A Point Cloud Based HRTF Technique for Analysing the Effects of 3D Sounds

Author

MUHAMMAD USMAN

NUST201464532MCEME35514F

Supervisor

DR. KHURRAM KAMAL

DEPARTMENT OF MECHATRONICS ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

JULY, 2017

A Point Cloud Based HRTF Technique for Analysing the Effects of 3D Sounds

Author

Muhammad Usman

NUST201464532MCEME35514F

MS-82

A thesis submitted in partial fulfilment of the requirements for the degree of

MS Mechatronics Engineering

Thesis Supervisor:

DR. KHURRAM KAMAL

Thesis Supervisor's

Signature:_____

DEPARTMENT OF MECHATRONICS ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY, ISLAMABAD

JULY, 2017

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student
Muhammad Usman
NUST201464532MCEME35514F

Signature of Supervisor
Dr. Khurram Kamal

Declaration

I certify that this research work titled "*A Point Cloud Based HRTF Technique for Analysing the effects of 3D Sounds"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student

Muhammad Usman

NUST201464532MCEME35514F

Copyright Statement

Acknowledgements

First and foremost, I thank Allah (SWT) for endowing me with health, patience, and knowledge to complete this work. Without health and facility given by Allah SWT unlikely I cannot done the thesis successfully. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual was Allah's will, so indeed none be worthy of praise but Allah.

I am grateful to my parents and sister who always helped and supported me during my education. Without them, I would never be able to complete this thesis. I am profusely thankful to my beloved parents for everything they have done for me.

I express my deepest gratitude to my supervisor Dr. Khurram Kamal for his always encouraging behavior. Before him, research was a boring and tough subject for me. He taught me how to conduct research effectively. His dedication to the teaching and research set a new standard for me. Throughout the thesis, he was more of a mentor than a supervisor to me.

I would also like to express my special thanks of gratitude to Dr. Javaid Iqbal and Dr. Umar Shahbaz khan for being on my thesis guidance and evaluation committee. I am especially thankful to them for their support and cooperation.

I would like to express my gratitude to all the friends specially Sohail Akram, Abdullah Rasheed and (Soon-to-be Dr.) Tayyab Zafer who have rendered valuable assistance to my study.

I acknowledge, the effort being made by the department in providing me all the facilities to attain the highest standards of excellence & the efforts of the faculty members in transferring their precious knowledge to us.

Dedicated to my exceptional parents and sister to express my unspoken love

"You don't really understand human nature unless you know why a child on a merry-go-round will wave at his parents every time around, and why his parents will always wave back" ~William D. Tammeus

ABSTRACT

3D Multimedia technologies such as virtual and augmented reality are in major focus nowadays due to advancements in multimedia technology. These systems are required to produce sound effects in 3D in order to accurately simulate the real world. This thesis proposes a novel system for realize 3D sounds in accordance with the head position for virtual reality and gaming purposes. The technique employed an Inertial Measurement Unit (IMU) mounted on a headphone and a depth sensor to calculate azimuth and elevation of head. Extended Kalman Filter is used to fuse the head tracking data from IMU and Kinect measurements. The estimated head position is fed to the CIPIC Head Related Transfer Function (HRTF) database in order to produce 3D sound effects. Results show a promising future for the proposed technique in 3D sound realization.

Realistic sound generation is an integral part of Virtual Reality (VR) and Augmented Reality (AR). Conventionally, sound systems are designed in such a way that sound realization is independent of the direction of the listener. Normally, this is not critical, but in the case of VR and AR, sound localization becomes compulsory [1-2]. The listener should be able to distinctly perceive the exact location and orientation of any sound source inside the simulated environment. Hence, the sound generation must account for all 6 degrees of freedom in nature by taking into account three positions and three orientations of the receiver with respect to the source. Although the modification of a sound based on the relative position is important for large simulated spaces, for small, confined areas the relative orientation takes precedence. Therefore, this work tackles the issue of relative orientation based modification of source sound. This is called realistic 3D sound generation in this paper. Commercially available virtual reality products, like Samsung Gear and Oculus Rift, have spawned many new researches. One of the major problems in VR systems is generating 3D sounds according to the listener's head position in 360 view of a simulation and orientation with respect to the simulated environment. The Challenge is to generate realistic sounds onto virtual world

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

This chapter gives a brief review of the 3D sounds, various methods with sounds and aims and motivation behind selecting the 3D sound research.

1.1. Introduction to Thesis

Multimedia industry is increasing day by day. Everyday there is a new product of multimedia in the market. 3D sound systems are not new in the world. A head tracking based 3D sound multimedia system can create a new trend in the market.

3D Multimedia technologies such as virtual and augmented reality are in major focus nowadays due to advancements in multimedia technology. These systems are required to produce sound effects in 3D in order to accurately simulate the real world. Realistic sound generation is an integral part of Virtual Reality (VR) and Augmented Reality (AR). Conventionally, sound systems are designed in such a way that sound realization is independent of the direction of the listener. Normally, this is not critical, but in the case of VR and AR, sound localization becomes compulsory. The listener should be able to distinctly perceive the exact location and orientation of any sound source inside the simulated environment. Hence, the sound generation must account for all 6 degrees of freedom in nature by taking into account three positions and three orientations of the receiver with respect to the source. Although the modification of a sound based on the relative position is important for large simulated spaces, for small, confined areas the relative orientation takes precedence. Therefore, this work tackles the issue of relative orientation based modification of source sound. This is called realistic 3D sound generation in this paper. Commercially available virtual reality products, like Samsung Gear and Oculus Rift, have spawned many new researches. One of the major problems in VR systems is generating 3D sounds according to the listener's head position in 360 view of a simulation and orientation with respect to the simulated environment. The Challenge is to generate realistic sounds onto virtual world.

Various methods of head tracking have been developed for head tracking in the immersive application that can be utilized with virtual reality (VR) and Augmented Reality (AR). Head tracking is mostly done either in indoor applications or outdoor applications. The technique

developed in the thesis is used for the indoor applications. Most common use of head tracking in immersive application is to implement the view of the virtual world with respect to the position in real world. The technique developed in the thesis employs the fusion sensor based head tracking that can be used to update the roll, pitch and yaw value in Head Related Transfer Function.

In multimedia the first ever product including the spatial sounds was the stereo sound systems. The basic concept behind the stereo system was to place left loudspeaker on left side while the right on the right side.

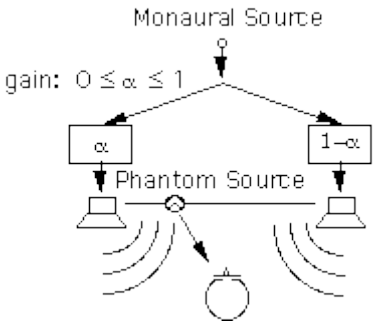


Figure 1.1. Basic Concept of Stereo System

Basically the stereo sound systems create a time lag between left and right speakers in such a way that a listener percept it as a spatial sound. If the sound on the left speaker is delayed 15 to 20 ms as compared to the right, the listener will percept the sound localization on the right side.

The tool for generating the digital 3D sound is called Head Related Transfer Function. The HRTF records the location-dependent spectral changes that happen when the audio wave propagates from an audio source to the ear drum. These spectral changes occur due to diffraction in waves by the head, torso and outer part of ear. The character of spectral changes depends on the azimuth, elevation and distance between listener and sound source.

The thesis is organized in the following sections. Chapter 3 explains the previous work done on the field of head tracking. The techniques that are already being used, i.e., the inertial, acoustical, and sensor fusion based tracking is explained. Chapter 4 describes the basic methodology of the technique. Chapter 5 and 6 explain the experimentation, tool and

techniques used in the thesis and in the chapter 7 explains the conclusion and future work discussion.

1.2. Summary

- Multimedia is a profitable industry. By producing a head tracking based 3D sound system, a new product can be introduced which can easily capture the market trend.
- Various methods for head tracking already have been developed. The head tracking is mostly done for the 3D view in immersive applications. Head tracking based 3D sound systems are still in development process.
- Sensor fusion based head tracking can be utilized in gaming application for estimating the exact location of the object and to create a realistic sound realization.

2. LITERATURE REVIEW

This chapter provides a brief discussion about the head related Transfer Function (HRTF) and a detailed discussion about the tracking techniques. In the last part of the chapter a brief comparison of all tracking techniques has been done.

2.1. Head Related Transfer Function

Human body contains two ears but our listening system has ability to listen in all three dimensions. Whenever a listener listens any sound around the location of sound can be estimated, this is because the brain and inner ear work together to estimate the position of sound source. This ability makes human to track anything with the sound makes any listener to localize sound source even without light. Humans percept the sound source by the difference of the monaural sound cues received from both ears these monaural cues arrive human ear at different intensity and different time with these intensity difference and arrival time difference human auditory system human brain estimate the location of the sound source. The ability of human brain to estimate the sound source location with respect to ear location can be defined by an impulse response that is called Head Related Impulse Response (HRIR). While Head Related Transfer Function (HRTF) is the Fourier Transform of HRIR.

HRTF of both ears tells us the filtering of the source of sound (t) before it is estimated at left ear as $yL(t)$ and at right ear as $yR(t)$.In linear system analysis the Transfer Function of any system is described as the ratio of output and input functions. The Transfer Function (f) of a system at frequency f is given as

$$(f) = Output(f) \,/\, Input(f) \qquad \textbf{(2.1)}$$

2.2. Magnetometer Sensors Based Head Tracking

Flux-gate sensors commonly known as Magnetometers basically work on the principal of faraday's law. According to the Faraday's law the potential difference (or current) can be created within a loop with the existence of varying magnetic field. A magnetometer consists of a magnetic core, a drive winding and a pair of sense windings. The core has the drive winding wrapped around it. The sense windings are often wound flat on the outside of the core. When un-energized, a fluxgate's permeability 'draws in' the Earth's magnetic field. When energized, the core saturates and ceases to be magnetic. The magnetic field of earth is released from the core as this switching occurs, ending in a small induced potential that is proportional to the strength and direction of the external field.

Magnetometers or commonly known as magnetic compass commonly used for heading references, are now also available for Augmented and Virtual reality applications. A hostile magnetic environment in surrounding environment of the user can seriously damage the accuracy of magnetometer's navigation.

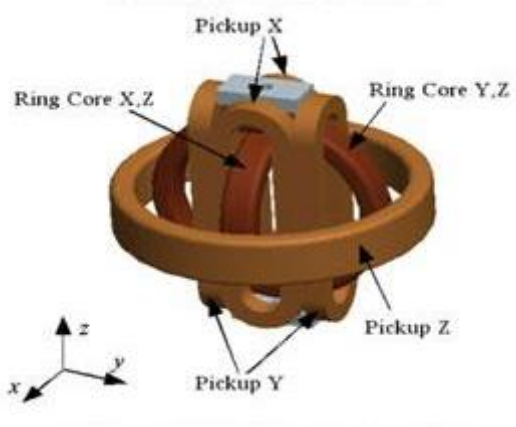


Figure 2.1. 3D Fluxgate Magnetometer [3]

E. F. Helbling et al. [4] use an on-board magnetometer for yaw and pitch calculation of movements of the robotic bee. For Calibration of pitch angle, a potentiometer is connected at the end of the axle as shown in Fig. 2.1 As the magnetometer is made to rotate at any particular angle, the voltage output received is proportional to the magnetic field of the sensor. Similar process is done for yaw calculations.

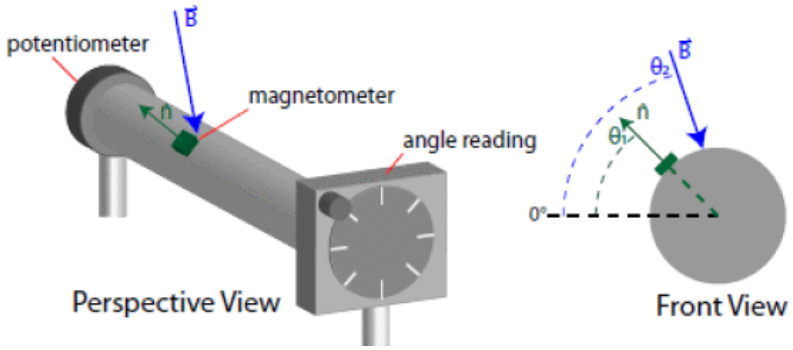


Figure 2.2. Calibration of magnetometer [4]

2.2. Inertial Measurement Unit based Head Tracking

All accelerometers usually have a zero-g level. Zero-g level is an offset or can be said as a bias level. This is the reading the sensor provides with no acceleration. The zero-g value can be found in data sheet of particular accelerometer. Accelerometers have also a sensitivity factor. Sensitivity is basically the measure of change between output angles and rotation of

the accelerometer. Sensitivity is usually expressed in mv/g. Divide the zero-g level with sensitivity value to get final readings. However, since the conversion to angles uses ratios of the acceleration vector components, this factor divides out. The calculated rotations along the X-axis (φ), Y-axis (-ρ) and Z-axis (θ) can be expressed as:

$$\rho = \arctan(A_X / \sqrt{A_Y^2 + A_Z^2})$$
$$\varphi = \arctan(A_Y / \sqrt{A_X^2 + A_Z^2})$$
$$\theta = \arctan(\sqrt{A_X^2 + A_Y^2} / A_Z)$$

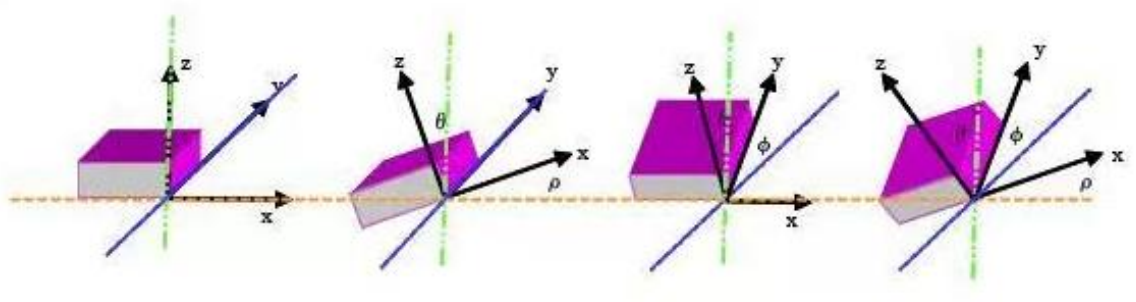


Figure 2.3. Tilt angles of accelerometer [5]

The tilt angles along X-axis express rotation around the Y-axis, and the tilt along the Y-axis express (negative) rotation around the X-axis. Computing the angles from gyroscope is bit difference. Since the gyroscope calculates the angular velocity, not angular orientation angles itself. For calculating angles, we have to initialize the gyroscope with some known point (can be calculated from accelerometer). After that, we have to calculate the angular velocity along X,Y and Z axis at time interval of (Δt). Then change in angle can be calculated as

$$\omega \times \Delta t = change\ in\ angle \quad \textbf{(2.4)}$$

The new angle will be the sum of original position and the change in angle. While taking separate angular data, either from accelerometer or from gyroscope can cause problems. As, the gyroscope provides accurate date in short intervals of time , but it causes drift in its angles over a long interval of time. While, accelerometer gives accurate readings over a long interval of time, but it is noisy in short terms. The standard way combine these two inputs is "Kalman Filter" , which is quite complicated to use. Luckily, there exist a simpler method for combining the gyroscope and accelerometer data, known as "Complimentary Filters". The formulae to calculate the angles from gyroscope and accelerometer data are given as:

$$\textbf{Filtered Angle} = \alpha \times (\textbf{Gyroscope Angle}) + (1 - \alpha) \times (\textbf{Accelerometer Angle}) \quad \textbf{(2.5)}$$

$$\alpha = \tau/(\tau + \Delta t) \quad \textbf{(2.6)}$$

(Gyroscope Angle) = (Last Measured Filtered Angle) + (ω×Δt)Δt = sampling rate **(2.7)**

τ = time constant which should have at least greater value than timescale of typical accelerometer noise.

J. H. Kim et al. [6] uses inertial measurement unit of 3D tracking of hand motions. Data from accelerometer and gyroscope are taken for 3D angles. Then the data is fed to the Kalman filter for the stability. The authors found out that the data from a tri axis accelerometer and three single axis gyroscopes can be fused using Kalman filter to generate more accurate and stabilized values of yaw, pitch and roll of elbow joint.

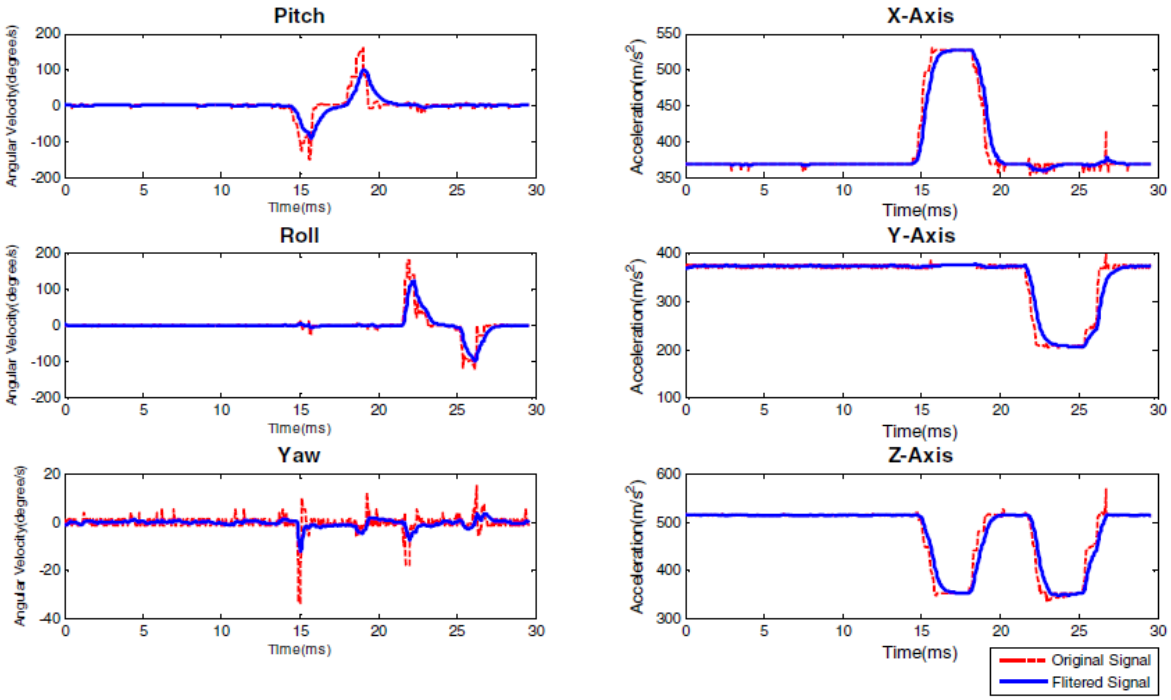


Figure 2.4. Yaw, Pitch roll and linear movements with filtered signals from accelerometer [6]

Fig 2.4 Shows pitch roll and yaw and movement along X, Y and Z axis measured by Accelerometer and also showing the filtered angle comparison with original signal.

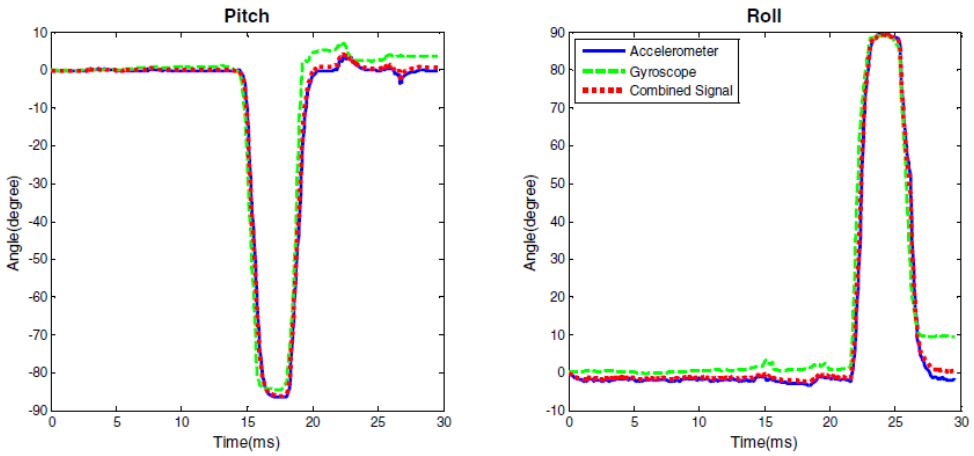


Figure 2.5. Estimated roll and pitch from fusion of gyroscope and accelerometer [6]

Fig 2.5 shows the filtered angle by averaging both of the accelerometer and gyroscope values. The yaw values can only be measured from gyroscope values that's why there is no combined signal of yaw in the fig. 2.4.

P. Gui et al. [7] performs the fusion of tilt angles measured from accelerometer and gyroscope of an inertial measurement unit by kalman and complementary filter. The authors found that both kalman and complementary filter showed promising results for stabilization of yaw, pitch, and roll values obtained from gyroscope and accelerometer. The algorithms performed efficiently both in static and in dynamic states. Given the fine-tuned filter co-efficients , complementary filter showed more efficient results as compared to the Kalman. Moreover, the complementary filter does not require any state variables so it is computationally economical.

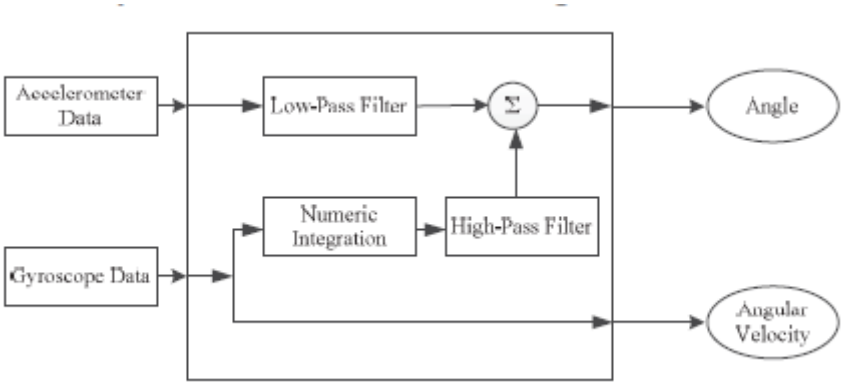


Figure 2.6. Complementary filter Algorithm [7]

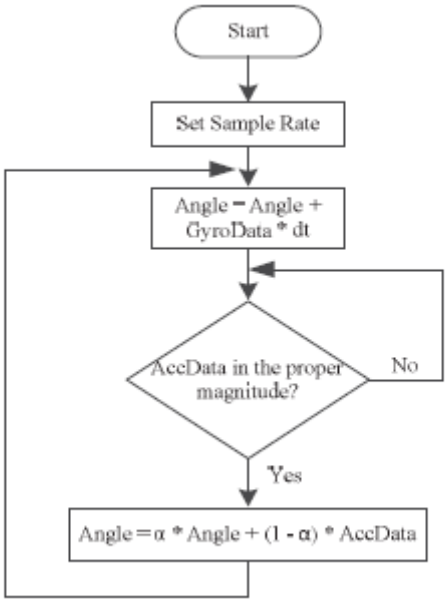


Figure 2.7. Complementary filter flow chart [7]

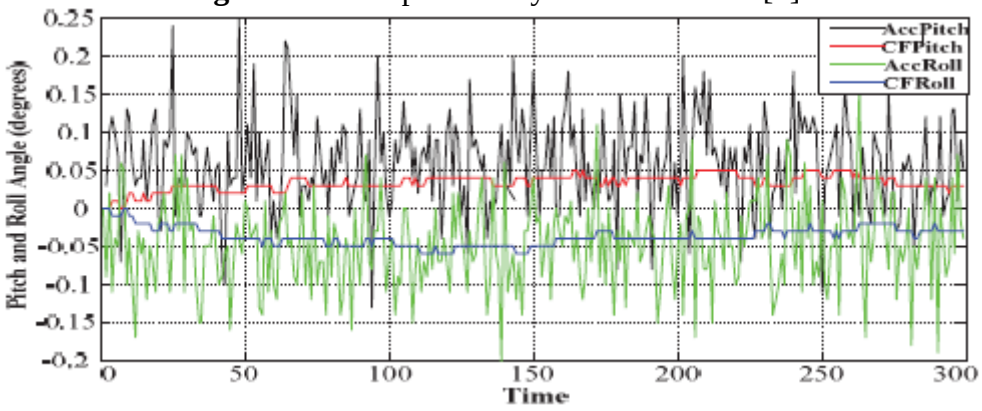


Figure 2.8. Complementary filter static test [7]

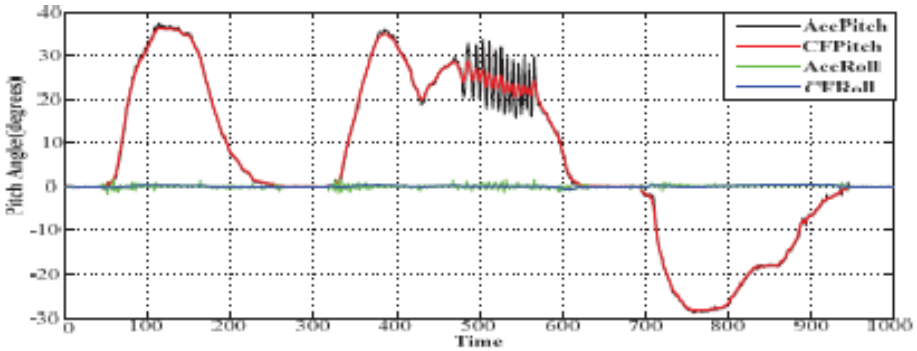


Figure 2.9. Complementary filter Dyanamic test Pitch [7]

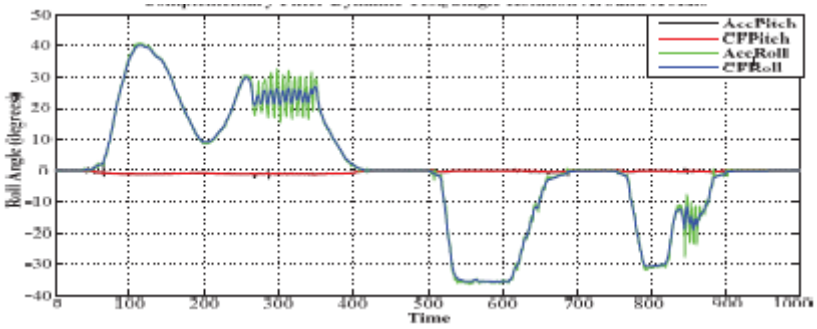


Figure 2.10. Complementary Filter Dynamic Test Roll [7]

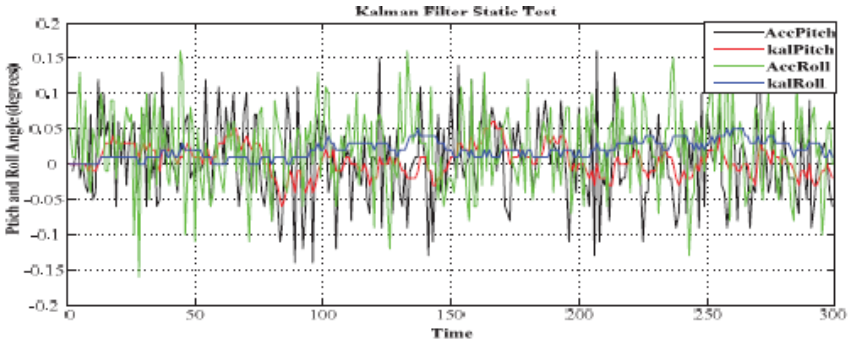


Figure 2.11. Kalman filter static test [7]

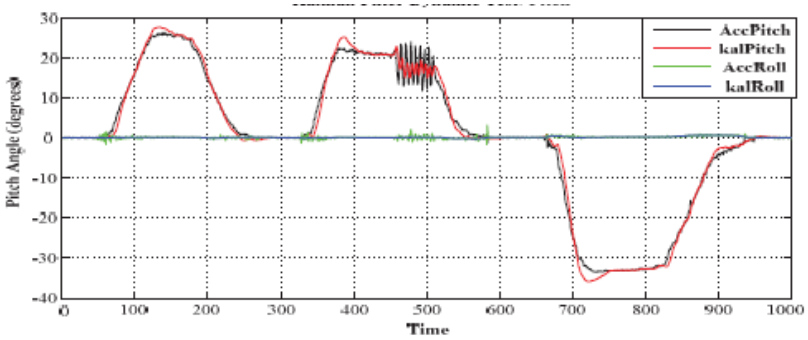


Figure 2.12. Kalman filter Dynamic Test Pitch [7]

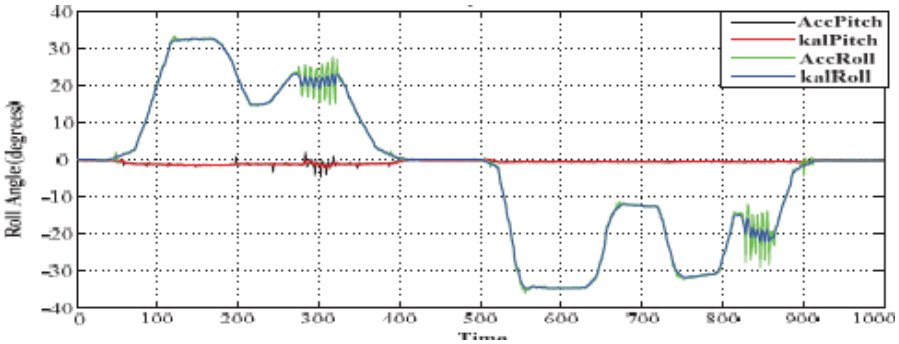


Figure 2.13. Kalman filter Dynamic Test Roll [7]

J. H. Kim et al. [8] explains the feasibility of using the low cost inertial sensors for analysing the human motion. Motion analysis of human body was done using Sony Move remote and Nintendo Wii. Tracking data obtained from both is then compared with a standard IMU

sensor developed by Xsens known as Xsens MTx . The authors found that MTx sensor with static angle accuracy of **0.3°** can easily be used to measure movement up to 10 cm with an error of 0.5cm. The inertial sensor can be used to measure the elbow movement up to **150°** with error of just **1°.** Moreover, it was found that Sony Wii can be used to track the motion of human body up to 10cm having an error of approximately 0.2 cm. Furthermore, the Nintendo Wii can also be used to measure the upper body movement with a great accuracy.

E. Foxlin [9] uses an inertial sensor for head tracking for virtual reality applications. The head tracking inertial sensor used in the technique consists of a gyroscope. The head tracking data obtained from gyroscope was taken as the sensor input in kalman and the measurement from gyroscope was then compared with the readings from an inclinometer using complementary Kalman filter and separate-bias kalman filter. The authors found that by using measurement and process noise matrices the output of kalman filter diverges just after a few moments when the sensor is still. It was then concluded by repeated experimentation that to prevent the divergence in the output the diagonal elements in the measurement noise.
covariance matrix should not be less than '1'. Moreover, by proper adjustment of the values of noise covariance matrices and time step values used in the technique, the output from kalman can be made stabilized.

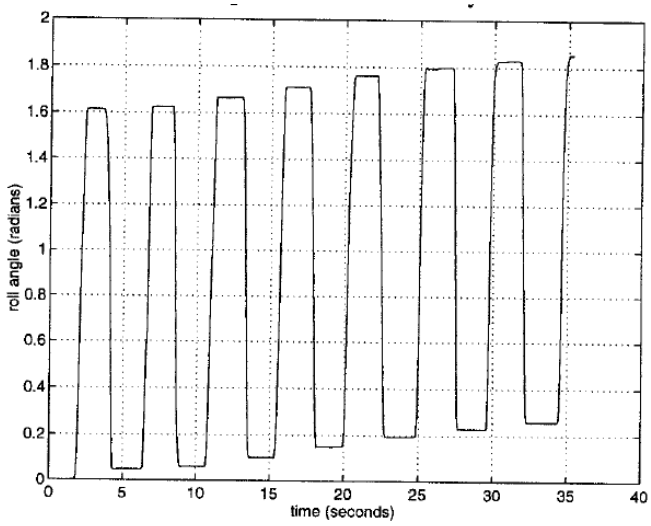


Figure 2.14. Test run without complementary Kalman Filter [9]

Two show the exact estimations two datasets were collected. One for the separate-bias kalman filter and one for complementary kalman filter. It was found that without complementary kalman filter the output is noise free but contains some noise. In the case of complementary kalman filter the output is drift free but contains noise.

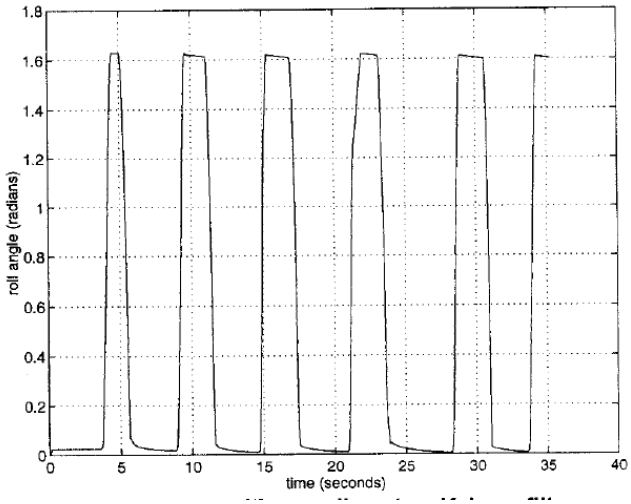


Figure 2.15. Test run with complementary Kalman filter [9]

2.4. Head Tracking and Acoustics

Acoustics is a branch of science that deals the study about the types of mechanical waves in any type of medium. Vibration, sound, ultrasound and infrasound are some examples of acoustic signals. This portion present a brief explanation of the application of an acoustic signal in head tracking.

Eric Foxlin [10] proposed a new technique for tracking known as "The CONSTELLATION tracking system", employs acoustic for position measurement rather than RF. Its basic principle is exactly same like inertial navigation system the only difference is that it operates for indoor navigation.

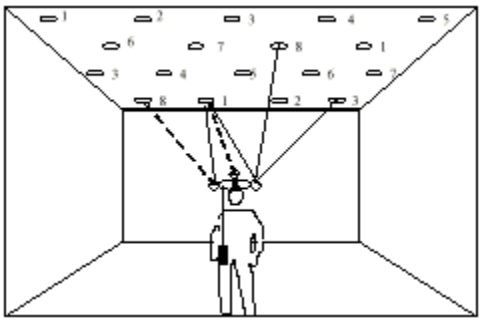


Figure 2.16. The Constellation System [10]

Fig 2.16. illustrates the basic concept of constellation system, which is designed for Head Mounted Display (HMD) in applications of VR and AR. The HMD is equipped with an inertial measurement instrument and three ultrasonic range-finder modules. The range finder modules communicate with the inertial sensors to find out the exact location of the tracked object.

Eric Foxlin utilizes the InertiaCube [11] for finding the linear acceleration and angular changes of the tracked object.The hardware overview of the technique is shown in the fig below.

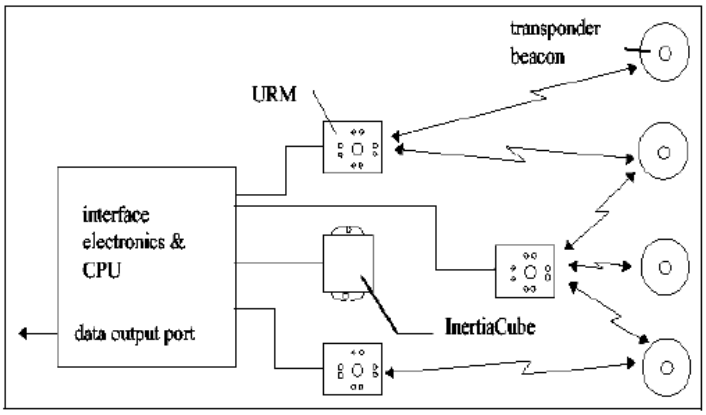


Figure 2.17. Hardware Overview [11]

Barnard, M et al. [12] use acoustic signal for head tracking purposes, the authors used direct to reverberant speech energy ratio (DRR) in a reverberant room for tracking the head positions. The authors found out that the speaker facing towards the microphone showed the higher DRR value, while the speaker pointing away from the microphone showed lower value of DRR.

R. N. Aguilar et al. [13] uses ultrasonic acoustic system for tracking several points on a human body. The targets points on human body are estimated using acoustic time of difference of arrival (TDOA) from multiple transmitters. The signal processing for estimating the TDOA from different sensors is done using Hilbert Transform, zero-crossing detection, parabolic interpolation, cross co-relation and post filter analysis. The authors found that while using the TDOA technique the effect of speed of sound is finished in tracking any part of the human body.

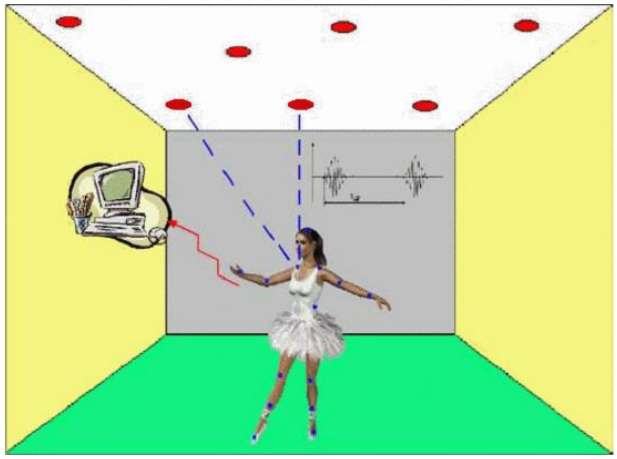


Figure 2.18. An overview of an ultrasonic tracking system [13]

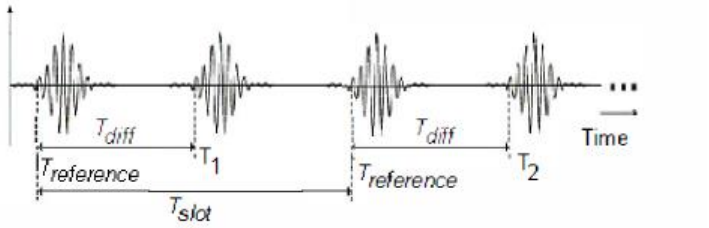


Figure 2.19. Sequence of waves coming out of transmitter [13]

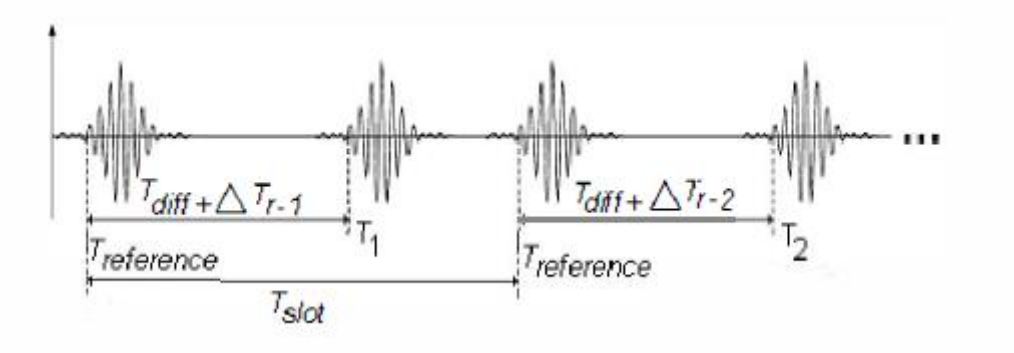


Figure 2.20. Sequence of waves at receiver end [13]

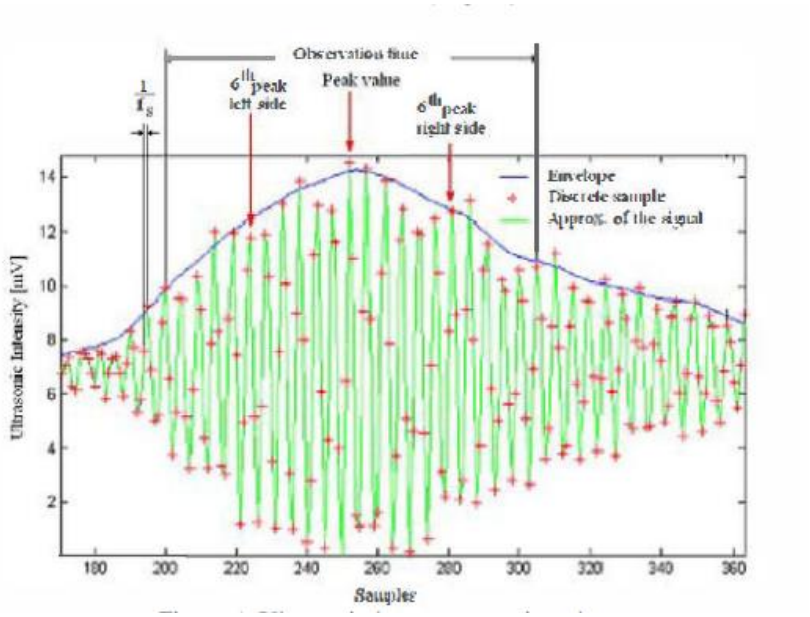


Figure 2.21. Ultrasonic Burst at Microphone [13]

2.5. Head Tracking Using the Optical Devices (Vision Based Tracking)

In optical head tracking, there are a lot of sensors available, from ordinary video cameras to LEDs, to detect either ambient light or light emitted under control of the position tracker. Infrared light (depth sensing) is also frequently utilized to prevent interference with other activities.

The tracking based on the optical sensors provides a tremendous amount of data regarding the person that is being tracked. Potentially the visual sensors can be said as the strongest competitor among all the sensors. The problem of localization has attained a remarkable attention from the researchers in field of tracking due to their high accuracy. The basic parts of tracking process are:

- Environment Representation
- Models for Sensing
- Algorithms for localization

Relative or the absolute position of the sensors are obtained in most of the tracking techniques. The technique developed may vary depending upon the nature of sensor, the geometric model and environment representation. A vision based sensor or the multiple

image based sensors should be able to obtain the features of an object (specially head in our case) for the tracking purposes.

2.6. Camera Model Localization

For estimating the position and localization and orientation of the geometric models of cameras have great importance. Pinhole camera model is one of the most used model in localization as shown in the fig. 3.22. The cameras with optical lens can be modelled by using pinhole camera model. The coordinates system employed by camera is three dimensional Cartesian coordinate system.i.e. (X,Y,Z) while the image is represented in two dimension (X,Y) coordinate system. A real life three dimensional object is represented onto a two dimensional image.

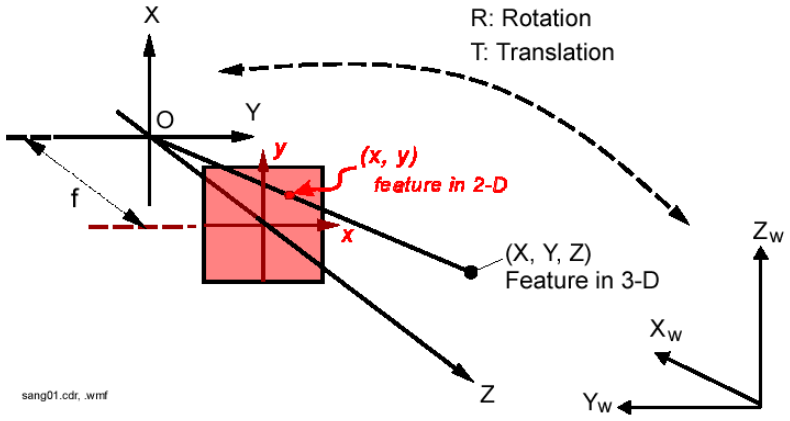


Figure 2.22. A real life three dimensional object is represented onto a two dimensional image [14]

Although a lot of information in pinhole camera model get distorted in such projection, the angle of the object and orientation of the object can easily be obtained from the obtained image if the focal length of the image is well known. The physical size of image plane [15] is obtained from the internal or intrinsic parameters of the image, .i.e. focal length, redial lens distortion factor, and image scanning parameters are known as the intrinsic parameters of an image. The image orientation and position can be described by six parameters, three describes position and three describes the orientation, and they are known as intrinsic and extrinsic parameters of camera. The relationship between real world coordinates and camera coordinates is represented by these parameters. The Main challenge to tackle is to find out the exact orientation and position of the object being tracked by extracting features from an image or multiple images. It is understood that a single feature cannot provide much

information to estimate the object localization. So, it is understood that we need multiple images for object tracking.

The Environment representation can be performed by extracting simple features like points or lines and for the sake of more accuracy this can be performed by extracting complex features like three dimensional models of environment surroundings or objects being tracked. This section explains the object tracking that is based on the extraction of simple land mark features.

A camera model localization based head tracking is used by C. Rougier et al. [16], the authors use a monocular cam for falls detection of elderly people using 3d head tracking. The head is detected using the features extraction. Then the falls detection is done by obtaining the trajectory of the head. The authors found out that the single camera can be used to detect falls and can also be used for estimating 3D velocity characteristics without any use of wearable sensors.

Similar technique is used by R. B. Mapari et al. [17] the authors of the paper use leap motion sensor which is basically a monochromatic stereo cam for the purpose of real time pose tracking. The stereo vision cam is used for extracting features. The features extracted from stereo vision is then used for head tracking. The technique developed by the authors utilizes leap motion for real time pose estimation for sign language recognition. The authors found that for exact recognition the leap motion sensor should be placed at an inclined surface. Moreover, the technique can further be improved if the features extraction is further accompanied with trained neural networks.

Moreover, F. A. Kondori et al. [18] also employs similar technique for the head tracking purposes the authors use Microsoft Kinect sensor which is basically a depth sensor, for real time head pose estimation. The features are extracted by depth images and then these features are used for estimating the real time pose estimation. The authors found that the technique can estimate the head pose efficiently in all 6 degree of freedom. Moreover, the algorithm developed can also be used for detecting the pose of multiple people in field of view of the Kinect sensor.

K. Satoh et al. [19] uses a bird eye view camera and a gyroscope for head tracking purposes. A gyroscope is mounted of a Head Mounted Display (HMD) and a bird's eye view camera perceives the HMD from a fixed point. The HMD consists of a marker, and a gyro sensor. The Gyro sensor intends the position of the user, so that the head pose parameters can be

reduced. While the other parameters are obtained from the bird eye view camera that observes the marker on the head mounted display.

Figure 2.23. Experimental setup used in the technique [19]

Shay Ohayon et al. [20] uses a single camera for head pose estimation. Head pose is obtained by using camera pose estimation formulation. Head tracking model is obtained by 3D features points.

2.7. Model Based Approaches

The other group of researchers that are working on the problem of tracking employ the model based or model-matching technique for tracking any object. The geometric features of the environment surrounding the moving object are extracted to compute the exact location of the object. Then the results received from sensing tracker, i.e., camera or depth sensor, is then matched with the geometric features.

In model base approaches, the data acquired from the sensor is then matched with a model of the environment. If the features extracted from the sensor and the object being tracked are matched, then it is an easy task to compute the exact position of the object.

The geometric model commonly employed in such techniques three dimensional model of buildings, indoor structure and can be floor models also. For estimating the localization of the moving object the two dimensional image from sensor should be able to extract the features from the environment. Then the information obtained from environmental model and sensor image are matched and object is tracked from that. The main challenge is that the environment geometric features are in three-dimensional models while the image is in two

dimensional observations. So, challenge is to recognize the object basically and whole problem in a nutshell is described as

- Identifying Objects
- Estimating the orientation of the identified objects

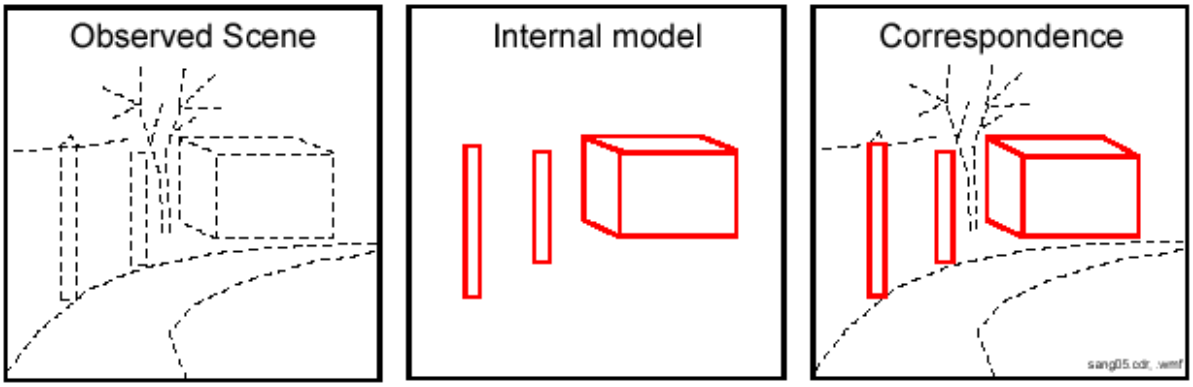


Figure 2.24. Finding the matched featured between the geometric model and sensor observations

2.8. Sensor Fusion Based Tracking

Normally, any tracking sensor may track the moving object with high accuracy, but there are always some drawbacks in using any type of sensor. By fusing two sensors, the accuracy of the whole tracking system can be improved. Furthermore, the accuracies of both sensors combines and make the overall system more accurate. For example, the inertial sensors have a drift problem and the optical sensors have the field of view problem. So, by fusing an inertial and optical sensor the overall accuracy of tracking system can be improved.

Main challenge in fusing multiple sensors data is to select an appropriate fusion algorithm for estimating the exact position and orientation of the object being tracked. The technique developed in the thesis utilize “Extended Kalman Filter” for fusing the data obtained from an inertial sensor and depth sensor.

A. O. Ercan et al. [21] fuses an inertial sensor and a camera for head tracking in immersive applications. The inertial sensor contains the accelerometer only. The accelerations data obtained from accelerometer is then fused with the visual data obtained from the camera using the “extended kalman filter”. The performance of overall system is measured by taking accelerometer as control input in extended kalman filter and camera and inertial both as measuring input, and in next step the accelerometer was taken as the

measurement input. The authors found that if accelerometer is taken as the control input while the camera as the measurement input the overall system performance is better as compares to all control and measurement inputs combinations used in the system.

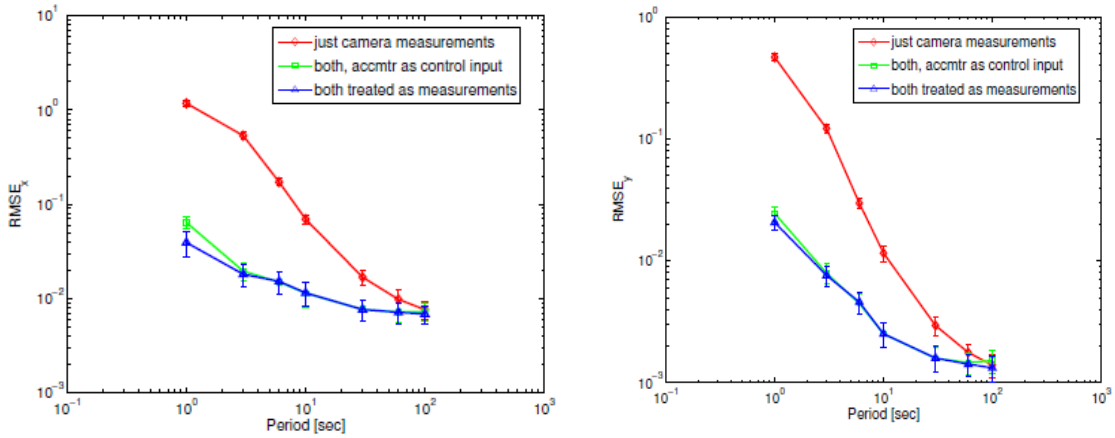


Figure 2.25. Standard deviation in C and Y direction in all three scenarios [20]

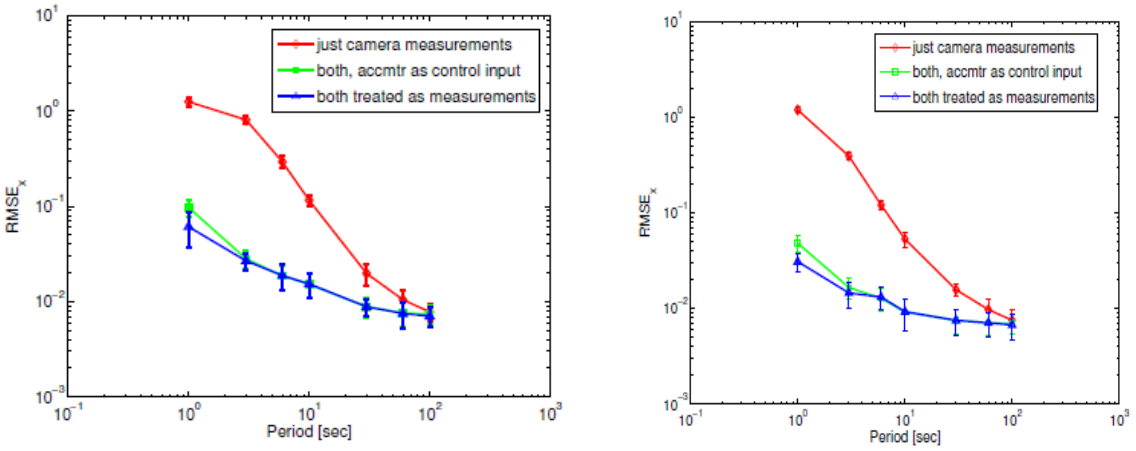


Figure 2.26. RSME tracking in X and Y direction in all three scenarios [21]

A. T. Erdem et al. [22] fuses the yaw, pitch and roll data obtained from the inertial sensor with data obtained from a moving camera for tracking the exact location of the camera in motion. Extended kalman was employed for fusing the camera and inertial sensors movements. Furthermore, the combination of measurement and control inputs were changed with accelerometer, gyroscope and camera readings. The authors found that the best combination for acquiring 3D data was taking both accelerometer and gyroscope data as the measurement input. Furthermore, the cases of gyro as measurement, gyroscope and accelerometer as control and accelerometer and gyro as measurement cases were compared it was found that accelerometer and gyroscope as control input case and taking gyroscope and

accelerometer both as measurement inputs case showed the same results and both provide better performance when the camera is in fast speed.

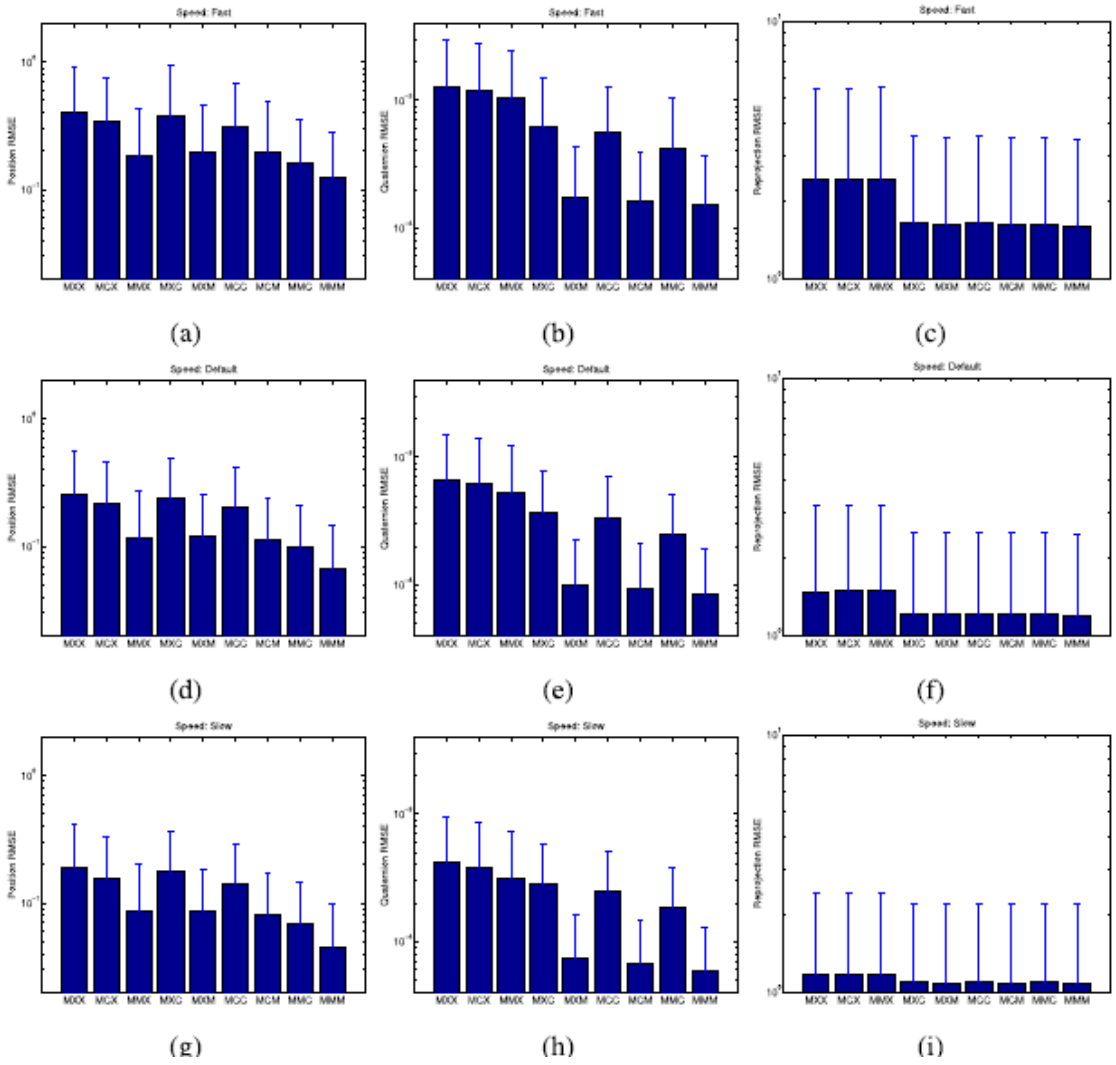


Figure 2.27. RSME measurements for all cases [22]

2.9. Thesis aims and objectives

The scope of thesis is quite broad, but according to the literature review the scope of the thesis can be defined as follows

- To develop a novel technique to measure 3D head tracking that can be used to update location of both ears in HRTF to change the effects of 3D sounds.
- To study state of the art techniques of studying 3D head Tracking.

Table 2-1: A brief comparison between the techniques used in head tracking applications

Method	Description	Strength	Weakness
Mechanical	This system depends on a physical link between a fixed reference point and the target.	Efficient in accuracy Low lag No issue of magnetic interference No issue of field of view	Not easy to interface with any electronic system.
Magnetic	The technique utilizes a set of magnetic coil in transmitter to produce magnetic fields. The sensor in receiver calculate strength and angles of magnetic field.	Economically efficient Accurate No field of view problem Good noise immunity	Field distortion due to magnetic or metal surfaces. Electromagnetic waves interference due to radio signals Accuracy decreases with increase in distance from receiver
Source less , Non-inertial	Use passive magnetic field sensors relative to the earth's magnetic field the values of yaw, pitch and roll are measured	No transmitter required Can be embedded to any system Low cost	Difficult to mark movement between magnetic hemisphere
Optical	Available in many types, From a camera to LED.	Easily available No issue of Magnetic interference Efficient Highly accurate	Field of view is the biggest problem Light intensity is also an issue Weight Expensive

Inertial	Consists of gyroscope and accelerometers. Orientation of the object containing the inertial sensors is measured by the angles measured by the gyroscope and accelerometer.	No range problem Quick No issue of magnetic interference Small in size Inexpensive	Drift Accurate only for high changes in angles
Acoustic	Commonly, three microphones are used as receiver and three sound sources are used as transmitter. The distance between source and sound is calculated by triangulation.	Cost effective No magnetic interference Light weight	Ultrasonic noise interferences Low accuracy Line of sight is a big issue

2.10. Summary

The chapter can be summarized as follows

- Many sensors based head tracking methods has been developed. The most widely used methods for tracking are inertial tracking and optical tracking.
- The inertial sensors have drift problems while the optical sensors have the problem of field of view.
- The drawbacks of both sensors can be minimized by the fusion of sensors.

3. METHODOLOGY

This chapter briefly describes the methodology of the technique developed. The flow diagram of the technique is shown in fig 3.1.

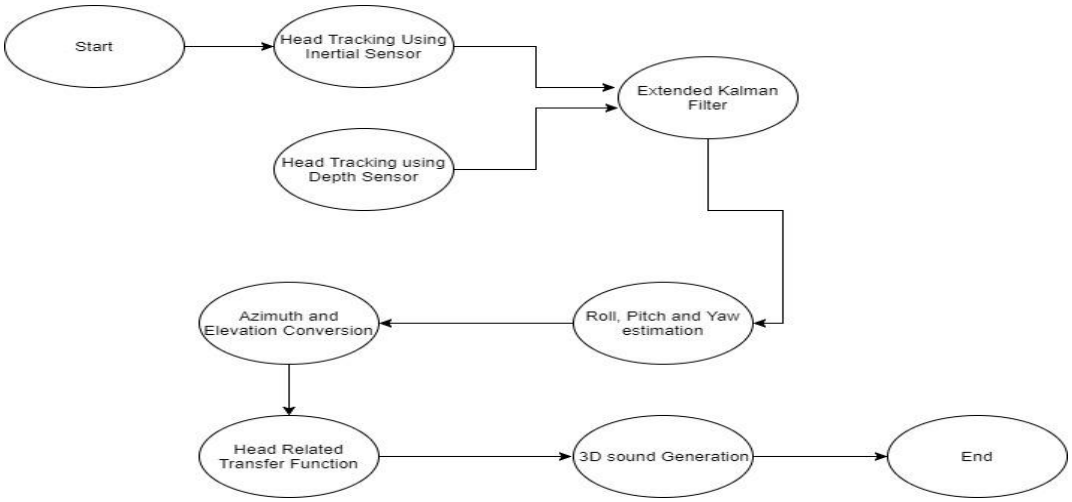


Figure 3.1. Flow chart of the methodology

First of all, head tracking is done by the inertial measurement unit mounted on the head. The inertial sensor consists of an accelerometer and gyroscope. The angle calculation from an inertial sensor is well explained in second chapter. The yaw, pitch and roll data is obtained from gyroscope and accelerometer and by applying complementary filter a fusion data of both accelerometer and gyroscope.

While the yaw pitch and roll data is obtained from the inertial sensor at the same time the head pose is also estimated using the Kinect sensor. Kinect is a depth sensor developed by the Microsoft. Kinect is basically a part of XBOX for interactive gameplay. The yaw, pitch and roll data is obtained from the kinect is fused with the data obtained from the inertial sensor by using the extended kalman filter. The extended kalman filter is well explained in the fourth chapter. The data from both sensors is fused by extended kalman filter in six simultaneous combinations of control and measurement inputs as follows.

1. Kinect as measurement and gyroscope as control
2. Kinect as measurement and accelerometer as control
3. Kinect as measurement and fusion of gyroscope and accelerometer as control
4. Gyroscope as measurement andkinect as control
5. Accelerometer as measurement and kinect as control
6. Fusion of gyroscope and accelerometer as measurement and kinect as control

The roll, pitch and yaw is obtained from the extended kalman output is then converted to the azimuth and elevation measurements. In our case the yaw measurement is considered as the azimuth while the pitch is considered as the elevation of the head pose.

Azimuth and elevation values of the head pose is then passed to the Head Related Transfer Function developed by the CIPIC . Head Related Transfer function is explained briefly in second chapter and derivation methods of HRTF are explained in detail in forth chapter. HRTF is basically a tool for generating the 3D sounds. In the end the 3D sound is generated by HRTF that correspond to the exact value of head pose.

3.1. Interaural Polar Coordinates

To explain the location of a sound, source relative to the listener's position, a coordinate system is required.

Since, the shape of head is approximately spherical, a spherical coordinate system is required. The standard coordinates in spherical system are azimuth, elevation and range. Unfortunately, there is not a single standard way to define these coordinates. In literature, different people define these coordinates in different ways. One of the most commonly used coordinates system is the vertical-polar coordinates system. Here, the azimuth is measured as the angle between the median plane and the vertical plane containing the sound source and z-axis. Furthermore, the elevation angle is measured in vertical direction up from the horizontal plane.

The technique utilizes an important alternative, i.e. the interaural polar coordinate system shown in Fig 3.2. In this system, the elevation is measured between the horizontal plane to a plane through the source and x-axis, this is defined as the interaural axis. While the azimuth is measured from the angle over from the median plane.

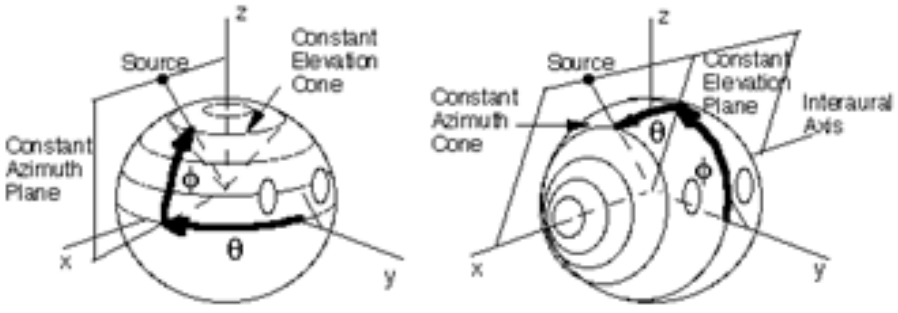


Figure 3.2. The Vertical Polar-Coordinates System and Interaural-Polar Coordinate System [23]

3.2. Summary

- Head tracking data is obtained from inertial sensor and Kinect.
- The tracking data is fused using the extended kalman filter using different combinations of control and measurement inputs.
- The tracking data obtained from the extended kalman is passed to HRTF to obtain the 3D sounds.

4. TOOLS AND TECHNIQUES

This chapter explains the scope of thesis in details and background theory regarding the tools and techniques used in the thesis is also well explained in this chapter. The chapter is sub-divided into three sections. The first one explains the Extended Kalman filter in detail, the second one explains the Head Related Transfer Function in details.

4.1. Extended Kalman Filter

As already explained that for head tracking an optical sensor i.e the Kinect is employed in the proposed method while head mounted inertial measurement unit is also measuring the head movement at the same time. It is well understood that every sensor has its own advantages in measurements and at the same time also have some limitations also. For example, the Kinect sensor has a limitation of field of view, while the inertial sensor has the limitation of the drift. So, to minimize the error obtained from the sensors, extended kalman filter is utilized to fuse the sensors and by fusing the Kinect and inertial sensor, the errors obtained in head tracking can be minimized.

Kalman filters minimizes the statistical noise or sensors noise by utilizing a number of measured values observed over a period of time. The Kalman filter was developed by Rudolf E. Kalman.

Kalman filter is enormously used in a lot of applications in navigation, tracking, and control of autonomous vehicles. The main application of kalman filters is time series analysis and signal processing. In the field of robotics, the kalman filter is mostly used for trajectory estimation.

Basic operation of Kalman filter consists of two steps. The first step is prediction step, where the filter estimates the values of current state variables with their possible uncertainties. As the next measurement is observed, these estimated values are updated by weighted averages. The algorithm of Kalman filter is highly recursive, so it is easy to implement the kalaman filter in real time applications. It utilizes the current input measurements and last measured states and its uncertainty matrix.

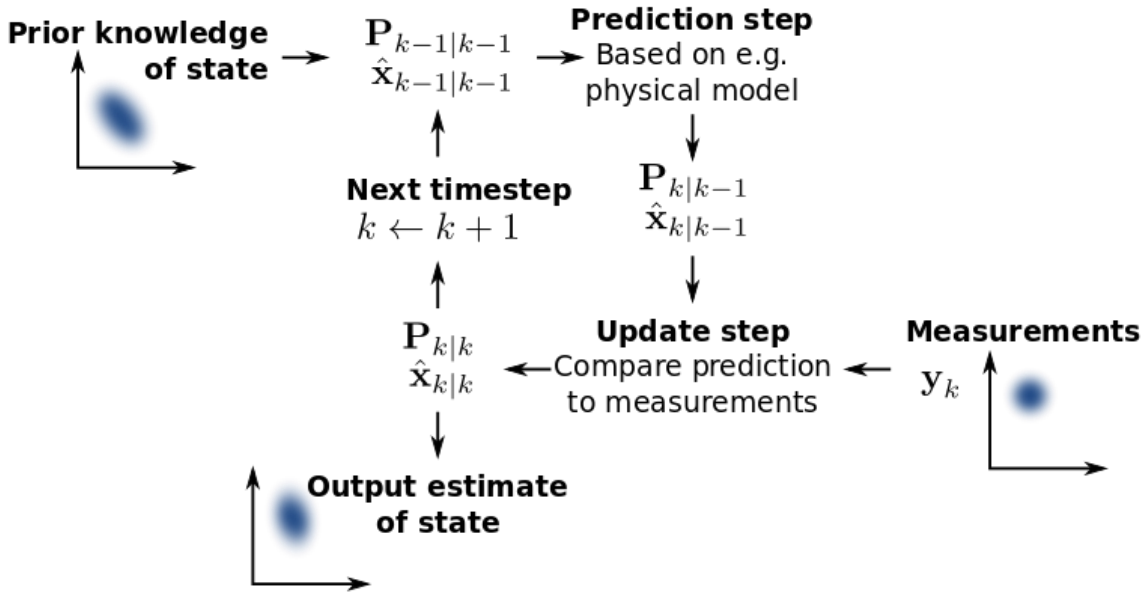


Figure 4.1. The flow of the Kalman Filter [24]

4.2. Discrete time Kalman Filter

The Kalman filter [25] was developed by **Rudolf** (**Rudi**) **Emil Kálmán** in 1960 to solve linear predictive problem. A brief derivation of kalman filter and extended Kalman filter is given below.

suppose the state $y \in \Re^n$ of a linear difference equation

$$y_k = By_{k-1} + C\mathbf{u}_k + \mathbf{w}_{k-1} \qquad \textbf{(4.1)}$$

process noise **w** is derived from *N*(0,**R**), as covariance matrix **R**.

with a measurement vector $\mathbf{z} \in \Re^m$

$$\mathbf{z}_k = \mathbf{H}y_k + \mathbf{v}_k \qquad \textbf{(4.2)}$$

Measurement noise vector **v** is drawn from *N*(0,S), with covariance matrix **S**.

B, R are *n*x*n*. **C** is *n*x*l*. **S** is *m*x*m*. **H** is *m*x*n*.

$y^{\wedge}_k \in \Re^n$ is the estimated state at time-step *k*.

$\hat{y}_k^- \in \Re^n$ after prediction, before observation.

Errors:

$$\mathbf{e}_k^- = y_k - \hat{y}_k^- \qquad \mathbf{e}_k = y_k - \hat{y}_k \qquad \textbf{(4.3)}$$

Error covariance matrices:

$$Q_k^- = E[\mathbf{e}_k^- \mathbf{e}_k^{-T}] \qquad Q_k = E[\mathbf{e}_k \mathbf{e}_k^T] \qquad \textbf{(4.5)}$$

Kalman Filter's task is to update $\hat{y}_k \quad Q_k$

Update expected value of **y**

$$\hat{y}_k^- = B\hat{\mathbf{x}}_{k-1} + C\mathbf{u}_k \qquad \textbf{(4.6)}$$

Update error covariance matrix **Q**

$$Q_k^- = BQ_{k-1}B^T + R \qquad \textbf{(4.7)}$$

Previous statements were simplified versions of the same idea:

$$\hat{y}(t_3^-) = \hat{y}(t_2) + u[t_3 - t_2] \qquad \textbf{(4.8)}$$

$$\sigma^2(t_3^-) = \sigma^2(t_2) + \sigma_\varepsilon^2[t_3 - t_2] \qquad \textbf{(4.9)}$$

Update expected value

$$\hat{y}_k = \hat{y}_k^- + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}\hat{y}_k^-) \qquad \textbf{(4.10)}$$

innovation is $\mathbf{z}_k - \mathbf{H}\hat{y}_k^-$ **(4.11)**

Update error covariance matrix

$$Q_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H})Q_k^- \qquad \textbf{(4.12)}$$

Compare with previous form

$$y^{\wedge}(t_3) = y^{\wedge}(t_3^-) + K(t_3)(z_3 - y^{\wedge}(t_3^-)) \quad \textbf{(4.13)}$$

$$\sigma^2(t_3) = (1 - K(t_3))\,\sigma^2(t_3^-) \quad \textbf{(4.14)}$$

The optimal Kalman gain $\mathbf{K}_k$ is

$$\mathbf{K}_k = Q_k^- \mathbf{H}^T (\mathbf{H} Q_k^- \mathbf{H}^T + S)^{-1} \quad \textbf{(4.15)}$$

$$= \frac{Q_k^- \mathbf{H}^T}{\mathbf{H} Q_k^- \mathbf{H}^T + S} \quad \textbf{(4.16)}$$

Compare with previous form

$$K(t_3) = \frac{\sigma^2(t_3^-)}{\sigma^2(t_3^-) + \sigma_3^2} \quad \textbf{(4.17)}$$

As most of the real engineering processes are **nonlinear in nature**. Some can be approximated by linear systems but some cannot.

This was recognized early in the history of Kalman filters and led to the development of the "**extended Kalman filter**" which is simply an extension of linear Kalman filter theory to nonlinear systems.

A Nonlinear system is simply a process that can be described by the following two equations:

$$y_k = f(y_{k-1}, \mathbf{u}_k) + \mathbf{w}_{k-1} \quad \textbf{(4.18)}$$

$$\mathbf{z}_k = h(y_k) + \mathbf{v}_k \quad \textbf{(4.19)}$$

process noise **w** is drawn from *N*(0,**Q**), with covariance matrix **R**.

measurement noise **v** is drawn from *N*(0,**R**), with covariance matrix **S**.

For a scalar function *y*=*f*(*y*),

$$\Delta L = f'(y)\Delta y \quad \textbf{(4.20)}$$

For a vector function **l**=*f*(**x**),

$$\Delta l = \mathbf{J}\,\Delta y = \begin{bmatrix} \Delta l_1 \\ \vdots \\ \Delta l_n \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial y_1}(y) & \cdots & \frac{\partial f_1}{\partial y_n}(y) \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial y_1}(y) & \cdots & \frac{\partial f_n}{\partial x_n}(y) \end{bmatrix} \cdot \begin{bmatrix} \Delta y_1 \\ \vdots \\ \Delta y_n \end{bmatrix} \qquad \textbf{(4.21)}$$

Let **B** be the Jacobian of *f* with respect to **x**.

$$B_{ij} = \frac{\partial f_i}{\partial y_j}(y_{k-1}, \mathbf{u}_k) \qquad \textbf{(4.22)}$$

Let **H** be the Jacobian of *h* with respect to **y**.

$$\mathbf{H}_{ij} = \frac{\partial h_i}{\partial y_j}(y_k) \qquad \textbf{(4.23)}$$

Then the Kalman Filter equations are almost the same as before!

Predictor step:

$$\hat{y}_k^- = f(\hat{y}_{k-1}, \mathbf{u}_k) \qquad \textbf{(4.24)}$$

$$Q_k^- = BQ_{k-1}B^T + R \qquad \textbf{(4.25)}$$

Kalman gain:

$$\mathbf{K}_k = Q_k^- \mathbf{H}^T (\mathbf{H} Q_k^- \mathbf{H}^T + S)^{-1} \qquad \textbf{(4.26)}$$

Corrector step

$$\hat{y}_k = \hat{y}_k^- + \mathbf{K}_k(\mathbf{z}_k - h(\hat{y}_k^-)) \qquad \textbf{(4.27)}$$

$$Q_k = (\mathbf{I} - \mathbf{K}_k\mathbf{H})\,Q_k^- \qquad \textbf{(4.28)}$$

4.3. Head Related Transfer Function

Human body contains two ears but our listening system has ability to listen in all three dimensions. Whenever a listener listens any sound around the location of sound can be estimated, this is because the brain and inner ear work together to estimate the position of sound source. This ability makes human to track anything with the sound makes any

listener to localize sound source even without light. Humans percept the sound source by the difference of the monaural sound cues received from both ears these monaural cues arrive human ear at different intensity and different time with these intensity difference and arrival time difference human auditory system human brain estimate the location of the sound source. The ability of human brain to estimate the sound source location with respect to ear location can be defined by an impulse response that is called Head Related Impulse Response (HRIR). While Head Related Transfer Function (HRTF) is the Fourier Transform of HRIR.

HRTF of both ears tells us the filtering of the source of sound (t) before it is estimated at left ear as $yL(t)$ and at right ear as $yR(t)$.In linear system analysis the Transfer Function of any system is described as the ratio of output and input functions. The Transfer Function (f) of a system at frequency f is given as

$$(f) = Output(f) \,/\, Input(f) \quad \textbf{(4.29)}$$

HRTF can be derived by different methods such as localization of sound in Virtual Auditory space [26], HRTF phase synthesis [27], HRTF magnitude synthesis [28]. In common HRTF does not depend on the distance but they do depend on direction For this reason most of the HRTFs that are reported in literature reported that source of sound is

1 m away from the listener let YL(θ,ϕ,ω) and YR(θ,ϕ,ω)be the Fourier transforms of the signals received at the entrance of the listener's left and right blocked ear canals, respectively, and the sound source is placed one meter away from the listener and the sound source has azimuth θ and elevation ϕ. Consider YC(θ,ϕ,ω) be the Fourier transform of the signal received. Hence the HRTF of left and right ear can be expressed as:

$$HL(\theta,\phi,\omega) = YL(\theta,\phi,\omega)/\ YC(\theta,\phi,\omega) \quad \textbf{(4.30)}$$

$$HR(\theta,\phi,\omega) = YR(\theta,\phi,\omega)/\ YC(\theta,\phi,\omega) \quad \textbf{(4.31)}$$

Some of the ways through which the HRTF can be derived are

1. Localization of sound in Virtual Auditory space [26]
2. HRTF Phase synthesis [27]
3. HRTF Magnitude synthesis [28]

4.4. Localization of Sound in Virtual Auditory Space

A normal assumption is made while calculating the HRTF is that the sound listened by the listener in virtual auditory space is similar to voice in free space.

Normally, the sound listened under the headphones seems to be originate inside the head. In virtual auditory space for realistic feelings, headphones must be able to “externalize” the sound produced within the virtual auditory space. Using HRTF the sound can be positioned in in special space as described below.

Now , suppose that $x_1(t)$ is an electrical signal driving a loudspeaker, and $y_1(t)$ is the signal received by the microphone placed inside the listeners eardrum. Now, let us suppose that $x_2(t)$ is the signal that is driving the headphones and $y_2(t)$ is the response of microphone to the signal. The aim is to achieve $x_2(t)$ such that $y_2(t) = y_1(t)$ is obtained. So, we get following equations after applying the laplace transforms to these signals.

$Y_1 = X_1$LFM, and [26] **(4.32)**

$Y_2 = X_2$HM, [26] **(4.33)**

Where L represents the transfer function of loudspeaker in free field. F represents the HRTF while H is transfer function between headphone and eardrum.So, by taking $Y_1 = Y_2$ we get the value of X_2 as

$X_2 = X_1$LF/H. [26] **(4.34)**

And by the observations, the required transfer function is given as

T= LF/H [26] **(4.35)**

4.5. HRTF Phase Synthesis

This method is pretty less reliable method for estimating the values of the head related transfer functions in the low part of the frequency band, while in the upper band frequencies are effected by the pinnae features. Earlier studies and most of the research published shows that the HRTF phase response is mostly linear.

A scaling factor is a function of the anthropometric features. For example, a training set of N subjects would consider each HRTF phase and describe a single ITD scaling factor as the average delay of the group. This computed scaling factor can estimate the time delay as function of the direction and elevation for any given individual. Changing the delay time to phase response for the left and the right ears is conventional.

The HRTF phase can be expressed by the interaural time difference scaling factor. This is in turn is quantified by the anthropometric data of a given individual taken as the source of reference. For a generic case take β as a sparse vector.

$$\beta = [\beta_1, \beta_2, \ldots, \beta_N]^T \quad \text{[27]} \quad \textbf{(4.36)}$$

for a non-negative shrinking parameter λ:

$$\beta = \underset{\beta}{\operatorname{argmin}} \left(\sum_{a=1}^{A} \left(y_a - \sum_{n=1}^{N} \beta_n X_n^2 \right) + \lambda \sum_{n=1}^{N} \beta_n \right) \quad \text{[27]} \quad \textbf{(4.37)}$$

The interaural time difference scaling factor H can be expressed as

$$H' = \sum_{n=1}^{N} \beta_n H_n. \quad \text{[27]} \quad \textbf{(4.38)}$$

4.6. HRTF Magnitude Synthesis

We solve the above minimization problem using Least Absolute Shrinkage and Selection Operator (LASSO). We assume that the HRTFs are represented by the same relation as the anthropometric features. Therefore, once we learn the sparse vector β from the anthropometric features, we directly apply it to the HRTF tensor data and the subject's HRTF values H' given by:

$$H'_{d,k} = \sum_{n=1}^{N} \beta_n H_{n,d,k} \quad \textbf{(4.39)}$$

4.7. Summary

The chapter summarizes as follows:

- Simultaneous localization and mapping filters like kalman filter and particle filter can be used to fuse the measured values of sensors used for head tracking.
- Different methods for calculating the Head related transfer function have been described. The standard CIPIC database has been utilized in the thesis.
- Sensor fusion using extended kalman filter can be possible. Depth and inertial sensor fusion using extended kalman filter has never been reported in literature so far.

5. EXPERIMENTATION

This chapter explains the experimental setup and data acquiring procedure in detail. The detail about the sensors used also presented in the chapter.

5.1. Experimental Setup

Experiments are conducted at PRISM LAB in College of Electrical and Mechanical Engineering NUST, Islamabad, Pakistan. Basically a depth sensor (Kinect) and an inertial measurement unit were used for data acquisition. Both sensors were used for head tracking purposes. A Kinect was placed at a one-meter distance from the test case while the inertial sensor was mounter on the headphones as shown in fig 5.2.

Figure 5.1. The Experimental Setup

Figure 5.2. The Experimental Setup and the test case

A test case wearing head phones and sitting at exact one-meter distance from the Kinect is shown in the fig. 5.2. Test of fifty test cases were performed at 5 different positions as normal, right, left, up, and down head poses.

Table 5-1 Head Tracking data obtained from a test case

	imu gyro			Kinect		
	roll	pitch	yaw	roll	pitch	Yaw
Left	-6.68	-4.02	37.16	-10	0	40
Right	-0.32	-4.3	-7.45	10	5	30
Up	-5.7	34.22	-5.09	5	35	5
Down	-1.89	-21.25	-3.28	10	-25	-5
normal	-0.85	0.57	0.57	5	5	5

Table 5.1 is the data obtained from the head poses of only one test case. A total of fifty test are conducted and similar fifty tables are made.

5.2. Inertial Measurement Unit

Fig 5.3 shows MPU 6050, which is an inertial measurement unit which is employed in our thesis for head tracking purpose. MPU 6050 is a InvenSense's 6-axis MEMS based motion sensors with a 3-axis gyroscope and a 3-axis accelerometer. The sensor is power efficient, low cost and has high accuracy for gyro and accelerometer values with a high refresh rate in a very compact size. It is widely used in electronic devices like robots and gadgets where it is needed to find out the position of the device. The sensor can easily be integrated with Arduino, MPU6050 Arduino Module and later on it can be used to communicate with MATLAB to find out the exact position of head which will be kept in continuous tracking state. This chip has 16-bit serial communication which increases the rate of data transfer from the MPU 6050 module to the Arduino. It transfers the data of all 6 sensors combined which increases the rate of data extraction from sensors to apply it in real time head tracking.

Figure 5.3. MPU 6050 [29]

Table 5-2 Data Obtained from Inertial Sensor

1. ACC:-5.33;-4.04;0.00#GYR:-0.85;0.57;0.55#FIL:-5.31;-3.91;0.55
2. ACC:-5.47;-4.20;0.00#GYR:-0.84;0.56;0.53#FIL:-5.31;-3.93;0.53
3. ACC:-5.80;-3.98;0.00#GYR:-0.83;0.56;0.55#FIL:-5.32;-3.93;0.55
4. ACC:-4.75;-3.78;0.00#GYR:-0.77;0.54;0.54#FIL:-5.24;-3.94;0.54
5. ACC:-4.78;-3.76;0.00#GYR:-0.79;0.48;0.47#FIL:-5.24;-3.99;0.47
6. ACC:-4.95;-4.52;0.00#GYR:-0.81;0.44;0.47#FIL:-5.25;-4.04;0.47
7. ACC:-5.57;-4.53;0.00#GYR:-0.84;0.48;0.56#FIL:-5.29;-4.03;0.56
8. ACC:-5.07;-3.60;0.00#GYR:-0.86;0.51;0.59#FIL:-5.30;-3.98;0.59
9. ACC:-5.86;-3.89;0.00#GYR:-0.86;0.54;0.57#FIL:-5.32;-3.95;0.57
10. ACC:-5.10;-3.33;0.00#GYR:-0.85;0.57;0.55#FIL:-5.30;-3.90;0.55
11. ACC:-5.17;-3.88;0.00#GYR:-0.85;0.55;0.53#FIL:-5.30;-3.91;0.53
12. ACC:-5.44;-3.80;0.00#GYR:-0.86;0.54;0.56#FIL:-5.32;-3.92;0.56
13. ACC:-5.61;-4.03;0.00#GYR:-0.85;0.53;0.58#FIL:-5.32;-3.94;0.58
14. ACC:-4.90;-3.97;0.00#GYR:-0.86;0.54;0.57#FIL:-5.30;-3.92;0.57
15. ACC:-5.34;-3.47;0.00#GYR:-0.87;0.55;0.55#FIL:-5.32;-3.90;0.55
16. ACC:-5.26;-3.95;0.00#GYR:-0.90;0.55;0.57#FIL:-5.34;-3.90;0.57
ACC:-5.55;-4.29;0.00#GYR:-0.91;0.54;0.60#FIL:-5.36;-3.93;0.60
ACC:-5.25;-3.57;0.00#GYR:-0.91;0.55;0.61#FIL:-5.36;-3.90;0.61

Table 5-2 shows the continuous data being obtained from the inertial sensor. As it is already explained that the inertial sensor consists of an accelerometer and a gyroscope. So, the roll, pitch and yaw values are obtained using gyroscope and accelerometer and by fusing both accelerometer and gyroscope using complementary filter values are also taken form roll, pitch and yaw.

5.3. Microsoft Kinect

Figure 5.4. Microsoft Kinect V2 [30]

Kinect a motion sensing input device made by Microsoft for Xbox gaming consoles, and windows PCs. It enables users to interact and control with devices through gesture and speech. Kinect is widely used in interactive games and windows applications.

Inside the covering, a kinect V2 contains:

- A depth sensor, which is used for 3D visualization of the objects. The resolution of depth sensor is 512*424, and it operates at frequency of 30 Hz. Its operational range is from 0.5 meter to 4.5 meter.
- A camera that is used for capturing color images. It operates at frequency of 30 Hz in normal condition, and in case of low light it operates ar 15 Hz. Its resolution is 1080p.
- An Infrared Camera, it allows the sensor to see in dark. In this new addition of the Kinect, the IR sensor and color camera, both can be used at a same time. Its operating frequency is 30 Hz and it operates at a resolution of 512 x 424.
- A multi-array microphone, that is used for capturing the sound, recording the sound, can also be used for sound localization.

By addition of Kinect for Windows support package of MATLAB®, Kinect can easily be integrated with MATLAB®.

Figure 5.5: Different head poses of the test cases

Head poses of test cases are shown in the fig 6.5 below. A test case is asked to sit in front of Kinect as shown in fig 6.2 and to wear the head phones on which an inertial sensor is mounted. After that the test case is asked to move his head in 5 different position one by one as explained earlier. Then the roll, pitch and yaw data obtained from both sensors is recorded.

5.4. Summary

The chapter can be summarized as follows

- The roll, pitch and yaw values are taken at 5 different head poses .i.e., front face, right, left, up and down using inertial sensor and the Microsoft Kinect ®
- Accelerometer, gyroscope and their fusion is used from head pose measurement in inertial sensors.
- Kinect has limitations in field of view, while there are drift problems in Inertial measurement unit.

6. RESULTS AND DISCUSSIONS

This chapter explains the results obtained from the complete research performed in the thesis. Moreover, the results from the different combinations of measurement and control inputs of the kalman filter are also explained in the chapter. Finally, the comparison of all possible combinations is given. As already explained that the tests were conducted on 50 test cases. Due to large quantity of data and results, result graphs of only 3 test cases are shown here. In the last part of the chapter, 3D sound results obtained from the CIPIC database also presented.

6.1. Test Case 1

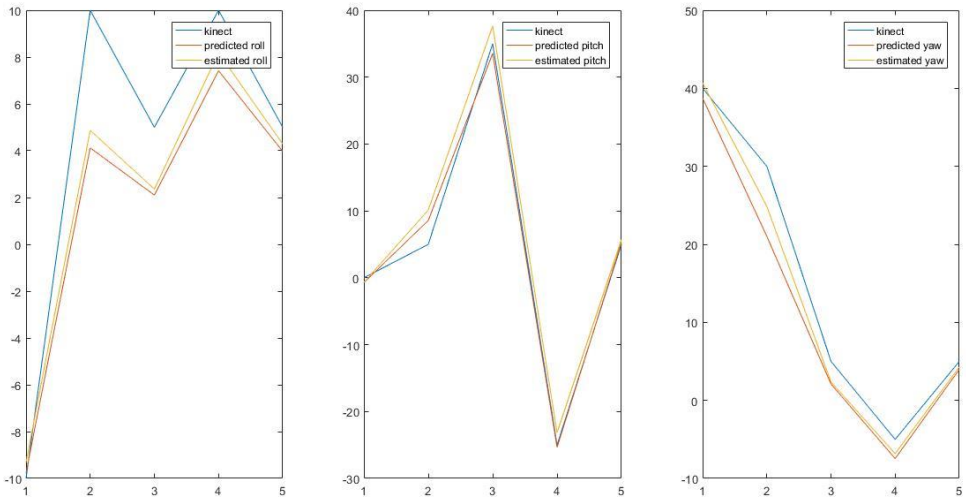


Figure 6.1. Taking Kinect as control and Gyroscope as the measurement value

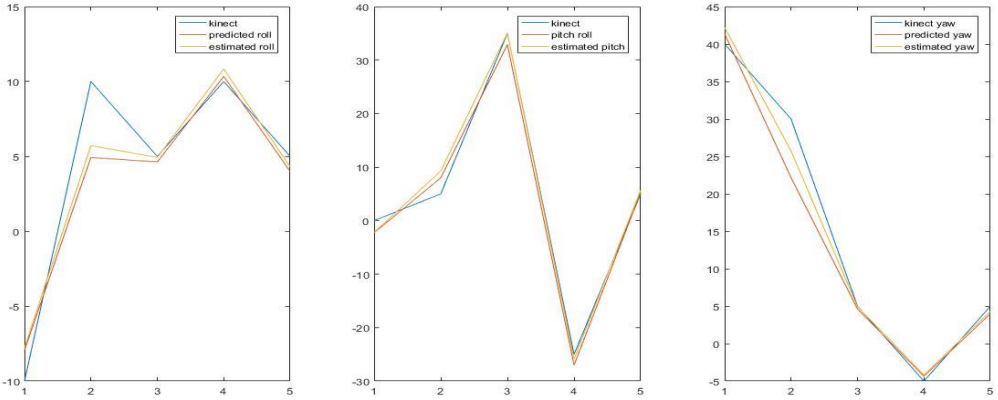


Figure 6.2. Taking Kinect as Control and Fusion of Gyro and Accelerometer as Measurement

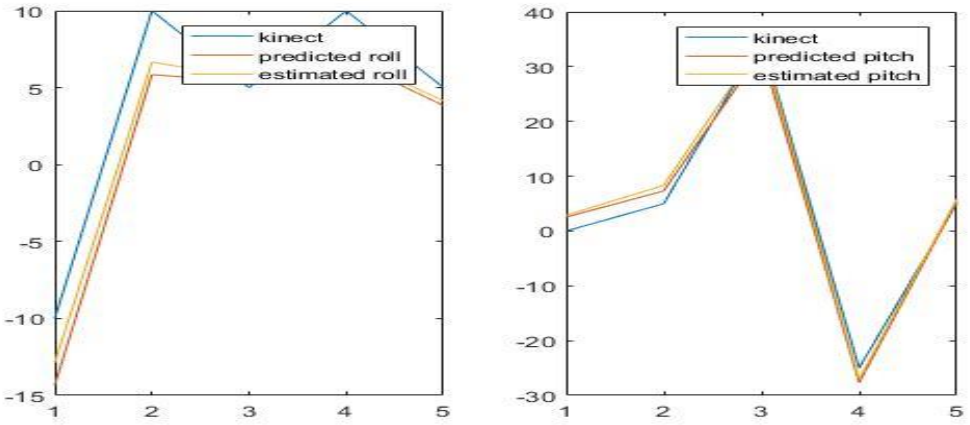


Figure 6.3. Taking Kinect as Control and Accelerometer as Measurement

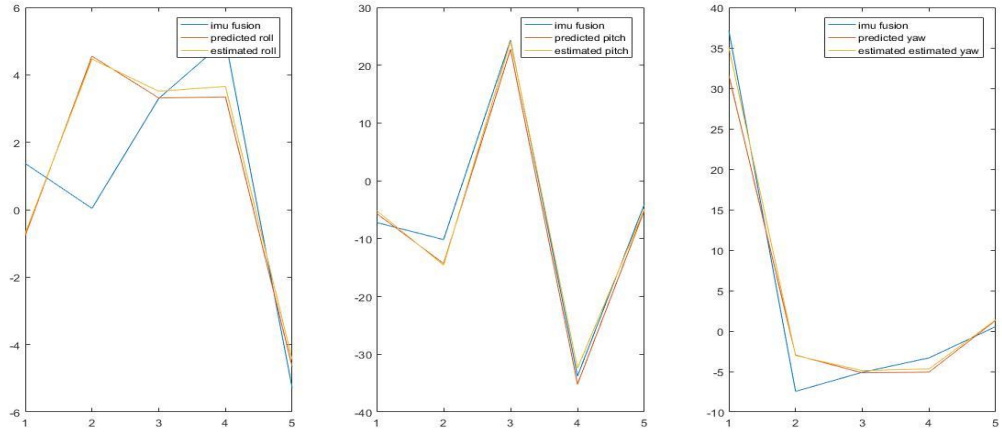


Figure 6.4. Taking IMU fusion as control and Kinect as measurement

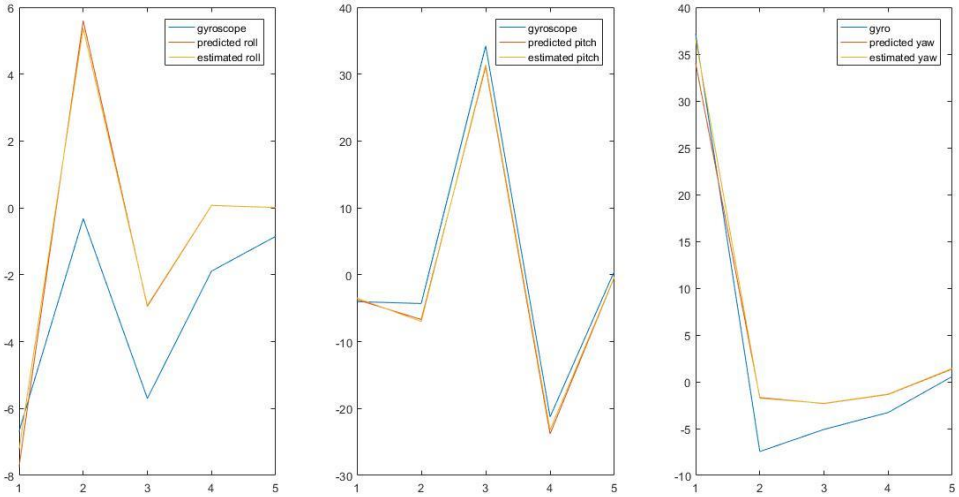


Figure 6.5. Taking gyroscope as control and Kinect as Measurement

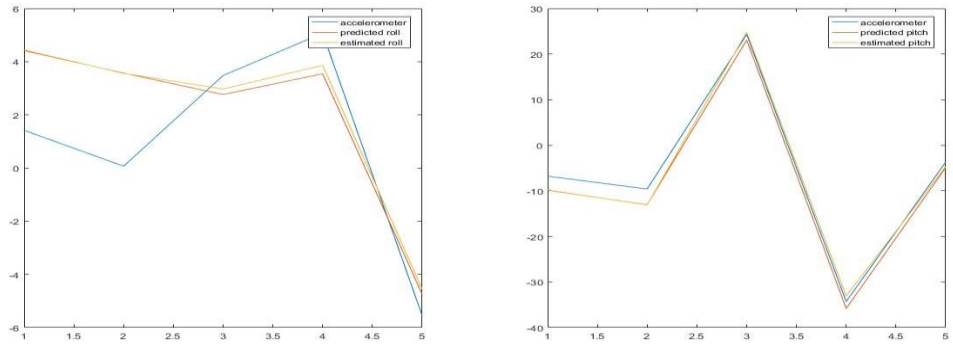


Figure 6.6. Taking accelerometer as control and Kinect as measurement

6.2. Test Case 2

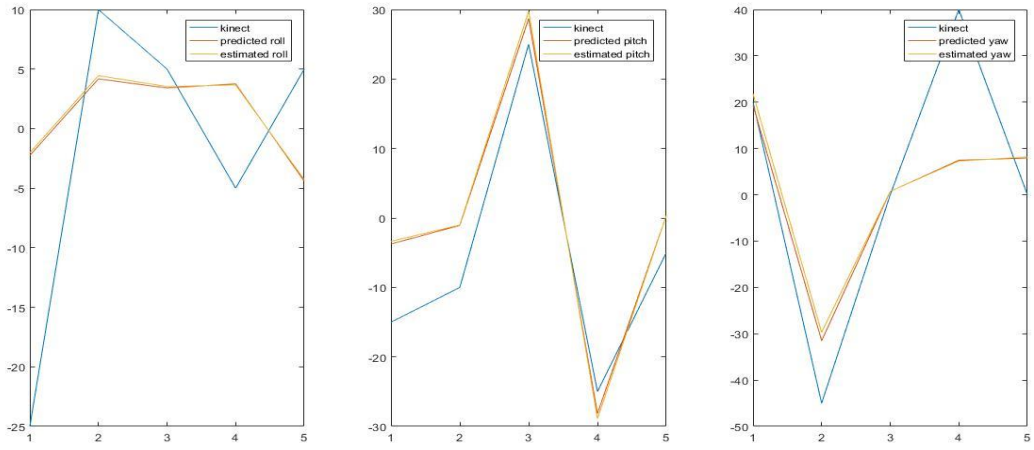


Figure 6.7. Taking Kinect as control and gyroscope as measurement

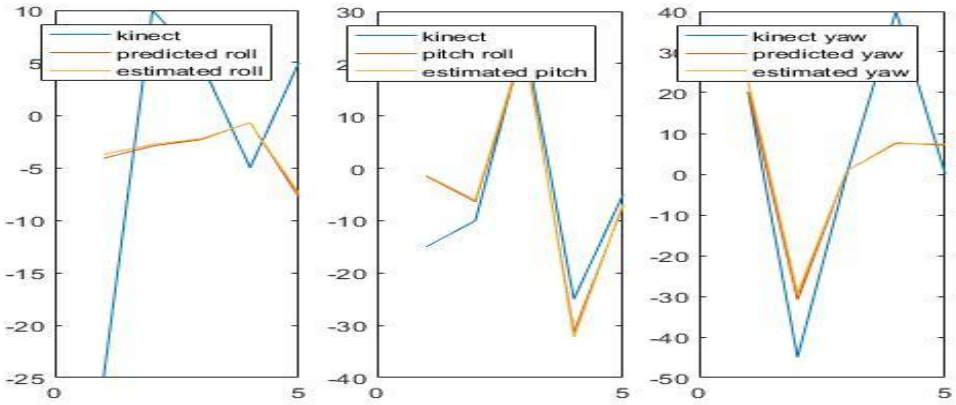


Figure 6.8. Taking Kinect as control and IMU fusion as measurement

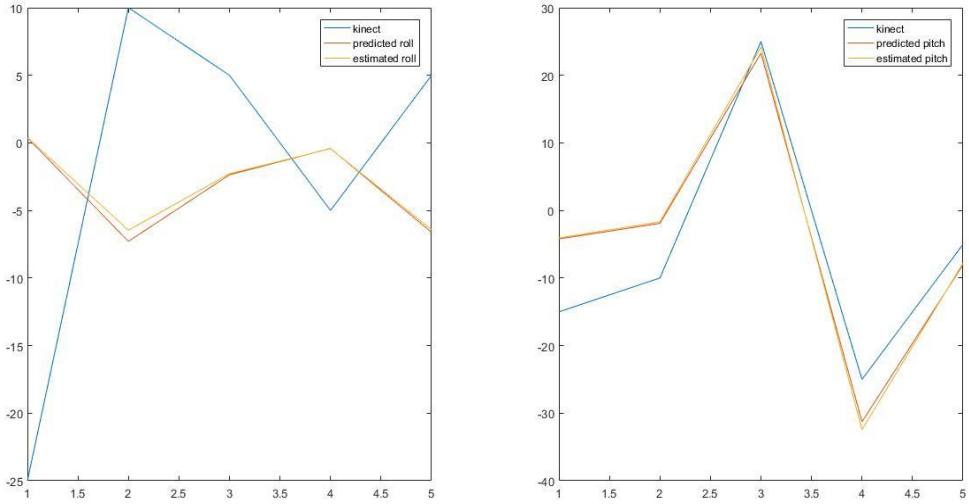


Figure 6.9. Taking Kinect as control and accelerometer as measurement

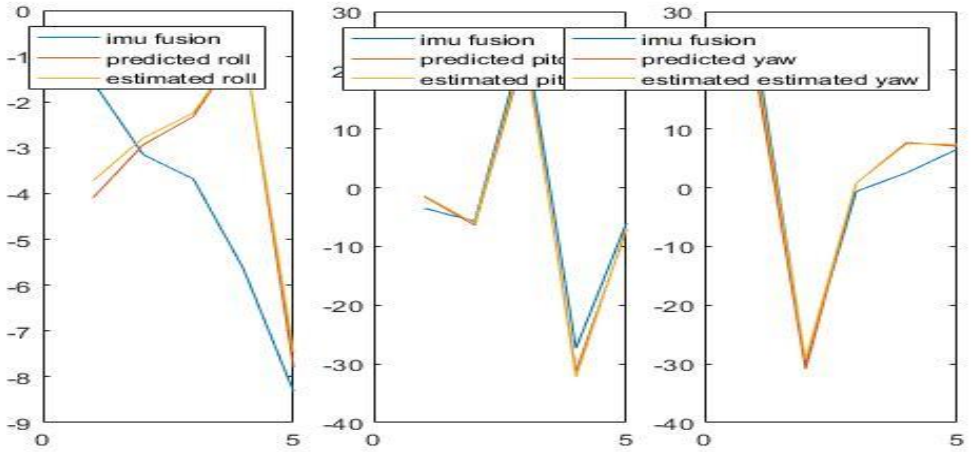


Figure 6.10. IMU Fusion as Control and Kinect as measurement

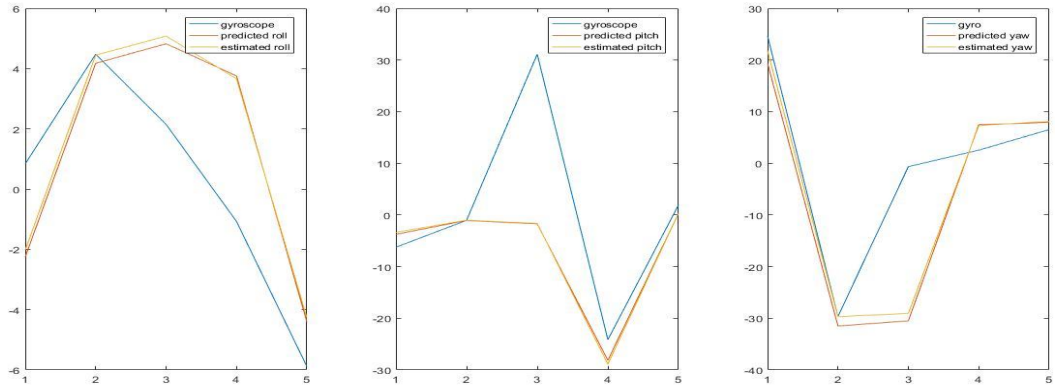


Figure 6.11. Gyroscope as control and Kinect as measurement

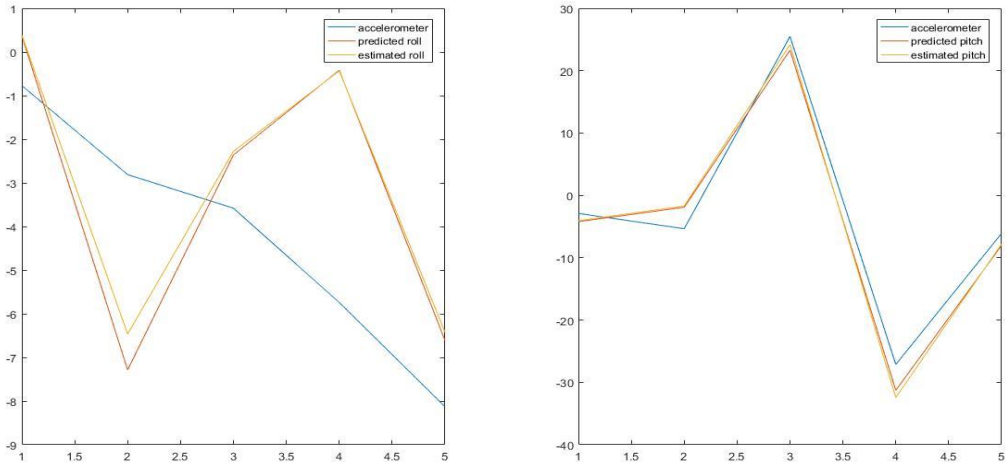


Figure 6.12. Taking accelerometer as control and Kinect as measurement

6.3. Test Case 3

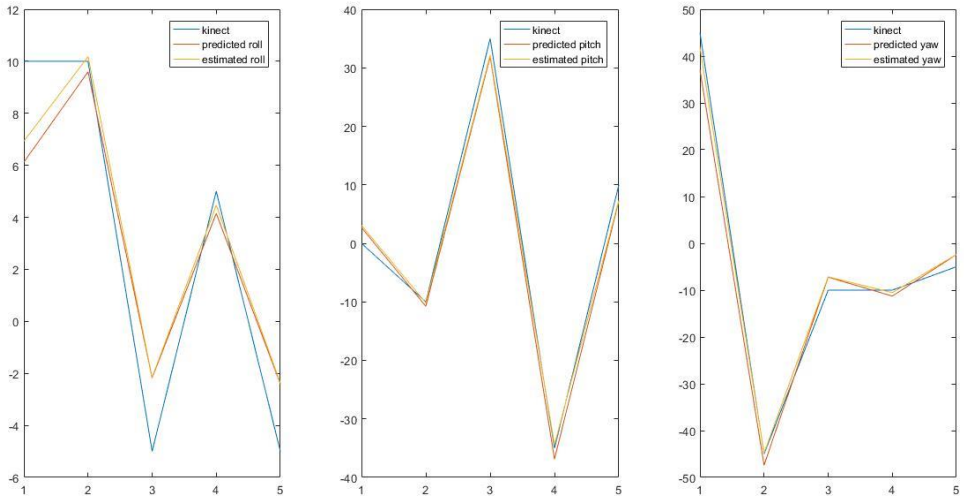


Figure 6.13. Kinect as control and gyroscope as measurement

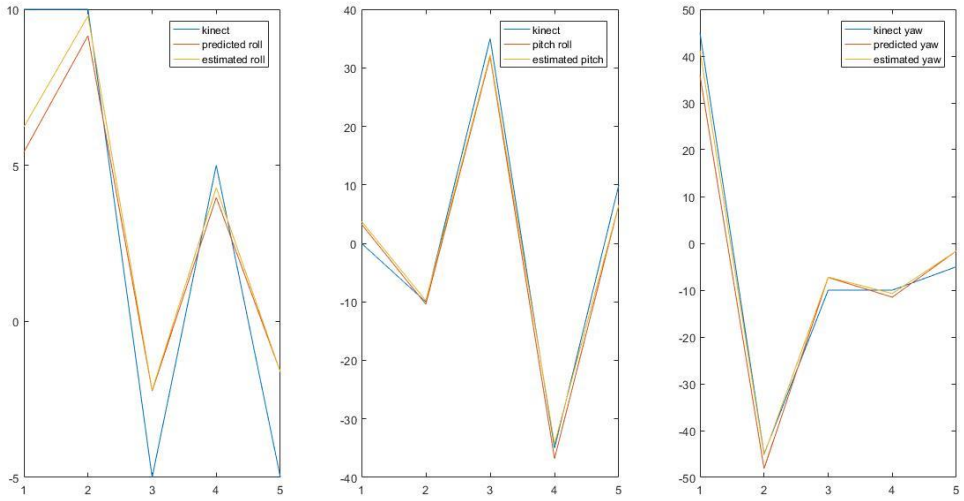


Figure 6.14. Kinect as control and IMU fusion as measurement

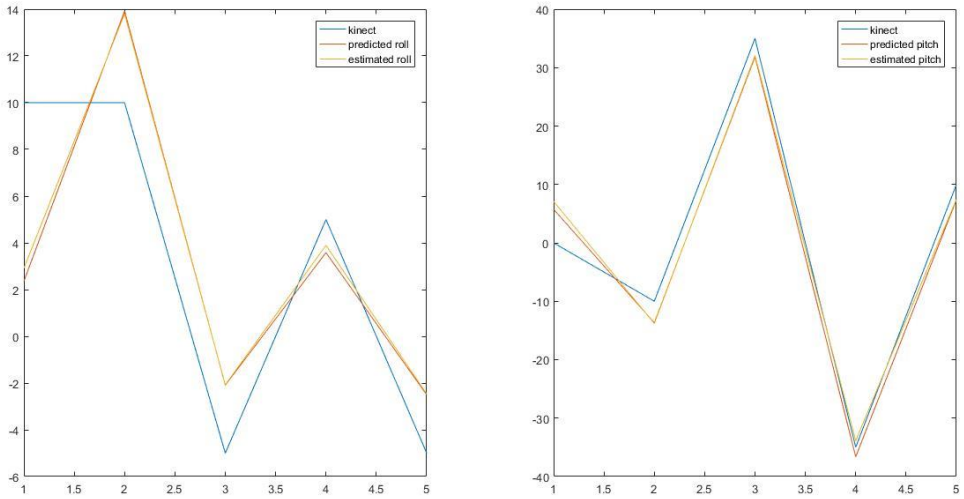


Figure 6.15. Kinect as control and accelerometer as measurement

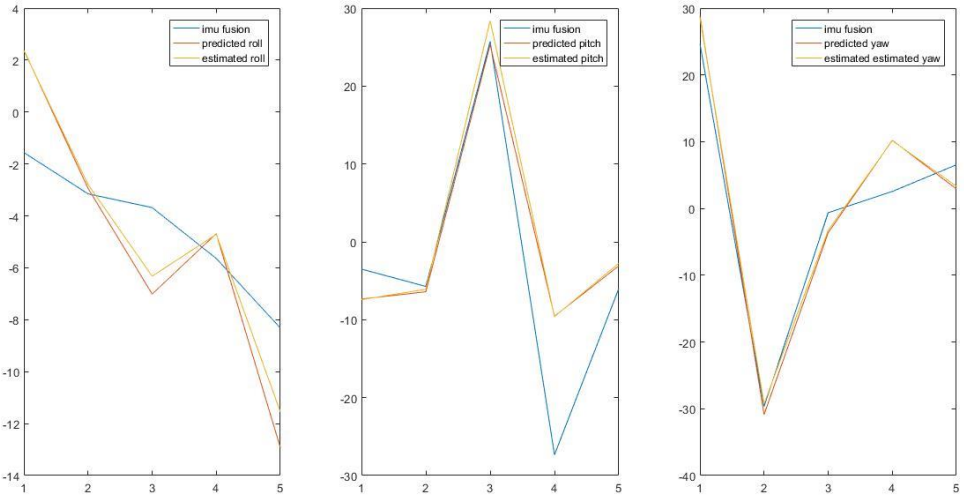


Figure 6.16. IMU fusion as control and Kinect as measurement

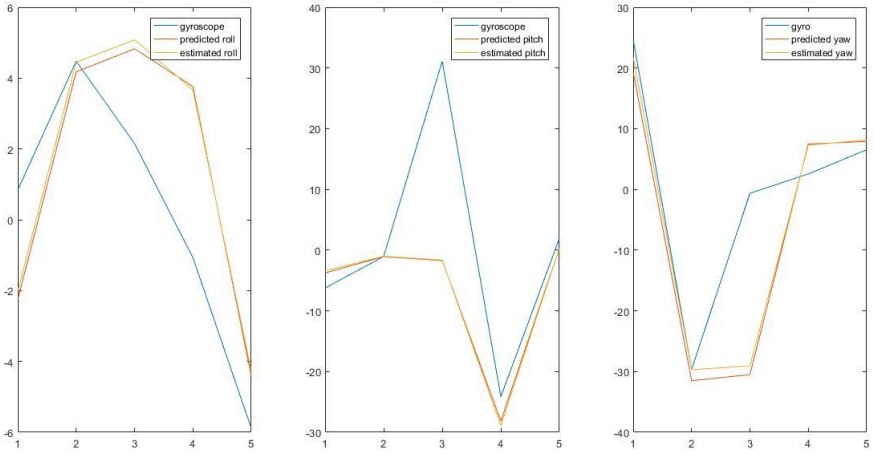


Figure 6.17. Gyroscope as control and Kinect as measurement

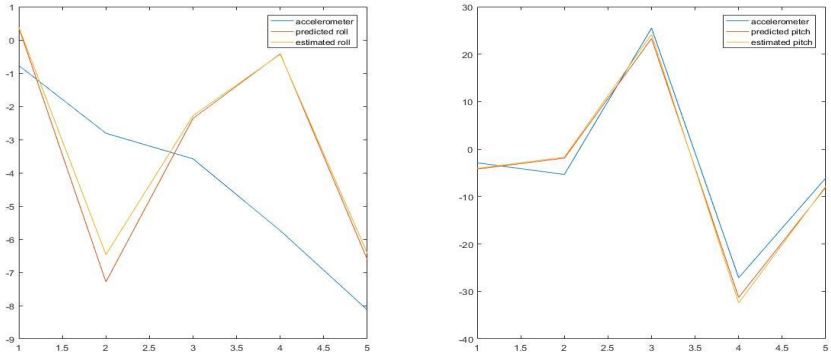
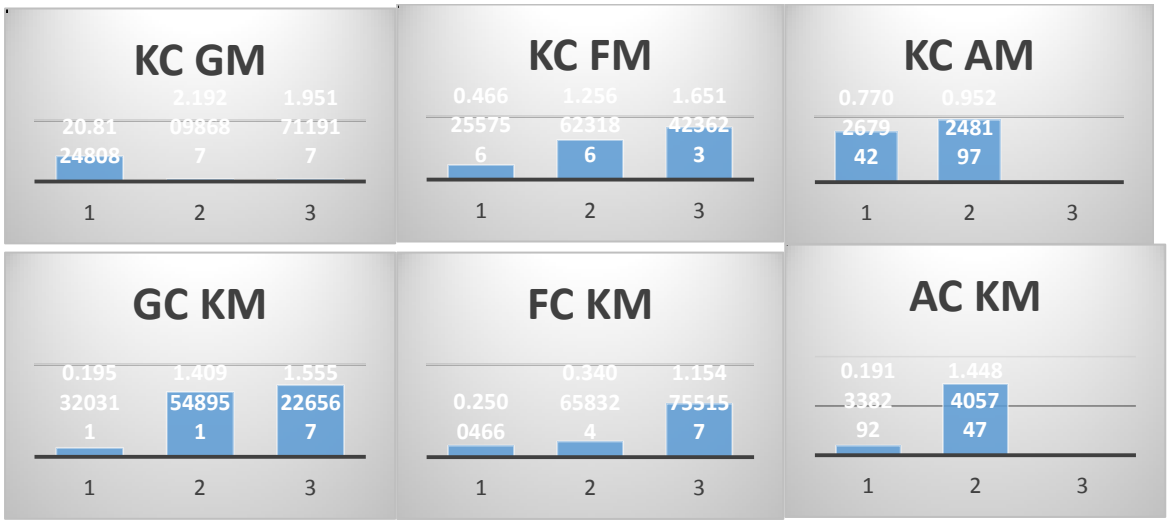
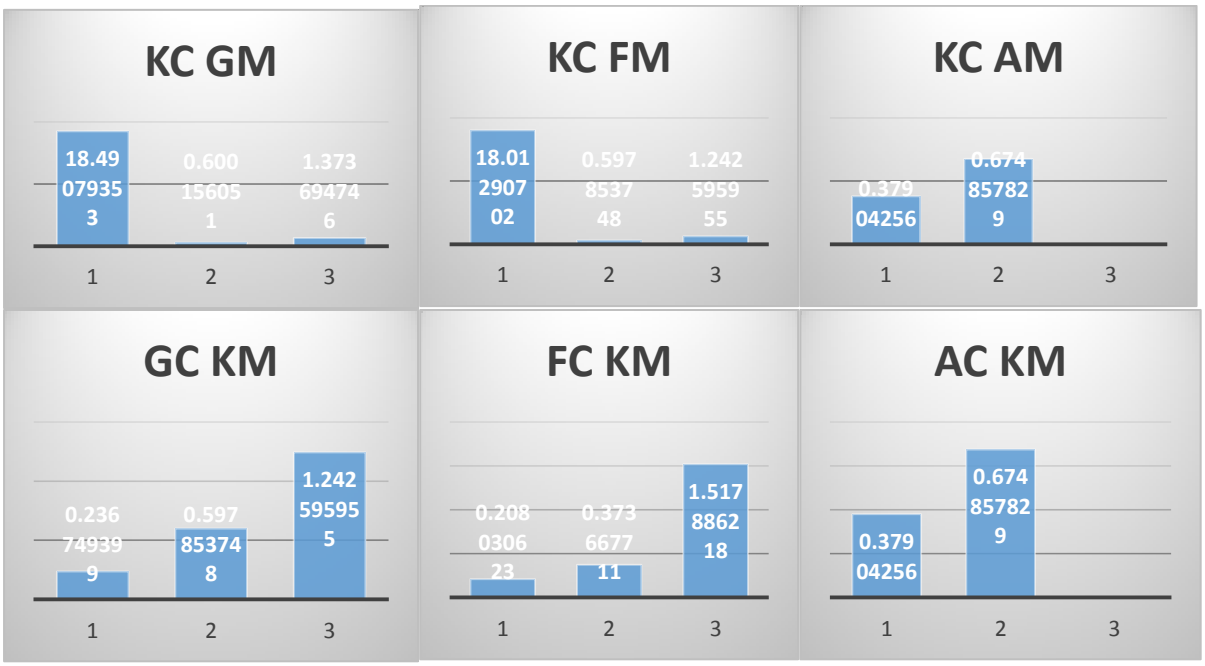


Figure 6.18. Accelerometer as control and Kinect as measurement

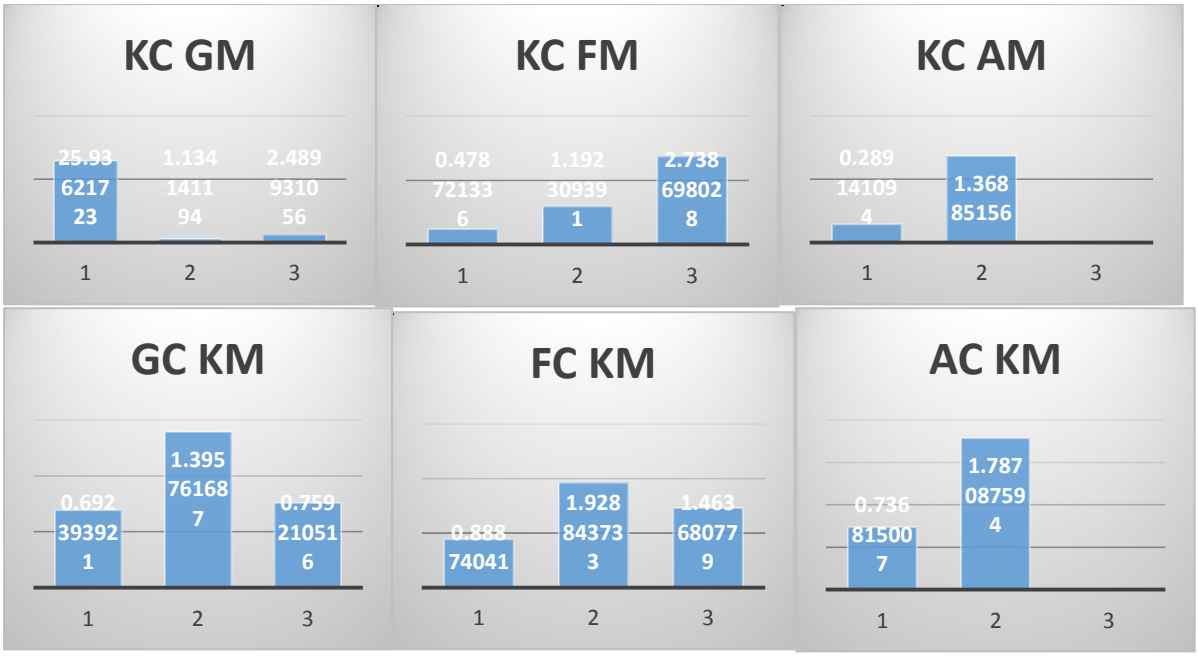
6.4. RMSE for Test Case 1



6.5. RMSE Test Case 2



6.6. RMSE Test Case 3



6.7 Discussion

This section of the chapter briefly discusses the results obtained from the experiments conducted in the thesis. As, it is well explained earlier that the experiments were conducted on 48 test cases. As, due to some errors and mistakes in data acquisition, twenty test cases were chosen after screening process for further implementation of the technique. The results are further. There are six possible combinations of extended Kalman Filter in the proposed technique. First part of the chapter is dedicated to the results obtained from the extended kalman filter in five different face poses. In the graphs, the X-axis is the position number of the face pose, while the y axis describes the angle of the head in degrees.

There are six different graphs according to different combinations of EKF. So, fig 7.1, 7.7 and 7.13 describes the yaw, pitch and roll obtained by taking the Kinect as a control input while taking gyroscope as the measurement input. Each figure from fig. 7.1, 7.7 and 7.13 further divided into three figures, that consists of roll figure, pitch figure and yaw from left to right. Each graph consists of three plots, in which the blue line shows the control input while the red coloured plots show the estimated values and the yellow coloured graph lines show the predicted values. So, it can be seen in all figures in which the EKF is applied on roll values,

there exist a lot of difference between the control values and values obtained from EKF. This is due to the presence of drift in gyroscopes. The drift in gyroscopes is relatively in greater in roll values amount as compared to the pitch and yaw values. While in pitch and yaw values, these is a very small amount of drift in gyroscope values.

In Fig 7.2, 7.8 and 7.13 the graphs are by taking Kinect as the control input while taking fusion of accelerometer and gyroscope as the measurement input. Each figure is similarly consisting of further three figures as it is in described before. It can be seen these figures also that the plots for the roll values show relatively more difference in control values and the values from the EKF. As, it is explained early that the gyroscope shows relatively more drift in roll measurement.

Fig 7.3, 7.9 and 7.15 shows the plots obtained by taking Kinect as a control and accelerometer as the measurement value. These figures further divided into two more figures containing graphs for roll and pitch values not for the yaw values as it was described for the previous cases. This is due to the limitation of accelerometer in yaw measurement. In these graphs it can be seen that the difference in roll values is relatively much more as compared to pitch values. This error may be due to the jerk errors in the accelerometer.

Fig 7.4, 7.10, 7.16 are measured by taking Gyroscope as control input and Kinect as the measurement values. Fig 7.5, 7.11 and fig 7.17 are plotted by taking IMU fusion values as control input while the Kinect as measurement value. Fig 7.6, 7.12 and 7.18 are plotted by taking accelerometer as the control input while taking Kinect as the measurement values. In all of the cases it was observed that the difference between the EKF values and control vector values is maximum in roll values. It is due to the fact that during head movements the changes occur in roll values are relatively much more than the pitch and yaw values. Further, after calculating the EKF values of all the cases the root mean square values of all cases were observed. The RMSE value between the estimated and the control values were taken for all cases. In RMSE value each figure consists of six parts named as

KCGM= Kinect as control and gyroscope as measurement

KCFM= Kinect as Control and IMU fusion as measurement

KCAM=Kinect as control and Accelerometer as measurement

GCKM=Gyroscope as control and kinect as measurement

FCKM=IMU fusion as control and Kinect as measurement

ACKM= Accelerometer as control and Kinect as measurement

So, the RMSE for 20 selected cases was measured, and in the end the average RMSE for all cases was measured. The RMSE for all 20 cases is shown as below

6.8 RMSE of 20 Test Cases

KCGM				KCFM				KCAM			GCKM				FCKM				ACKM	
20.81248	2.192099	1.951712		0.466256	1.256623	1.651424		0.770268	0.952248		0.19532	1.409549	1.555227		0.250047	0.340658	1.154755		0.191338	1.448406
20.81248	0.600156	1.373695		18.01291	0.597854	1.242596		0.379043	0.674858		0.236749	0.597854	17.18244		0.208031	0.373668	1.517886		0.379043	0.674858
20.81248	1.134141	2.489931		0.478721	1.192309	2.738698		0.289141	1.368852		0.692394	1.395762	0.759211		0.88874	1.928844	1.463681		0.736815	1.787088
20.81248	2.404598	2.783348		0.874328	1.815592	1.760147		1.218381	1.248617		1.243375	1.670729	4.342875		0.900849	1.51996	3.512467		0.516265	1.255445
20.81248	0.786843	1.173589		0.555421	0.66634	0.926509		0.240397	1.116726		0.157076	1.934279	30.24327		0.899592	2.938441	2.839414		0.640749	1.595582
20.81248	1.057826	1.166929		0.208255	2.409249	0.555032		0.605439	0.879407		0.669928	2.434113	3.834799		1.921151	0.565646	2.900909		0.817471	2.505046
20.81248	0.588294	0.949386		0.15456	0.447458	0.30626		0.77247	0.693582		0.276681	1.558869	2.454102		0.661886	1.472746	1.759644		0.356484	1.73929
20.81248	24.64078	13.84548		0.508577	1.084029	0.509291		0.250082	0.497095		0.170736	2.64278	2.394053		10.76086	89.52491	8.068514		0.396009	1.269857
20.81248	2.512756	2.488002		0.432078	0.452461	1.594546		0.469872	1.33241		0.498959	2.312595	1.765149		6.75692	1.246027	2.085584		0.473656	1.173589
20.81248	1.434833	2.728718		0.65995	1.435362	2.579944		1.186295	1.365721		0.157544	0.916867	1.158606		0.098339	0.971067	1.011483		0.345988	1.135524
20.81248	3.803887	4.265595		1.085489	2.340045	3.900328		0.701289	1.249083		1.470804	2.718733	3.395666		3.123256	6.629472	3.489328		0.771101	1.501656
20.81248	0.493321	0.671929		0.184727	0.267764	0.736531		0.384486	0.654142		0.369645	1.839127	3.891545		0.35469	1.370938	1.847456		0.481009	1.895339
20.81248	1.941926	1.803752		0.317144	0.632657	0.59484		0.363898	0.629476		1.233953	1.58581	2.566815		14.05463	1.188291	7.421356		0.79975	1.482477
20.81248	1.735004	1.345282		0.130212	0.569885	0.953802		0.450645	1.093481		0.699299	2.268735	3.297463		0.50429	0.868617	1.87365		0.949606	2.335123
20.81248	1.735004	1.345282		0.130212	0.569885	0.953802		0.450645	1.093481		0.699299	2.268735	3.297463		0.50429	0.868617	1.87365		0.949606	2.335123
20.81248	0.594884	1.388467		0.140153	0.3529	1.040995		0.390749	1.414607		0.484351	1.290784	2.330052		0.355001	1.307662	2.239794		0.876627	1.20581
20.81248	2.18741	1.888676		0.520329	2.141566	1.901238		0.41215	2.270976		0.587157	1.053211	1.259056		0.625378	1.378482	1.388379		0.622581	1.158439
20.81248	1.497112	1.522592		0.249551	1.181902	1.069967		0.196506	1.445273		1.429617	7.226875	2.807145		0.246244	0.932865	1.108864		0.302695	1.4898
20.81248	0.503216	1.870171		0.408727	0.73986	1.207287		0.213253	0.39361		0.285438	1.853405	1.73953		1.020163	2.467184	1.448004		0.2745	1.471607
20.81248	1.164848	2.379257		0.270006	0.557706	1.760485		0.738428	0.498462		0.5985	1.888101	1.554791		0.29695	1.220681	0.964653		0.582325	1.980786
AVERAGE RMSE				AVERAGE RMSE				AVERAGE RMSE			AVERAGE RMSE				AVERAGE RMSE				AVERAGE RMSE	
20.81248	2.650447	2.471589		1.28938	1.035572	1.399186		0.524172	1.043605		0.607841	2.043346	4.591463		2.221566	5.955739	2.498474		0.573181	1.572042

After observing the RMSE values for all the test cases it was revealed that

- Over all best combination of EKF is taking accelerometer as control input while taking Kinect as measurement input, but accelerometer cannot measure yaw values.
- Second best combination was Kinect as control input with accelerometer but again problem with accelerometer is yaw measurement
- Overall best system with least root mean square error is taking IMU fusion data as control with Kinect as measurement

6.9. ITD (Interaural Time Difference) GRAPH

Fig.7 shows the interaural time difference. Interaural time difference defines the difference in arrival time of sound for one ear with respect to the other ear. The graph shows the magnitude value of ITD only. From the graph it is clear that the ITD depends mostly on azimuth value. However, the ITD also shows little variations with elevations. The change of ITD with the change of elevation is very small as shown in Fig. 7.19. The graph is drawn between interaural time difference on x-axis and elevation on y-axis.

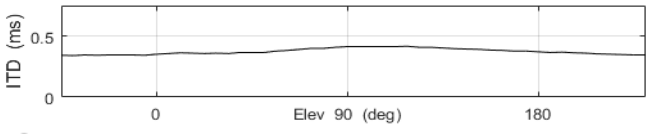


Figure 6.19. Graph Between ITD and Elevation

6.10. HRIR Graph

The graph is plotted between time (x-axis) and HRIR (y-axis) as shown in Fig. 6.20. HRIR for the left ear while red graph is for right ear.

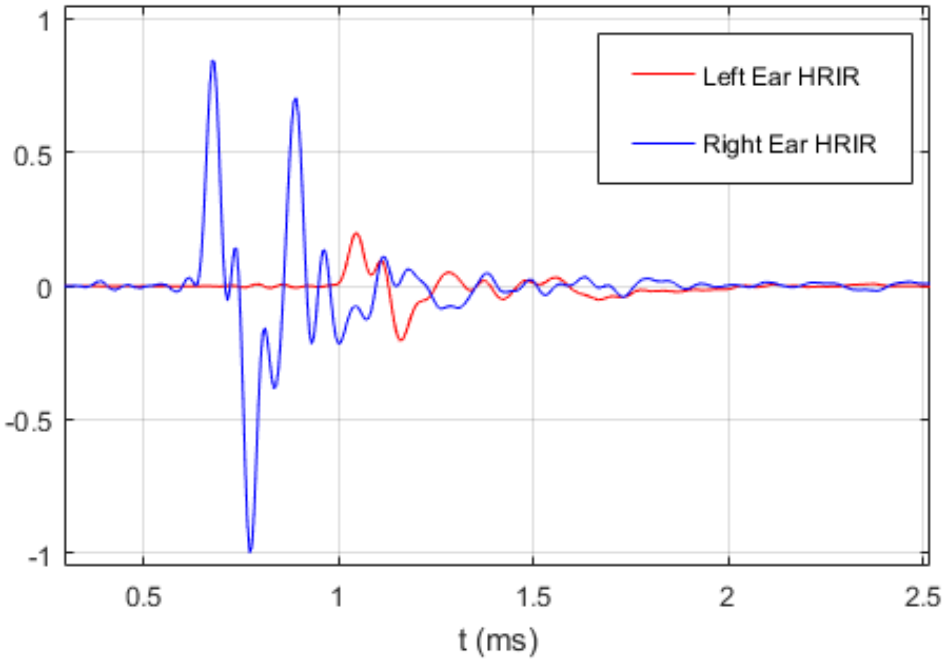


Figure 6.20. Graph Between Time and HRIR

6.11. Left and Right HRTF Graph

Fig. 6.21 shows the HRTF of the left and right ear, the magnitude of HRTF is represented in decibels.

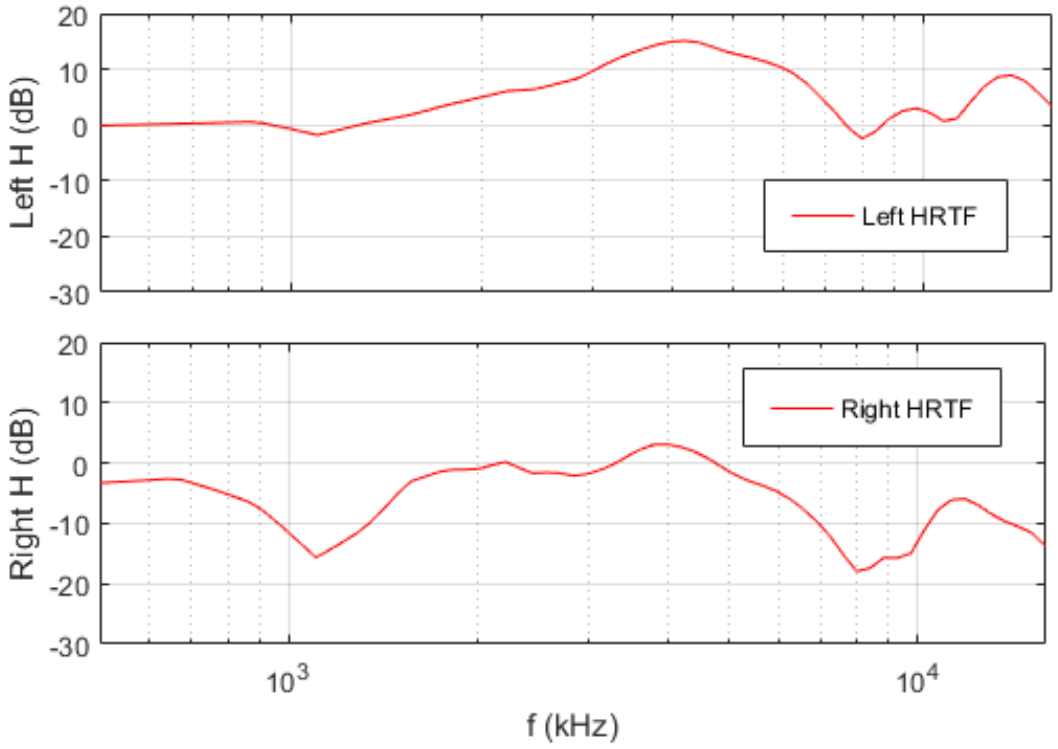


Figure 6.21. Graphs of HRTFs of Right and Left ears

6.12. Summary

- While using Kinect and accelerometer for tracking purposes, both sensors can be fused using six different combinations of EKF.
- Inertial sensor may produce a lot of error in measurement due to gyroscope drift
- Error in roll movement measurement is relatively much higher than the pitch and yaw measurement
- RMSE by taking gyroscope as measurement with Kinect as control is much higher than all other cases

7. CONCLUSION AND FUTURE WORK

The thesis proposed a technique that can be used to generate 3D head tracking based 3D sound generation using IMU and kinect values. Extended kalman is employed to fuse the values from Kinect and inertial sensor. For the fusion of the gyroscope and accelerometer from the inertial measurement unit the complementary filter is used.

The technique was implemented on MATLAB® using a Core i3 machine. The technique utilized every possible combination of control and measurement input between the Kinect and inertial measurement unit. Initial results show the promising future of the technique in head tracking based sound systems.

By analyzing the all combination of the sensors it was noticed that by taking accelerometer as a control input and Kinect as a measurement input the technique showed the minimum root mean square error, but due to limitation in accelerometer for the yaw measurement, this combination cannot be suitable. So, overall the best system for all six degree of freedom is taking the Kinect as the control input and Fusion of gyroscope and accelerometer as the measurement input.

In the future technique can be used to improve many of sound systems in gaming consoles, virtual reality, home theaters and cinemas. As we know that extended kalman filter is computationally expensive so in future the methodology can be further improved by comparing the extended kalman filter results with other techniques like Particle filters, ARMA etc. Although the system is good enough for head tracking to generate 3D sounds, more accuracy for head pose angle measurements is still required that can be further improved by adding more inertial sensors for head tracking in the system or by adding multiple computer vision devices for increasing the field of view.

REFERENCES

1. Tepljakov, s. Astapov, e. Petlenkov, k. Vassiljeva and d. Draheim, "sound localization and processing for inducing synesthetic experiences in virtual reality," 2016 15th biennial baltic electronics conference (bec), tallinn, 2016, pp. 159-162.
2. W. Fitzpatrick, m. Wickert and s. Semwal, "3d sound imaging with head tracking," 2013 ieee digital signal processing and signal processing education meeting (dsp/spe), napa, ca, 2013, pp. 216-221.
3. https://www.engineersgarage.com/articles/magnetometer?page=4&ui=desktop
4. E. F. Helbling, s. B. Fuller and r. J. Wood, "pitch and yaw control of a robotic insect using an onboard magnetometer," 2014 ieee international conference on robotics and automation (icra), hong kong, 2014, pp. 5516-5522.
5. http://www.geekmomprojects.com/gyroscopes-and-accelerometers-on-a-chip/
6. J. H. Kim, n. D. Thang, h. S. Suh, t. Rasheed and t. S. Kim, "forearm motion tracking with estimating joint angles from inertial sensor signals," 2009 2nd international conference on biomedical engineering and informatics, tianjin, 2009, pp. 1-4.
7. P. Gui, l. Tang and s. Mukhopadhyay, "mems based imu for tilting measurement: comparison of complementary and kalman filter based data fusion," 2015 ieee 10th conference on industrial electronics and applications (iciea), auckland, 2015, pp. 2004-2009.
8. J. H. Kim, n. D. Thang, h. S. Suh, t. Rasheed and t. S. Kim, "forearm motion tracking with estimating joint angles from inertial sensor signals," 2009 2nd international conference on biomedical engineering and informatics, tianjin, 2009, pp. 1-4.
9. E. Foxlin, "inertial head-tracker sensor fusion by a complementary separate-bias kalman filter," proceedings of the ieee 1996 virtual reality annual international symposium, santa clara, ca, 1996, pp. 185-194, 267.
10. Eric foxlin, michael harrington, and george pfeifer, (1998). Constellation™: a wide -range wireless motion -tracking system for augmented reality and virtual set applications.intersense incorporated.intersense is-600 user manual.
11. http://www.isense.com/products/prec/is600.
12. Barnard, m. And wang, w., 2016. Audio head pose estimation using the direct to reverberant speech ratio. Speech communication, 85, pp.98-108.
13. R. N. Aguilar, h. M. M. Kerkvliet and g. C. M. Meijer, "high-resolution low-cost ultrasonic tracking system for human-interface systems," 2005 ieee instrumentationand measurement technology conference proceedings, ottawa, ont., 2005, pp. 878-882.
14. http://homepage.tudelft.nl/c7c8y/Theses/PhDThesisPersa.pdf

15. Abidi, m. And chandra, t., (1990). Pose estimation for camera calibration and landmark tracking, proceedings of ieee international conference on robotics and automation, cincinnati, oh, may 13-18, pp. 420-426.
16. C. Rougier, j. Meunier, a. St-arnaud and j. Rousseau, "monocular 3d head tracking to detect falls of elderly people," 2006 international conference of the ieee engineering in medicine and biology society, new york, ny, 2006, pp. 6384-6387.
17. R. B. Mapari and g. Kharat, "real time human pose recognition using leap motion sensor," 2015 ieee international conference on research in computational intelligence and communication networks (icrcicn), kolkata, 2015, pp. 323-328.
18. F. A. Kondori, s. Yousefi, h. Li, s. Sonning and s. Sonning, "3d head pose estimation using the kinect," 2011 international conference on wireless communications and signal processing (wcsp), nanjing, 2011, pp. 1-4.
19. K. Satoh, s. Uchiyama and h. Yamamoto, "a head tracking method using bird's-eye view camera and gyroscope," third ieee and acm international symposium on mixed and augmented reality, 2004, pp. 202-211.
20. Shay ohayon and e. Rivlin, "robust 3d head tracking using camera pose estimation," 18th international conference on pattern recognition (icpr'06), hong kong, 2006, pp. 1063-1066.
21. Abidi, m. And chandra, t., (1990). Pose estimation for camera calibration and landmark tracking, proceedings of ieee international conference on robotics and automation, cincinnati, oh, may 13-18, pp. 420-426.
22. A. O. Ercan and a. T. Erdem, "on sensor fusion for head tracking in augmented reality applications," proceedings of the 2011 american control conference, san francisco, ca, 2011, pp. 1286-1291.
23. http://interface.cipic.ucdavis.edu/sound/tutorial/psych.html
24. en.wikipedia.org/wiki/Extended_Kalman_filter
25. Kalman RE. A New Approach to Linear Filtering and Prediction Problems. ASME. *J. Basic Eng.* 1960;82(1):35-45. doi:10.1115/1.3662552.
26. Carlile,s (1996). "virtual auditory space and applications". Austin, tx, springer.
27. Tashev, ivan (2014). "hrtf phase synthesis via sparse representation of anthropometric features". Information technology and applications workshop,san diego, ca, usa, conference paper: 1–5.
28. Bilinski,piotr; ahrens, jens; thomas, mark r.p; tashev, ivan; platt,john c (2014). "hrtf magnitude synthesis via sparse representation of anthropometric features". Ieee icassp, florence, italy: 4468–4472.
29. https://www.invensense.com/products/motion-tracking/6-axis/mpu-6050/
30. https://developer.**microsoft**.com/en-us/**windows**/**kinect**

Completion Certificate

It is to certify that the thesis titled "A Point Cloud Based HRTF Technique for Analysing the effects of 3D Sounds" submitted by registration no. NUST201464532MCEME35515F, NS Muhammad Usman of MS-82 Mechatronics Engineering is complete in all respects as per the requirements of MainOffice, NUST (Exam branch).

Supervisor: _____
Dr. Khurram Kamal
Date: ____ July, 2017