Gesture Controlled Audio System

ECE445 Design Document

Kehang Chang (kehangc2) Ruofan Chen (ruofanc2)
Ruohua Li (ruohual2)
Team 37
TA: Schroeder, AJ
3/3/21

1. Introduction

1.1 Problem and Solution Overview:

1.1.1 Problem Statement

When a person is cooking or working, it can be hard to interact with an audio system. Specifically, a person's hands are rarely free when cooking or working. Additionally, a noisy environment can render traditional "smart assistants" and their voice commands useless. A gesture-based system would be much more useful in such a situation. It is also very common when a person is trying to connect more speakers to enhance the listening experience during social gathering events, but he or she just doesn't have the right types of smart speakers to pair several speakers together. Smart speakers that are able to be paired together are usually expensive as well. We designed a cheaper way to distribute music without requiring any modern smart speakers. In other words, a basic magnetic speaker would be sufficient to bring the stereo effect to the end users. Thus a gesture controlled audio system with full stereo capability would be appealing to many users. There has not been an existing product in the market right now which would offer the convenience of both features.

1.1.2 Solution Overview

Gesture Controlled Audio System is an audio sharing and coordination system. This system aims to provide users with a handy way of controlling audio systems by enabling remote control using human gestures. Our proposed system consists of three subsystems: (1) human gesture capturing and recognizing system (vision subsystem) which employs a camera along with an embedded system to segment human gestures and convert them to control signals in real time, (2) distribution and receiving system (transmission subsystem) which contains one broadcaster and multiple receivers, and (3) signal processing and output system (audio subsystem) which process the data received by each receiver and send the signal to speakers. The setup requires no pairing procedure and music tracks are automatically synchronized. The whole audio system is controlled by human gestures from the master node.

1.1.3 Justification Reference

Gesture Controlled Audio System aims to provide an easy-to-control solution for the wireless speaker control problem. There actually exist some commercially available integrated solutions like Alexa from Amazon and HomePod from Apple. These solutions commonly receive user control signals via vocal instructions and distribute audio signals via Bluetooth. Based on CNET review, Alexa is awoken by using one of these three words: "Alexa, Amazon or Echo" [5]. Currently, voice control is still the most dominant way of interacting with these kinds of smart speakers. Also these kinds of speakers are not cheap. The newest released version of Alexa: All-new Echo (4th Gen) costs 80 dollars [6]. The newest released version of HomePod costs 299 dollars [7]. For music lovers, these expensive smart speakers offer the same music quality as a

15-dollars full-range 8’’ speaker does, and fancy features such as smart sharing and auto sync offered by those smart speakers are definitely overpriced [8]. Gesture Controlled Audio System utilizes all kinds of speakers and offers the same quality of music and easy way of distributing music at a far cheaper price.

Compared to these products, we plan to take different paths by designing the system to receive user input via captured and recognized human gestures and distribute audio signals via external RF receivers plugged into the speakers. These choices give us numerous advantages. (1) Our design offers better robustness in noisy environments since it is vision-enabled. (2) Our design provides better accuracy controlling from long distance since given enough resolution the accuracy of vision recognition will remain undiminished while vocal control accuracy will be impaired. (3) Our design is compatible with more devices since its connection is external for each device. (4) Our design offers the potential of human triangulation with respect to each audio device, which can lead to further and more accurate amplitude distribution to provide better stereo effect for users.

1.2 Visual Aid

The visual aid Figure 1 is used to provide a general overview of the design.

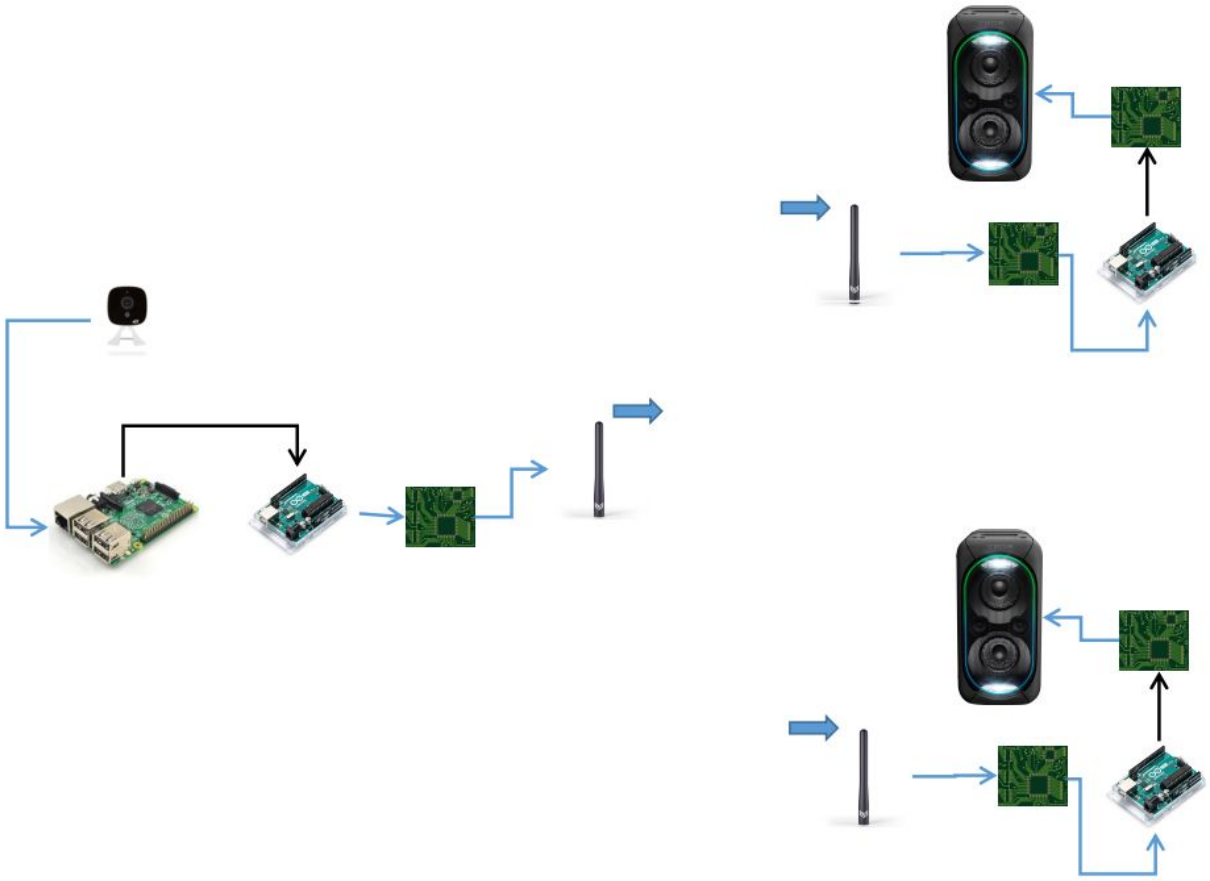

Figure 1. Visual Design of GCAS (Gesture Controlled Audio System)

1.3 High-Level Requirements List

- Visual Characteristics
  The vision subsystem must show its ability to recognize human gesture with precision and recall > 80% to accurately determine gestures. The gestures so far including play/pause, next track. And our vision subsystem should have a responding time less than 5 seconds.

- Audio Characteristics
  The audio subsystem should be able to distribute and play music with 8-bit resolution. Noise out of [20Hz, 20KHz] range should be filtered properly such that the noise/authentic signal power ratio < 25%.

- Transmission Characteristics
  The transmission subsystem should be able to deliver correct control signal and audio signal with lagging < 1s. To be more precise, the system should complete its entire signal processing pipeline within 1 second: transmit control signal to receiver and execute control signal.

2. Design

2.1 Physical Design

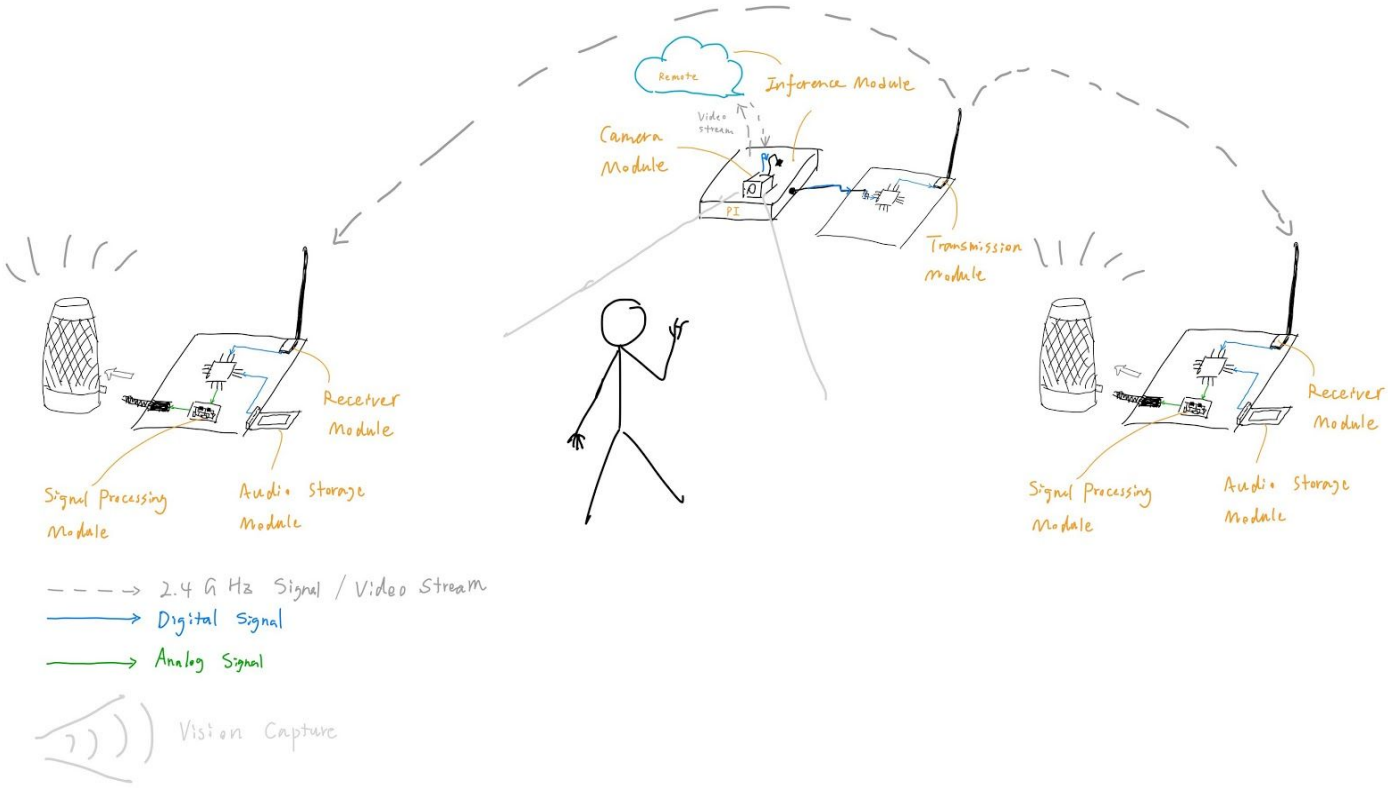

Figure 2. Physical Design of GCAS (Gesture Controlled Audio System)

2.2 Block Diagram

The vision subsystem is designed to capture human gesture by constantly observing humans in its eyesight and doing inference to control the audio signal that should be played. The transmission subsystem will preprocess the control signal determined by the vision subsystem and transmit it to audio subsystems. Audio subsystems will then process the signals transmitted and play audio signals from the audio storage module accordingly.

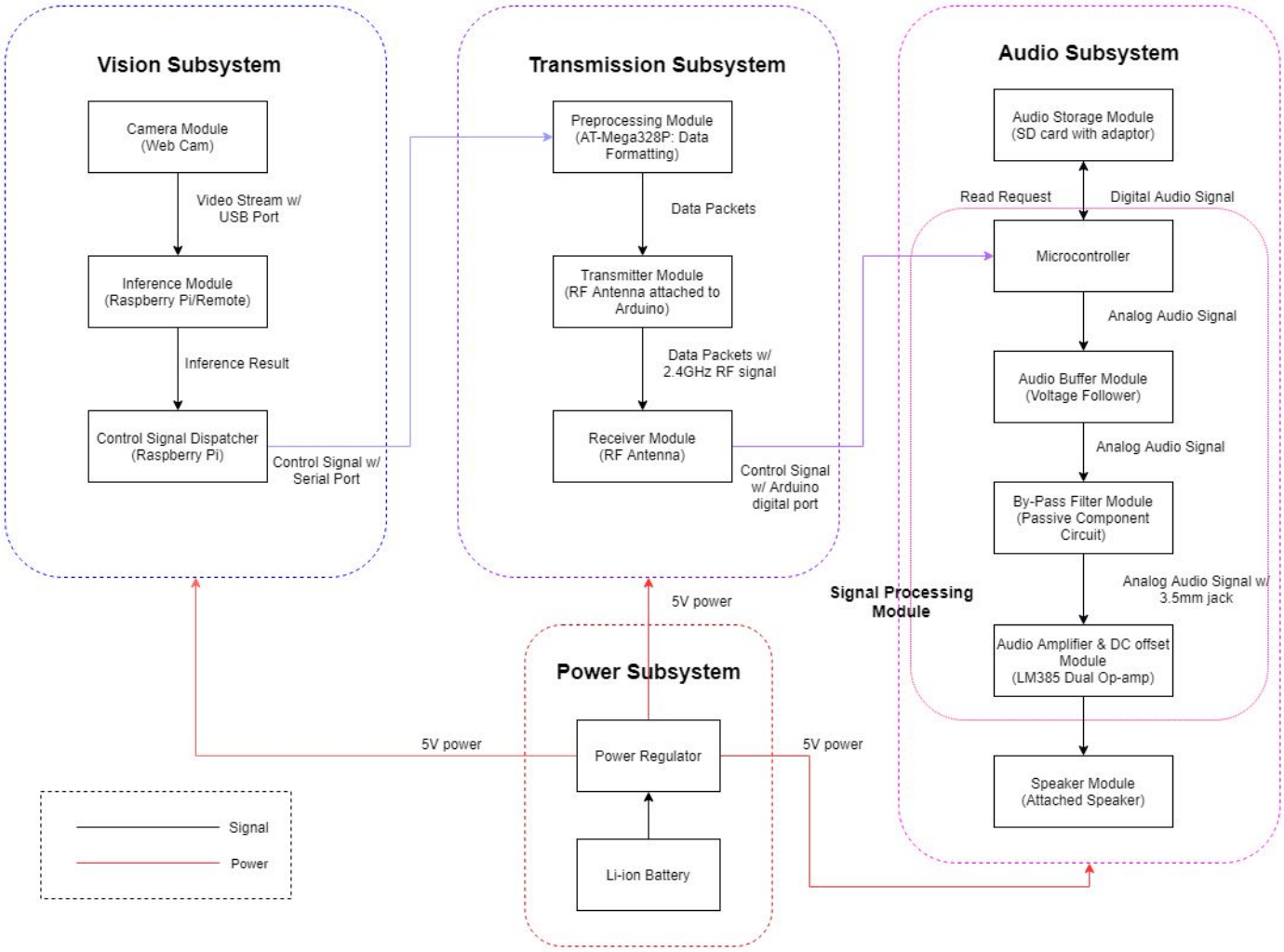

Figure 3. Block Diagram of GCAS (Gesture Controlled Audio System)

2.3 Subsystem Description

2.3.1 Vision Subsystem

Our vision subsystem will accurately detect and recognize human gesture via video stream and detection pipeline. The camera module will intake a video stream consisting of each frame with frame per second determined by the specific model of camera. Each frame, then, will be fed into the inference module in order to recognize any potential control gesture made by users. The inference result will be used to process the control signals sent to the transmission subsystem. The signal will be sent by serial port (usb transmission).

The vision subsystem is composed of a camera module attached to a processing module that is constantly capturing video data in its visual range and doing inference on each frame. Once a

valid gesture is captured, the vision subsystem will change the control signal sent to the transmission subsystem. So the ability of correctly differentiating valid gestures from invalid ones is critical to the entire system. If the vision subsystem fails to correctly recognize valid gestures with precision and recall > 80%, the system will fail to adjust its behavior when the user tries to control it.

2.3.1.a Camera Module

The camera module will capture a video stream at 30fps in a 1280x720 resolution. The camera module is mounted on the raspberry pi to feed input into the inference module (either locally on the raspberry pi board or remote server).

Requirement	Verification
The camera is fix-mounted and capture video stream at 30fps	A python script is used to trigger the camera to start to take video inputs

2.3.1.b Inference Module

The inference module is either a raspberry pi board or a remote server. The inference module will analyze the images captured by the camera module to output proper control signals based on its recognition of valid gestures.

Requirement	Verification
Detect a valid gesture (pause, resume, next song) based on images captured in a 30fps-capable surface-mounted camera	Do 100 gestures with 50 valid gestures and 50 invalid gestures, and according to Monte-Carlo method if we can obtain 40/50 correct response for valid gestures as well as at least 80% of valid gesture detected is actually valid, then the model is running as expected.

2.3.1.c Control Signal Dispatcher

The control signal dispatcher is handled by a software code running on the Atmega328P microcontroller to interpret the output from the inference module and send out control signals to the transmitter module.

Requirement	Verification
The dispatcher should complete the delivery	Input valid gestures such as play, pause, next

pipeline in less than a second.	to the camera module, and the delivery process to end device speakers should be less than in a second.

2.3.2 Transmission Subsystem

Then our transmission subsystem will receive signal from the vision subsystem and process it to transmittable data packets. Then it will send each audio subsystem their respective data packets, e.g. if there are two audio subsystems, two data packets representing each track for a piece of stereo music will be sent to each audio subsystem. The data packet will be sent by a designed data structure via 2.4GHz RF signal.

The transmission subsystem takes signal from the vision system and reorganizes the signal into data packets then sent to each speaker. The transmission system must be able to process data and transmit the data fast enough, otherwise the system will experience lagging and distortion.

2.3.2.a Preprocessing Module

This module is functioning on an ATMega chip which takes in the control signal sent by the control signal dispatcher and reformats it into wireless-transmittable data packets.

Requirement	Verification
This module will be able to take in serial port data and transmit it into transmittable data packets.	Print out the data sent by the control signal dispatcher and output of this module to see if the functionality is correct.

2.3.2.b Transmitter Module

This module is a circuit involving a RF antenna that sends control signal wirelessly over to the receiver module. It takes signals from preprocessing modules by wires.

Requirement	Verification
This module will be able to send the data packets sent by the preprocessing module via 2.4GHz RF signal.	Print out the data sent by the preprocessing module and received by receiver module to see if the data is correctly sent.

2.3.2.c Receiver Module

This module is a mirrored version of the transmission module in a way that it takes in data packets instead of sending them out.

Requirement	Verification
This module will be able to receive the data packets sent by the transmission module via 2.4GHz RF signal.	Print out the data sent by the transmission module and received by receiver module to see if the data is correctly received.

2.3.3 Audio Subsystem

Our audio subsystem, upon receiving control signals, will process the data packets sent from the transmission subsystem. Then it will perform D/A conversion and filtering to produce the desired analog signal for the speaker. Then the analog signal will be sent to the speaker by a 3.5mm audio jack and become output of the system.

The audio subsystem will receive digital packets from the transmission system and perform D/A conversion then send the analog signal to the speaker. The audio subsystem must filter out unwanted noise from other sources, otherwise the output of the system will lose the authenticity.

2.3.3.a Audio Storage Module

This module stores the audio signal that will be played by speakers, usually songs on a SD card. It is connected to the microcontroller via a specially made adaptor circuit.

Requirement	Verification
This module will be able to be read by the microcontroller file by file in order.	Check the content of this module and the result of reading by the microcontroller to see if it can be read correctly.

2.3.3.b Microcontroller

This module is an ATMega chip that takes in control signals from the receiver module as well as audio signal from audio storage module. The reading will be performed per the request of the control signal. It should output raw analog audio to the signal processing module.

Requirement	Verification
This module will be able to read audio signals according to the request of the control signal.	Apply a fixed control signal to this module and print out the data reading in by this module to check if this module is performing reading according to control signal' request.

2.3.3.c Signal Processing Module

The signal processing module consists of a DAC buffer, a by-pass filter, an audio amplifier, and a DC offset circuit. The DAC buffer is a voltage follower that is used to protect the audio signal coming out from the microcontroller. If an audio signal is used to drive the speaker and then it will distort the signal. The by-pass filter will filter out both high and low frequencies noises. The audio amplifier is used to amplify the signal and the DC offset circuit is used to let the audio signal oscillate around 2.5V rather than 0V.

Requirement	Verification
The audio signal coming out from the signal processing module should be without high and low frequency noises and oscillate around 0V.	Probe the PWM port on the Atmega328P board and output connection port from the signal processing module to compare the signals on the oscilloscope.

2.3.3.d Speaker Module

This module is a speaker connected to the signal processing module via 3.5mm audio jack.

Requirement	Verification
This module will be able to play analog audio signals outputted by the signal processing module.	Apply an audio signal from a device e.g. phone and computer with a 3.5mm jack then play music to see if the music is played accordingly.

2.3.4 Power Subsystem

Our power subsystem will source a constant 5V for camera module and inference module and 3.3V for transmission module. The power subsystem will try to incorporate a secure protection mechanism to prevent a short circuit happening.

The voltage source is provided by a four-pack Li ion battery. The 3.3V and 5V power supply is regulated by Atmega328P chips and other microcontrollers. The safety mechanism will incorporate a fuse to protect any potential hazards caused by a short circuit.

2.3.4.a Power Regulator Module

The power regulator module should be able to take a voltage supply ranging from 7V~14V and output a steady 5V voltage for any module on the PCB board.

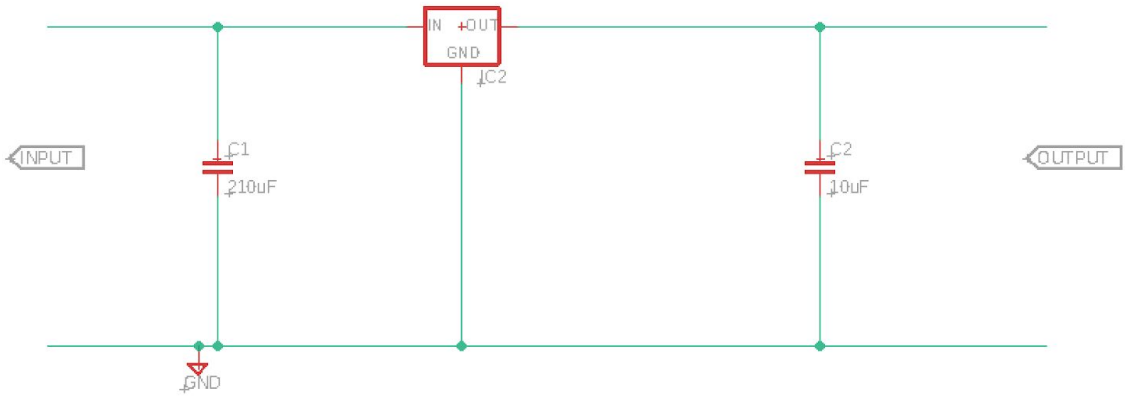

Figure 4. Power Regulator Schematics

Requirement	Verification
Output a 4.95~5.05V voltage with a maximum current of 50mA when a 7.2~14V input voltage is applied	A. Connect a 5 pack Li-ion battery pack to the power regulator module B. Probe the power line coming out of power regulator module to measure the supply voltage

2.3.4.b Li-ion Battery

This module is a battery case with 5x1.5V batteries that supply power to all modules.

Requirement	Verification
This module needs to output 6.5V~8.5V voltage.	Measure the output voltage to check if it is supplying power correctly.

2.4 Tolerance Analysis

2.4.1 Problem Formulation

Our inference module tries to infer based on multiple consecutive frames taken by the camera module. Specifically, let $f(.)$ denotes our inference module that takes in a frame I_{t_i} that in R^{mxn} ,

where m and n denotes the height and width of a frame and t_i denotes the time when the frame is taken by the camera. To simplify the problem for the sake of brevity, let $f(I_{t_i})$ output a Boolean value indicating whether frame I_{t_i} contain a valid gesture, i.e., $f(I_{t_i})$ in $[True, False]$. And let's define inference on multiple frames to be denoted by $f(I_{t_{1,\ldots,}} I_{t_r})$ and it is equal to $f(I_{t_1}) AND ... AND f(I_{t_r})$, so if a single frame from time step t_1 to t_r does not contain a valid gesture, then the output will be $False$. Also let $FP(.)$ denote the false positive probability of an inference and $T(.)$ denote the time of inference.

2.4.2 Tradeoff

So it is obvious that there exists a trade off between responding robustness and inference time (overall responding time). Our inference module will decide the next control signal based on r consecutive frames, i.e., $f(I_{t_{1,\ldots,}} I_{t_r})$. If we respond on fewer frames to decide what the next command is going to be, i.e., r is relatively small, we can respond faster to each user input, since $T(f(I_{t_{1,\ldots}} I_{t_r})) = T(f(I_{t_1})) + ... + T(f(I_{t_r}))$. However this strategy will also potentially increase the false positive rate of the model since users, from time to time, could accidentally pose valid gestures temporarily when they do not mean to trigger a control signal. Precisely, $FP(f(I_{t_{1\ldots}} I_{t_r})) = FP(f(I_{t_1}))...FP(f(I_{t_r}))$.

2.4.3 Design Choice

Currently we are choosing $r = 3$, since we have an inference per frame about 1.5 to 2.5 seconds and false positive rate per frame from 30% to 40%. So this gives us

$$FP_{min} = 30\%*30\%*30\% = 2.7\%$$
$$FP_{max} = 40\%*40\%*40\% = 6.4\%.$$
$$T_{min} = 1.5*3 = 4.5 \text{ seconds}$$
$$T_{max} = 2.5*3 = 7.5 \text{ seconds}$$

3. Cost and Schedule

3.1 Cost Analysis

Labor: based on the graduate student wage at University of Illinois, we are getting paid $20/hour. Our total labor cost is

3 people * $20/hour *10hour/week *16weeks = $9,600

Parts:

Part	Cost
Atmega328p-pu from microchip	$20.39

CONN PLUG STEREO 3.5MM RA 3COND (earphone jack)	$13
HiLetgo 5pcs Micro SD TF Card Adater Reader Module	$6.99
SanDisk 32GB Ultra SDHC UHS-I Memory Card	$8.99
ELP megapixel Super Mini 720p USB Camera Module	$29.9
E-Projects B-0004-H15 Ceramic Disc Capacitor, 50V, 0.01uF, 103 (Pack of 25)	$5.69
National Semiconductor LM386N-1 Semiconductor, Low Voltage, Audio Power Amplifier, Dip-8, 3.3 mm H x 6.35 mm W x 9.27 mm L (Pack of 10)	$8.50
OCR 24Value 500pcs Electrolytic Capacitor Assortment Box Kit	$15.99
Total	$109.45

Our grand total will be $ 9,600 (labor cost) + $109.45 = $9709.45

3.2 Schedule

Week	Goal
3/1-3/7	Finalize computer vision software design; working on image capturing and inference
3/8-3/14	Interface audio storage device with microcontroller & sound signal processing circuit testing
3/15-3/21	Interface transmission system with microcontroller and prototype all electronics on breadboard & power circuit and protection circuit design and testing
3/22-3/28	PCB design and testing
3/29-4/4	Working on synchronization problem of distributed nodes
4/5-4/12	Power circuit testing and protection circuit testing

4/13-4/19	Final product review and testing
4/20-4/25	Mock Demo

4. Discussion of Ethics and Safety

4.1 Ethics

We understand the importance of ethical and safety concern during the process of designing this product and put “safety, health, and welfare of the public” from #1 of the IEEE Code of Ethics at the forefront of our design thinking [1]. Since this product contains a camera module which will capture image input of users at real time, it is important to protect the privacy of end users. We plan to provide users the full disclosures about how we handle the user inputs and give users the ability to choose if captured images should be stored temporarily or erased immediately. Based on the Consumer Data Privacy and Security Act, we would “directly obtain the individual’s consent” before we start to collect user information [2]. We will also maintain a security program to maintain “security, confidentiality, and integrity of personal data” from malicious usages [2].

4.2 Electric Hazards

The audio system needs to process the signal from its receiver module and output the signal into the speaker. There is the possibility for a potential electric hazard caused by a short circuit. The audio system is powered by an atmega328P chip which will source a 5V voltage. The max transmission operating current is 115mA and max receiving operating current is 45mA for a typical RF communication module [3]. An electric current as low as 30 mA current could potentially induce a ventricular fibrillation [4]. Thus special circuitry has been implemented to cut down the power immediately once a short circuit has been detected.

4.3 Fire Hazards

The RF communication module and its peripheral signal processing units could be affected by surging of current due to a short circuit. The RF communication module has a total power dissipation of 60mW [3]. When power is not cut off promptly after a short circuit, the temperature of the PCB board will potentially induce a fire. Thus, we have taken the power regulation of the whole module into consideration during the PCB design of our product.

References

[1] “IEEE Code of Ethics.” *IEEE*, www.ieee.org/about/corporate/governance/p7-8.html.

[2] Moran, Jerry. “Text - S.3456 - 116th Congress (2019-2020): Consumer Data Privacy and Security Act of 2020.” *Congress.gov*, 12 Mar. 2020, www.congress.gov/bill/116th-congress/senate-bill/3456/text#toc-ida8f663638e07477a8a47c52bb9e5f876.

[3] “Single Chip 2.4 GHz Transceiver.” *Sparkfun*, www.sparkfun.com/datasheets/Components/nRF24L01_prelim_prod_spec_1_2.pdf.

[4] “Electrical Injury.” *Wikipedia*, Wikimedia Foundation, 15 Jan. 2021, en.wikipedia.org/wiki/Electrical_injury.

[5] Martin, Taylor. “34 Alexa Tips and Tricks for Beginners.” *CNET*, www.cnet.com/how-to/amazon-echo-alexa-tips/.

[6] *All-New Echo (4th Gen) | With Premium Sound, Smart Home Hub, and Alexa | Charcoal*. www.amazon.com/All-New-Echo-4th-Gen/dp/B07XKF5RM3/ref=asc_df_B07XKF5RM3.

[7] “Buy HomePod.” *Apple*, www.apple.com/shop/buy-homepod/homepod?afid=p238%7CsE8EkhnRj-dc_mtid_1870765e38482_pcrid_487811458778_pgrid_54140720810_&cid=aos-us-kwgo---slid-SUuEsuxd--product-.

[8] “Full-Range 8’ Speaker Pioneer Type B20FU20-51FW.” *Parts Express*, www.parts-express.com/GRS-8FR-8-Full-Range-8-Speaker-Pioneer-Type-B20FU20-51FW-292-430?gclid=Cj0KCQiAvbiBBhD-ARIsAGM48bxiUtTOtZXyLCI_CmjwfI6am72oGwLDXQFNWUhIsYm-aY3hDzTdWGAaAsEpEALw_wcB