference.

Attitude Reference Frame: The reference frame with respect to which the vehicle attitudes (pitch, yaw and roll) are measured.

Body Frame: It is a right handed frame to which inertial sensors are either physically or computationally aligned in a navigation system.

Circular Error Probability: It is a statistical method of defining error in two-dimensional navigation where the error is expected to be within a circle whose radius defines the error magnitude.

Coriolis Force: It is the force that explains the changes in the motion of a particle when the motion is viewed in a rotating co-ordinate frame as against viewing the same motion in an inertial co-ordinate frame.

Differential Operation: It is a method by which measurement errors inherent to the same satellites, particularly the ephemeris errors and those caused by reduction in the velocity of light during travel through the atmospheric layers, are eliminated to a great extent. The terminology is used in satellite navigation system.

Dilution of Precision: A dimensionless number that defines degradation factor in satellite navigation computation based on pseudo range measurement.

Earth Centred Inertial Frame: It is an inertial frame with origin located at the centre of earth with one axis pointing towards pole star, the second along vernal equinox and the third conforming to right handed frame.

Ellipsoid Geometry of Earth: A mathematical model of the geometry of earth that describes the earth as an ellipsoid, which is frequently used in inertial navigation.

Error Compensation Model: Mathematical model that defines inertial sensor error characteristic which is used to compensate for navigation error.

Failure Detection and Isolation: A scheme incorporated in a redundant inertial navigation system, which detects sensor failure and then isolate it so that the failed sensor data is not used in navigation.

False Alarm: The alarm raised in software when the level of signal, primarily due to short duration noise, crosses the detection threshold. The terminology is mostly used in redundant inertial system.

Frame of Reference: The right handed orthogonal frame with respect to which all measurements of position, velocity and attitudes are carried out in a navigation system.

Geographic Frame: A non inertial reference that mechanises local north direction and local direction of gravity with the origin at the location of the navigation system. Latitude, longitude and altitude are the position output of such a system.

Geometrical Dilution of Precision: A dimensionless number that relates error in navigation with the geometry of the sensor layout used in inertial navigation system. The terminology is also used to describe the effect of visible satellite geometry on the satellite navigation error.

Global Navigation Satellite System: Multiple and independent satellite navigation systems comprising GPS, GLONASS and Galileo which will provide global navigation coverage.

Gravitation Ellipsoid: A mathematical model of earth's Newtonian gravitation that defines gravitation based on the ellipsoid model of earth.

Gravity Gradiometer: An instrument system based on accelerometers to measure gradient of gravity when mounted on a flying vehicle.

Gyrocompassing: An autonomous method of measurement of the angular relationship between the sensor body frame and the navigation reference frame by using gyros and accelerometers.

Horizontal Dilution of Precision: The degradation factor in navigation while computing the horizontal position from range measurement.

Inertial Frame: A non rotating and non accelerating frame of reference on which Newton's laws of motion are valid. For most operational inertial navigation, such a frame is conceived as stationary with respect to fixed stars.

Inertial Measurement Unit: A system that measures specific force and angular rotation with respect to a predefined reference frame.

Inertial Navigation System: A system that measures position, velocity and orientation by instrumenting any of the defined reference frames using gyros and accelerometers.

Innovation Sequence: Defined as difference between actual measurement and predicted measurement, which finds usage in integrated inertial navigation.

Integrated Inertial Navigation: A better estimate of navigation information derived from independent measurements carried out by a set of complementary navigation sensors one of which is an inertial navigation sensor.

Local Gravity: It is the combined acceleration effect of Newtonian gravity of the earth mass and the inward centripetal acceleration due to rotation of earth about its polar axis, which are acting on a stationary mass. This is also referred as Plumb bob gravity.

Missed Detection: It is the inability of the failure detection algorithm to detect a soft failure. The term is used in redundant inertial system.

Navigation Update: It is the execution of onboard navigation task computation where the digital computation is carried out and updated at predefined time intervals during the entire navigation period.

Pseudo Range: Inaccurate range measurement obtained primarily due to synchronisation error between the receiver clock and the navigation satellite clock.

Satellite Navigation System: A system that measures position and velocity of a vehicle using measurement of ranges to the orbiting satellites.

Schuler Frequency: A frequency that is observed in geographic frame navigation error propagation, which has a periodicity of 84.4 minutes that is similar to Schuler pendulum time period of oscillation.

Self Alignment: It is a method of performing autonomous alignment in a strapdown navigation system by using the inertial sensors meant for navigation during flight.

Skewed Sensor Configuration: It is a configuration used in redundant strapdown system where the sensor input axes are preferentially skewed with respect to an orthogonal frame.

Soft Failure: It describes a performance shift of the inertial sensor when such shift marginally exceeds the allowable soft detection threshold in a redundant system.

Specific Force: The acceleration that is sensed by a vehicle borne accelerometer, which does not include gravitational acceleration.

Stabilised Platform System: A system that maintains its reference frame by isolating the angular rates of the carrier vehicle using gyros and gimbals.

Strapdown Navigation System: A system that analytically maintains its reference frame and derives navigation information using gyros and accelerometers which are strapped to the carrier vehicle.

Time Dilution of Precision: The degradation factor in computing the receiver clock bias in satellite navigation.

Vernal Equinox: It is the direction of the sun from the earth when the sun is crossing the equator during March 21–23.

Vertical Dilution of Precision: The degradation factor in computing vertical distance in satellite navigation from pseudo range measurement.

World Geodetic System 84: Datum for the earth issued in the year 1984 that defines various parameters such as reference ellipsoid, gravitational constants, rotation rate and ellipticity.

Inertial Sensors

Accelerometer: A sensor that senses the inertial acceleration of proof mass for measuring linear acceleration that acts on its sensitive axis.

Atom Interferometer Inertial Sensors: Interferometer that utilises the quantum mechanical wavelike properties of atom which are made to counter propagate and interfere. Detection of the phase difference of the interfered waves due to acceleration or angular rate results in either an accelerometer or a gyro.

Bulk Micromachining: The term applied to a micromachining process when a part of the substrate is etched to form the mechanical structure.

Capturing: It is a scheme that holds the gyro rotor or the accelerometer proof mass at its specified point through use of servo loop.

Circular Polarisation: It is the result of the superposition of two linearly polarized light waves having phase difference of $\pi/2$ or its odd multiples. The resulting wave can be right circularly polarised if the sense of motion to an observer is clockwise and left circularly polarised when the sense of motion to an observer is counterclockwise. These are the features seen in multi-frequency optical gyros.

Cluster (Orthogonal): A machined structure that provides three-dimensional orthogonal mounting provision to assemble the gyros and the accelerometers.

Coriolis Acceleration: The increment of acceleration relative to inertial space, which arises from the velocity of a mass relative to a rotating co-ordinate system.

Coriolis Vibrating Gyro: A vibrating gyro that measures inertial rotation by sensing the Coriolis force induced motion of the vibrating proof mass or the associated structure.

Degree of Freedom: An allowable freedom in a gyro spin axis with respect to the case. The freedom number represents the number of orthogonal axes about which the spin axis is free to rotate.

Dithered Gyro: The ring laser gyro that overcomes lock-in effect using mechanical dither about its input axis.

Drift Rate: The time rate of output deviation of gyro from the expected output and expressed as an equivalent input angular displacement per unit time with respect to inertial space. It is normally expressed as degrees per hour.

Dynamic Range: It is the ratio of maximum input operating range to minimum detectable threshold. This term is used to describe an important characteristic of inertial sensor.

Dynamically Tuned Gyro: A two-degree of freedom spinning rotor gyro that frees its rotor from the angular motion of the case by tuning the gimbal flexure spring torque with the negative dynamically induced spring torque.

Faraday Bias: It is a magneto-optical bias introduced in a multifrequency ring laser gyro to overcome the effect of lock-in.

Gyro (Gyroscope): A device to sense angular motion with respect to inertial space without external reference.

Hemispherical Resonator Gyro: A gyro working on the principle of a Coriolis Vibrating Gyro having the resonating structure shaped like a hemisphere.

Input Axis: The sensitive axis along or about which an input causes maximum output.

Interferometric Fibre-optic Gyro: A gyro which instruments Sagnac effect by measuring the phase difference of two interfering counterpropagating light waves.

Lock-in: It is a characteristic seen in ring laser gyro due to frequency locking between the counter propagating beams. This results in a dead band around zero angular input rate.

Mass Unbalance Coefficient: It is a characteristic in a spinning rotor gyro due to assembly error that gives rise to a drift under acceleration.

Micro electromechanical Inertial Sensor: Inertial sensor that is produced using micro-fabrication technique which has been adapted from integrated circuit manufacturing.

Micromachining: A fabrication process that realises three-dimensional structures primarily in silicon and quartz using process and the related facility, which has its origin in integrated circuit manufacturing.

Modeleble Residual: A term that is often used in inertial sensor model evaluation data analysis to describe the statistical residue in the model after the fitting process. The term normally indicates the non compensatable part of sensor error compensation that will propagate as navigation error.

Monolithic Integration: In micromechanical sensor fabrication when the sensor electronics is realised along with the sensor structure, the process is called monolithic integration.

Multi-axis Sensor: A sensor that is designed to measure input acting along or about more than one axis at a time.

Multisensor: A sensor that measures both linear acceleration as well as angular rate.

Null: It is the electrical output observed typically in a sensor when the physical input along its sensitive axis is zero.

Nutation Frequency: A frequency that is observed in a two-degree of freedom gyro that characterises periodic wobbling motion of the rotor spin axis that results from a transient input to the rotor.

Optical Gyro: A gyro that works on the principle of Sagnac effect by detecting the phase difference or the frequency difference between the counter propagating light beams.

Optical Path Length: The characteristic path length that is traversed in a single pass by a optical beam, taking into account the index of refraction of the medium supporting the propagation.

Photons: Light particles, which have no rest mass in the usual sense or which define quantum of electromagnetic wave energy.

Polarisation Maintaining Fibre: A single mode optical fibre that preserves the plane of polarisation of light coupled into it as the light beam propagates through the fibre length.

Polariser: This is an optical filter that converts unpolarised light into polarised type. This type of filter finds use in fibre-optic gyro.

Precession: It is the rotation of the spin axis in a spinning rotor gyro when a torque is applied perpendicular to the spin vector.

Pulse Rebalance: A servo control scheme that uses discrete torque pulses for generating rebalance torque in accelerometer or gyro.

Quantisation: The digital output of inertial sensor that changes in discrete steps as the input is varied continuously.

Random Walk in Angle: The angular error build up with time in a gyro due to white noise in angular rate. The error is typically expressed as $^{o}/\sqrt{h}$.

Reciprocal Path Length: A reciprocal optical path length is one whose properties are independent of the direction of propagation of light waves such as clockwise and counterclockwise light waves in fibre coil gyro.

Ring Laser Gyro: A gyro that instruments Sagnac effect by measuring the frequency difference of an active ring laser resonator.

Sagnac Effect: It is a rotation induced optical path length difference in inertial space between electromagnetic waves that counter propagate around a closed path.

Shot Noise: It is the noise seen in the output current of a semiconductor based photodetector that limits the rate detection threshold in a fibre-optic gyro.

Shupe Effect: It is a thermal gradient effect across a fibre gyro coil to produce bias in the gyro output.

Superluminescent Light Emitting Diode: It is a p-n junction semiconductor emitter based on stimulated emission with amplification, but insufficient for feedback oscillation to build up. It is used as a source in fibre-optic gyro.

Surface Micromachining: The term applied to a micromachining process that involves the use of sacrificial layers to produce elements largely confined to the vicinity of the silicon surface. Silicon substrate acts as rigid support base.

Tunneling Effect Inertial Sensor: Inertial sensor that uses quantum electron tunneling phenomenon to detect the sensing motion of the accelerometer or the gyro.

Vibrating Beam Accelerometer: A linear accelerometer whose no load beam vibrating frequency changes under acceleration that is proportional to input acceleration.

Index