Dynamic Audio System Based On Listener's Position For Surround Sound Effect

Nachiket Kamod (12), N V Roshni (31),
Ganeshsai Muppasani (30), Jagrati Chowdhari (9),

Guided By,
Prof. Deepali Yewale

Academic Year
2020 - 21

Contents

List of Figures

Chapter 1

1.1 INTRODUCTION

Automation plays a crucial role in the world economy and daily experience. In the last few decades, we have witnessed rapid development in audio systems.

The journey of audio systems begins with a single channel audio system (monaural audio system) in 1877. Later in the year, 1931 a two-channel audio system (Stereo audio system) was introduced and in the year 2005, the most advanced air audio or surround sound system (Multichannel audio system)was introduced.

Modern sound systems are increasingly gaining popularity day by day, remarkably since technological advances have lowered their prices, increased their qualities and features. One barrier to the more extraordinary experience while using these sound systems is its static nature in the surrounding sound. Surround sound involves three or more speakers surrounding the listener to give a surround sound effect by changing the sound source from various speakers.

Although high-end audio systems provide good sound quality but to achieve the surround sound effect, the user must configure the system manually depending upon his current position, which is a very tedious task. Whenever we are settling up a complex home theatre bundle, understanding the art and science of placement of the speaker channels and placement is the most crucial step while setting up a sound system.

The current sound system needs manual setup according to the ideal sitting position to achieve a good sound effect at the fixed position. This manual setup consists of speaker angles and speaker sound adjusted to create and sound pocket around a fixed position. However, this effect varies when we move away from the surround sound pocket created by speakers. This project aims to develop a real-time system to determine the listener's position and distance from the speaker system and adjust the orientations and volume levels accordingly.

1.2 ABSTRACT

To explain the problem statement briefly, consider a real-life scenario. A person configured the orientation and volume levels of my sound system to get perfect surround sound at some position where he usually sit. However, he wants to change his sitting position or even arrangement to some other part of room which can be far right or left or may be forward or backward from the last sitting arrangement. In this case, to get perfect surround sound, he will need to reconfigure speakers again (their orientation and volume levels) as per the new seating position, either with the assistance of a technician or on his own, which are mostly manual adjustments.

To overcome this scenario, we experimented with a combination of stereo vision and hardware technology which responds to real-time movements of the listener and dynamically adjust the sound pocket. This system uses the OpenCV face detection algorithm and simple geometrical formulae to calculate depths and angles for an individual speaker to introduce dynamically adjusted surround sound. Since the system avoids the heavy usage of hardware, complex algorithms, and machine learning approaches, it can be implemented on low-powered microprocessors and the current processors which are being used by sound

systems.

Chapter 2

LITERATURE SURVEY

A. Surround sound systems
(United States Patents On, September 16, 2014)

This paper proposes an idea of the development of a system that comprises of receiver for receiving a multichannel spatial signal that comprises at least one surround channel.

This system comprises a directional ultrasonic transducer for emitting ultrasound towards the surface to reach a listening position via a reflection of the surface and a driver circuit to drive the ultrasonic transducer.

The proposed system is capable of producing virtual surround sound without requiring a speaker to be located.

B. Shadow Sound System Embodied with Directional Ultrasonic Speaker
(ICISA.2013 on 2013)

The paper talks about the usage of the ultrasonic speaker and computer vision system installed on a motorized mount that can freely change the speaker's directions and altitude for a specific registered user.

The resulting system is proven to be able to track the registered user for providing user-selected sound contents without being interfered with by other people.

This method seems promising, but it requires individual hardware for each speaker, and the solution does not cover the implementation on multi-channel audio system efficiently.

C. An Efficient Implementation of Acoustic Crosstalk Cancellation for 3D Audio Rendering
(IEEE China SIP on July 2014)

In this paper, the given method makes the use of ultrasonic speaker and computer vision system installed on a motorized mount that can freely change the speaker's directions and altitude for a specific registered user.

The resulting system is proven to be able to track the registered user for providing user-selected sound contents without being interfered by other people.

This method seems promising, but it requires individual hardware for each speaker and the solution does not cover the implementation on the multichannel audio system efficiently.

D. Multi-rate adaptive filtering for immersive audio
(IEEE Xplore on February 2001)

This paper describes a method for implementing immersive audio rendering filters for single or multiple listeners and loudspeakers.

In particular, the paper focuses on the case of a single or two listeners with different loudspeaker arrays to determine the weighting vectors for the necessary FIR and IIR filters using the LMS (least-mean-squares) adaptive inverse algorithm.

It describes the transform-domain LMS adaptive inverse algorithm that is designed for crosstalk cancellation necessary in loudspeaker-based immersive audio rendering.

The algorithm used in this paper is only for a single listener and for only two loudspeakers.

2.1 CONCLUSION

High-end audio systems provide very high-quality sound, and it also provides the user with the feasibility to use them for multiple events. But even the best has some drawbacks.

- The complexity of hardware, as per the research study, we can observe the research is based on mono channel and not multi-channel.
- Research suggests the use of Kinect for object tracking, which comes with its drawbacks.
- The Algorithm's complexity to achieve the effect, researchers suggested very complex algorithms regardless of room geometry. Hence, they are challenging to understand and hard to implement on low-cost hardware.

Chapter 3

AIM AND OBJECTIVES

3.1 Aim

To develop a real-time self-adjusting Audio system based on listener's position to achieve a high-quality air sound effect.

3.2 Objectives

1. To introduce automation and artificial intelligence into the current trend of audio systems.

2. To make the audio system compatible with adjusting it's orientation and sound intensity based on the listener's position.

3.3 Methodology

From the abstract, we can conclude that speaker angles and sound intensities of individual speakers are essential.

Speaker angles define how the sound is going to reach the listener. Like is it reflecting from any surface, or the sound sources are directly pointed towards the listener.

Sound is nothing but oscillations of particles (typically air) in vibrational motion, which transports energy through a medium.

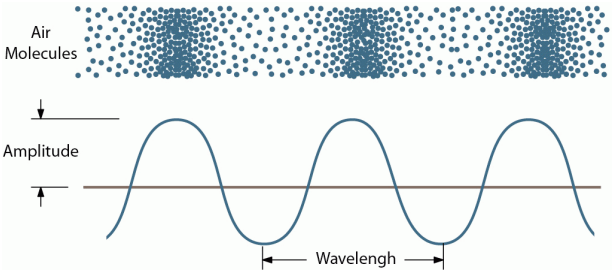

Figure 3.1: Sound Waves

Speakers push and pull surround air molecules in waves to generate a sound wave using a diaphragm. Typically, this diaphragm is in conical shape; hence it oscillates the molecules in the oval field.

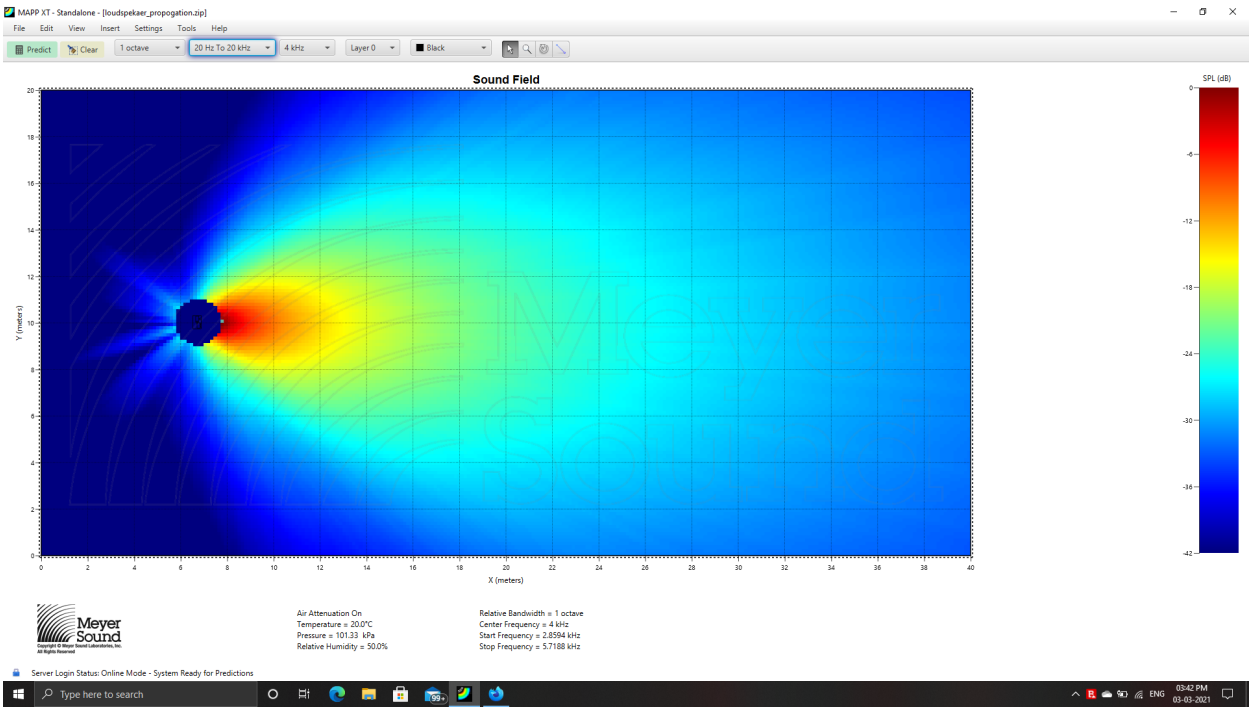

Figure 3.2: Sound propagation of a speaker

Figure 3.2 shows the oval propagation of the sound field from the speaker. The speaker's sound intensity at some depth is denoted with a heat map (dB). In front of the speaker, the sound intensity is maximum, and it fades away as we go far away from the speaker. Where, on the other hand, it is much less at the back of the speaker. Since it is oval, the propagation to left and right is also less than that of the front.

Figure 3.3 shows the reflection and reverberation of sound due to misalignment and the speaker's excessive sound intensity due to the collision of sound waves onto walls. These reverberations decay as they get absorbed by the surfaces of objects and walls in the room. In this case, the listener tends to hear the direct and repeated sound waves, which might sound muddy and grabbled.

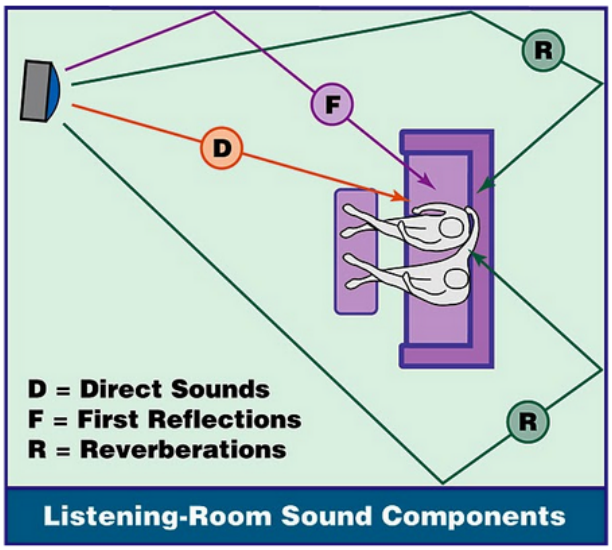

Figure 3.3: Reflection and reverberation of sound

Hence, it becomes necessary to align the speakers and adjust the sound levels in proper amounts to get the best surround sound.

To overcome this scenario, we experimented with a five-block system, which consists of the following.

1. Depth estimation unit (Cameras), to measure the listener's depth from one reference point and feed these variables to the microprocessor for further calculations.
2. Microprocessor, to measure depth and calculate panning and tilting angles and listener's depth from each speaker.
3. Mechanical unit, to pan and tilt the speakers.
4. Audio Processor Unit (Digitally controlled amplifier) adjusts the individual speaker gain using calculated results from the microprocessor.
5. Speakers, to sound individual 4-channeled output.

3.4 Specifications of the System

3.4.1 Depth estimation unit

Web Cam

(LAPCARE LAPCAM)

Figure 3.4: Web cam

1. 1280 x 720 pixels @ 720p resolution
2. Automatic low light correction
3. Plug and play Linux compatible, High-Speed USB 2.0

3.4.2 Microprocessor

Microprocessor serves an important role in DAC application, data processing estimation and controlling the response hardware in real time.

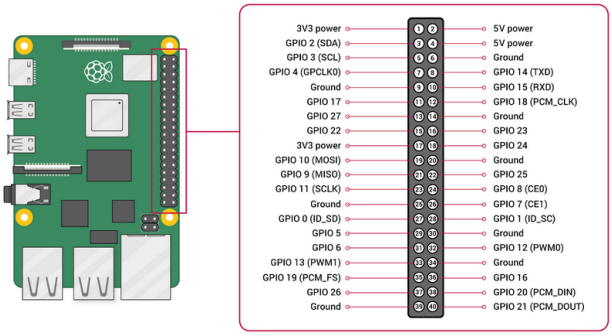

Figure 3.5: Microprocessor (Raspberry Pi 4B 4GB)

Raspberry Pi 4B 4GB RAM model comes packed with,

1. Quad core Cortex-A72 64-bit @ 1.5 GHz clock and uses ARM v8 architecture, with 4GB LPDDR4-3200 SDRAM.
2. 2.4 and 5 GHz IEEE 802.11ac wireless wifi hardware.
3. 2 Micro HDMI ports.
4. H.265 (4kp60 decode), H264 (1080p60 decode, 1080p30 encode).
5. OpenGL ES 3.0 graphics.
6. Micro-SD card slot for loading operating system and data storage.
7. 4 USB ports.
8. Software PWM on all pins and Hardware on GPIO12, GPIO13, GPIO18, GPIO19.

9. SPI

- SPI0 : MOSI (GPIO10), MISO (GPIO09), SCLK (GPIO11), CE0 (GPIO08), CE1 (GPIO07)
- SPI1 : MOSI (GPIO20), MISO (GPIO19), SCLK (GPIO21), CE0 (GPIO18), CE1 (GPIO17), CE2 (GPIO16).

3.4.3 Mechanical unit

Servo Motors

(SG90 Servo)

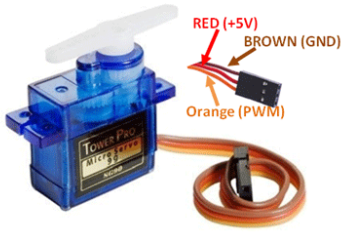

Figure 3.6: SG90 Servo

1. 180˚ rotation (90˚ in each direction).

2. Torque 2.5 kg-cm

3. Voltage 4.8-6 V

4. Speed 0.12 sec/60˚

3.4.4 Audio processor unit

Audio Amplifier

(LM386N-1)

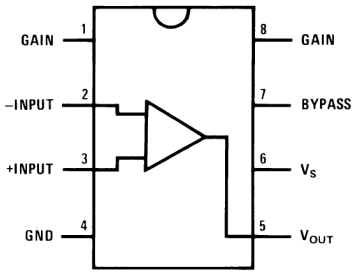

Figure 3.7: LM386N-1 pin-out

1. Operating Supply Voltage (Vs) 4 - 12 V

2. Voltage gain 20 - 200

3. Output power 325 mW

Digital Potentiometer

(MCP42010)

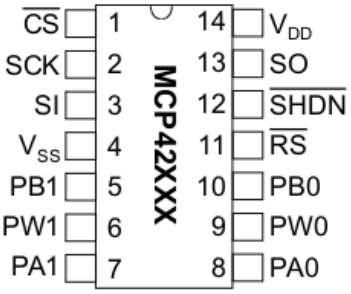

Figure 3.8: MCP42010 pin-out

1. Potentiometer values 10 kΩ

2. 256 taps for each potentiometer

3. 2 channel

4. SPI serial interface (mode 0, 0 and 1, 1)

5. Single power operation (2.7V - 5.5V)

6. Industrial temperature range: -40˚C to +85˚C

7. External temperature range: -40˚C to +125˚C

3.4.5 Speakers

The final component is the speakers, which are mounted on servos and the sound is dynamically controlled using Audio Processor. They have mounted on four corners of the room to form a four-channel audio system.

For this application, we are using 4Ω speakers to deliver four-channel output.

Chapter 4

BLOCK DIAGRAM OF THE SYSTEM

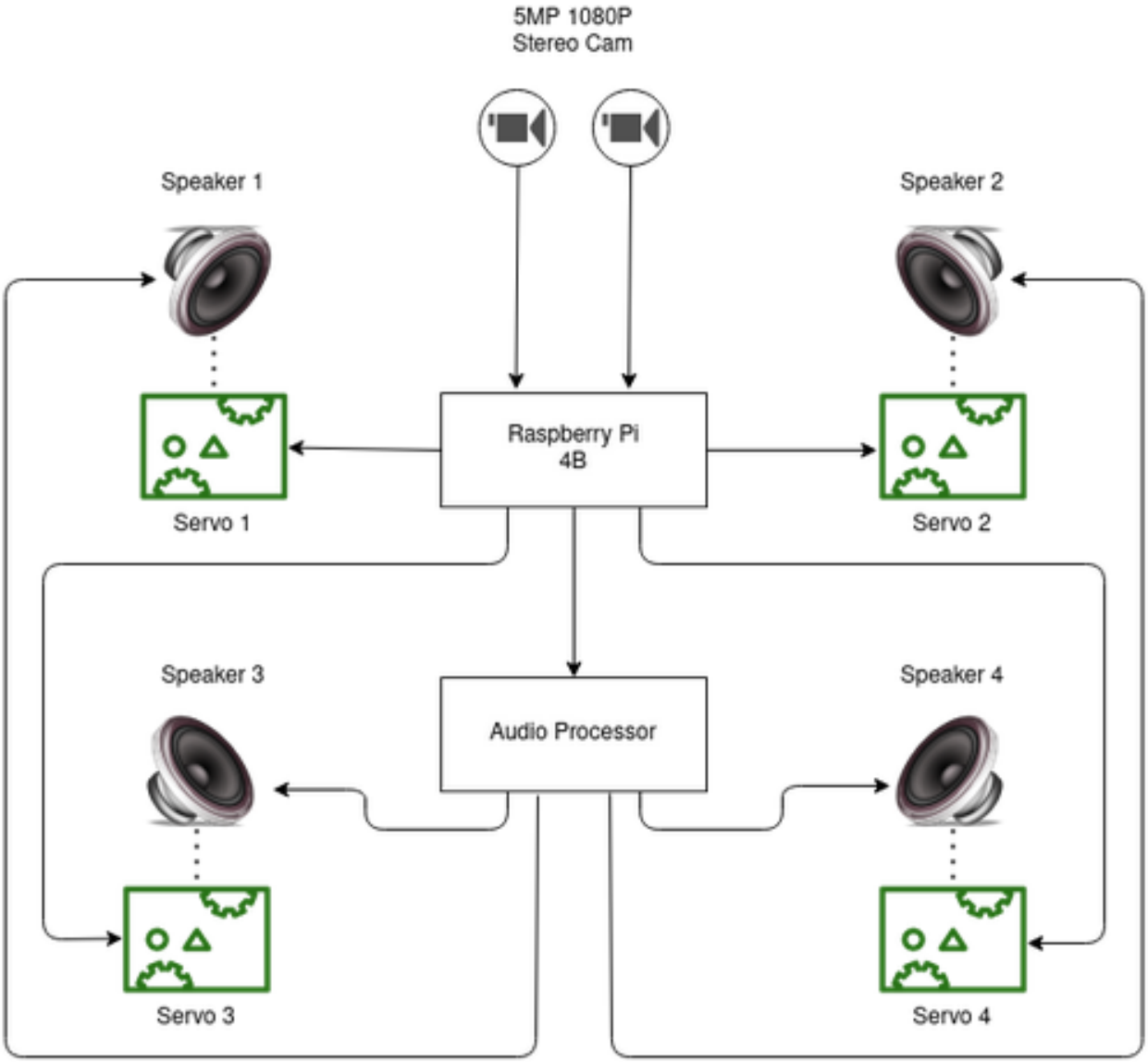

Figure 4.1: Block diagram

The dynamic audio system can be simplified as five measure blocks, each serving its own application in order to provide a dynamic surround pocket over the listener's head,

4.1 Depth estimation unit

Depth estimation unit includes stereo vision assembly of two web cams with similar (known or unknown) focal length (In our case 18 cm).

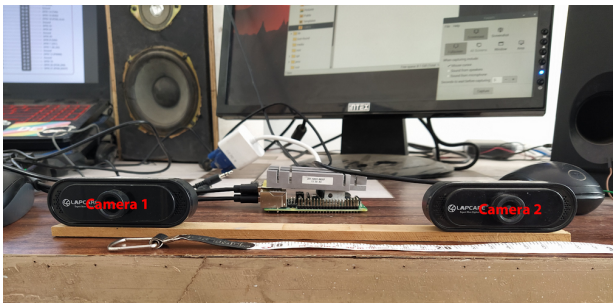

Figure 4.2: Assembly of the depth estimation unit

The cameras are placed at some known distance x from each other (24 cm) and computer vision (opencv) and geometrical equations have been implemented to measure the depth, deviation and height of an object from a reference point, where the reference point is the center of the stereo vision system.

4.2 Micro-processor

The Microprocessor acts as middleware between DAC and the hardware to control speakers. In this project we are using raspberry pi 4B (Quad core Cortex-A72 64-bit @ 1.5 GHz clock) as our microprocessor. Webcams are connected through USB 2.0 of RPi.

First we find depth, deviation and height of the listener's face from the reference point using OpenCV's frontal face Harr cascades classifier followed by open source AA symmetry algorithm of geometry, where OpenCV also helps to classify between listener and other objects.

The above process provides us three real-time variables,

1. Depth of the listener's face from the reference point.
2. Deviation of the listener's face from center of the axis.
3. Height of the center point of listener's face from the origin (reference point).

Further using this real-time variables, and some constants (room dimensions and speaker positioning), using custom designed geometrical algorithm we can calculate the panning, tilting angles and depth of the listener from each speaker.

Using panning and tilting angles we can rotate the servos to the required angles to direct the sound field towards the listener. And using depth we can adjust the sound levels of the speakers by controlling input voltage of amplifiers digitally.

4.3 Mechanical Unit

As discussed in the methodology, the speakers propagates sound in elliptical. Hence we need to align the major axis of the sound field towards the listener.

The mechanical unit is consisting of two servo motors for each speaker (channel), one for panning and the second for tilting,

Servos are connected to hardware PWM pins of the RPi and controlled in real time using feedback of the angle algorithm.

4.4 Audio Processing Unit

Usually, a surround sound system contains two or more speakers to generate a sound effect of moving objects from one place to another.

Even if we direct the speakers towards the listener's direction, it is necessary to adjust each speaker's sound levels according to the depth of the listener from each speaker.

The audio processor unit assembles with 4 class AB audio amplifiers driven by two 2-channel digital potentiometers for controlling each speaker's sound levels to adjust the sound pocket over the listener's head (ears).

4.4.1 Audio amplifier

An audio amplifier is a circuitry designed to increase the applied signal's magnitude to power a low resistance load (speakers).

Sound signals are applied to the non-inverting terminal of an amplifier through a voltage divider circuitry (potentiometer). This voltage divider adjusts the input signal's voltage levels, resulting in a change in volume levels at the output. This change is inversely proportional to the resistance at the wiper terminal of the voltage divider.

For this application, we are using LM386N-1 as our amplifier.

4.4.2 Digital potentiometer

Digital potentiometers mimic the analog functions of a mechanical potentiometer where microcontrollers or microprocessors control the resistance.

As we discussed, to adjust the audio amplifier's sound output, we adjust the input voltage given to the non-inverting terminal of the amplifier. Hence, we supply the audio signal to the amplifier through a digital potentiometer to increase and decrease input voltage and, hence, the speaker digitally using a microcontroller or a microprocessor.

For this application, we are using SPI-compatible MCP42010 Digital POT.

4.5 Speakers

Speakers serve the four channelled dynamically adjusted surround sound to the listener. Usually, they are mounted on four corners of the room, either at the listener's ear levels or near the ceiling.

Chapter 5

HARDWARE DESIGN

Hardware design is carried out in four phases,

1. Mathematical model.

2. Simulation and verification of algorithm.

3. Calibration of sensor.

4. Hardware Design.

5.1 Mathematical model

The mathematical model serves the most important role throughout the project, as it is intended to solve the issues that persisted in previous research.

This model is further divided into two sub-parts,

5.1.1 Listener tracking model

The listener tracking model is a combination of face detection and stereo vision technique for estimating depth.

Face detection

We must classify and sort out the entities from the rest of the objects from the surroundings to align the speakers properly.

In our case, these entities are people who are listening to the system. To classify them from other objects from surroundings, we implement the Harr cascade face detection algorithm to sort and cluster out these entities.

Harr cascades is a cascade classifier that implements a machine learning approach based on the Adaboost meta-algorithm.

The rectangular shape of the face is meaningful in initializing the classifier. Further, the algorithm focuses on the property that the eyes region is often darker than the face and nose region. The second feature proposes that the eyes are darker than the bridge of the nose. Similarly, this approach finds the entity's possible relations and features and records the features for further prediction.

Once the face is detected, we can obtain the face location from the origin (center of the image).

Stereo Vision

Stereo vision compares the information about a scene from two vantage points and examining relative positions of objects in the two panels.

An image can be termed as a grid of pixels within some range of indices. Using face detection, we narrowed down the object's position in the grid of pixels (x, y).

Stereo vision gives two images of the same scene from different positions in the same plane. Each image gives the disparity of the face from the origin of that respective image.

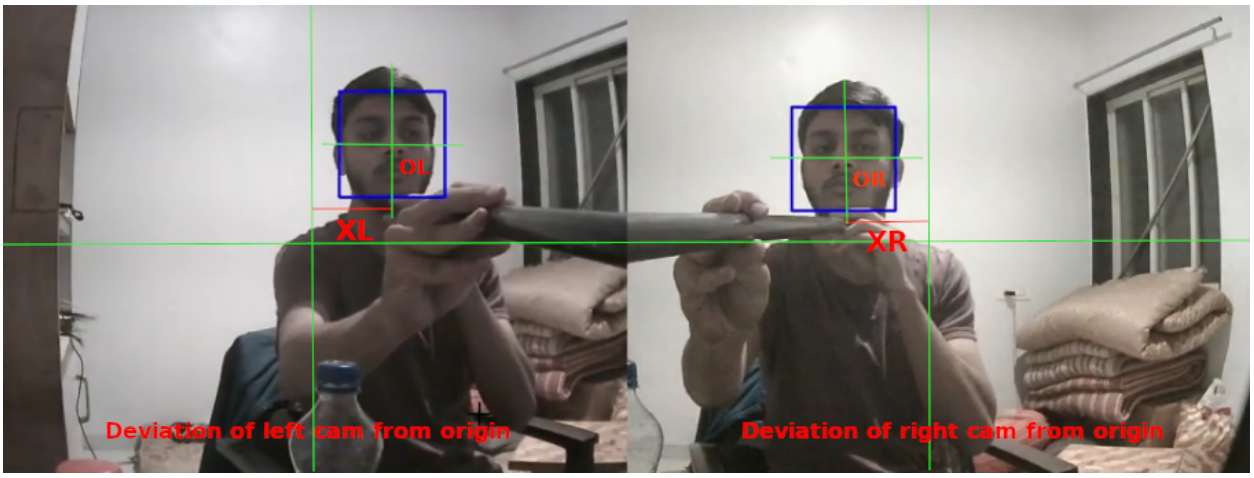

Figure 5.1: Stereo vision showing the disparity from origins of both the cameras

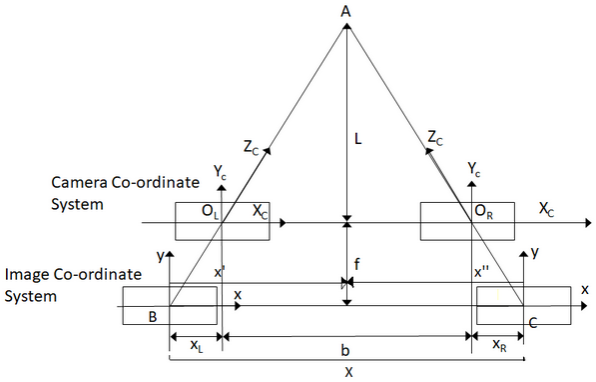

Figure 5.2: Depth sensing geometrical model

Figure 5.1 shows us the geometry behind the stereo vision method for depth sensing.

Here,

x = Distance between two webcams.

f = Focal length of the webcams.

X_L = Disparity of the image from the origin of the left webcam.

X_R = Disparity of the image from the origin of the right webcam.

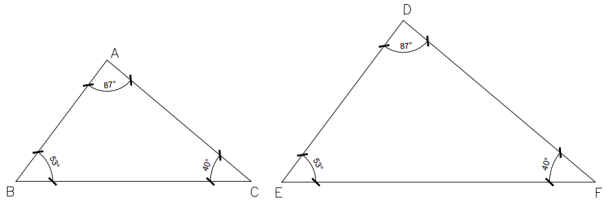

Figure 5.3: AA symmetry criterion

Stereo vision depth-sensing works on the principle of angle-angle symmetry (AA symmetry criterion) of two triangles. Using the AA symmetry criterion, we can state that if angle of the two triangles are congruent, then the third angle of both triangles must be the same. Hence, the ratio of each parallel side of triangles is equal.

i.e.,

$$\frac{AB}{DE} = \frac{BC}{EF} = \frac{AC}{DF} \tag{5.1}$$

Hence from figure 5.2, we can prove that,

$$\frac{f}{z} = \frac{X_L}{X'} \tag{5.2}$$

Where, $z = f + L$,

Similarly,

$$\frac{f}{z} = \frac{X_R}{X''} \tag{5.3}$$

From equation 5.2 and 5.3, we can say that,

$$x' = \frac{z \times X_L}{f} \tag{5.4}$$

$$x'' = \frac{z \times X_R}{f} \tag{5.5}$$

From figure 5.2 we can say that $x = x' + x''$,

$$\therefore x = \frac{z \times X_L}{f} + \frac{z \times X_R}{f}$$

$$\therefore x = \frac{z}{f} \times (X_L + X_R)$$

Finally, we get depth (z),

$$\boxed{z = \frac{x \times f}{X_L + X_R}} \tag{5.6}$$

Focal length

In the formulae, we came across the term 'f' (Focal length). The focal length of the lens is the distance between the lens and the image sensor when the subject is in focus, usually stated in millimeters (e.g., 28 mm, 50 mm, or 100 mm).