GESTURE CONTROLLED AUDIO SYSTEM

By

Kehang Chang

Ruofan Chen

Ruohua Li

Final Report for ECE 445, Senior Design, Spring 2021

TA: Anthony Schroeder

May 2021

Project No. 37

Abstract

The Gesture Controlled Audio System aims to provide users with an easy-to-use yet effective gesture-based way to control multiple speakers simultaneously. In this report, we demonstrate that with accurate vision processing, wireless communication, and audio processing systems, a well-designed and integrated system can be implemented to accomplish our goal.

Contents

1. Introduction

1.1 Motivation

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.

1.2 Solution Overview

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.

Our solution is designed 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.

2 Design

2.1 Design Overview

Our design is composed of three subsystems. 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. The modularity will be demonstrated by the physical diagram and block diagram as follows.

2.2 Physical Diagram

The following is a diagram shows the physical layout and data flow of the whole system.

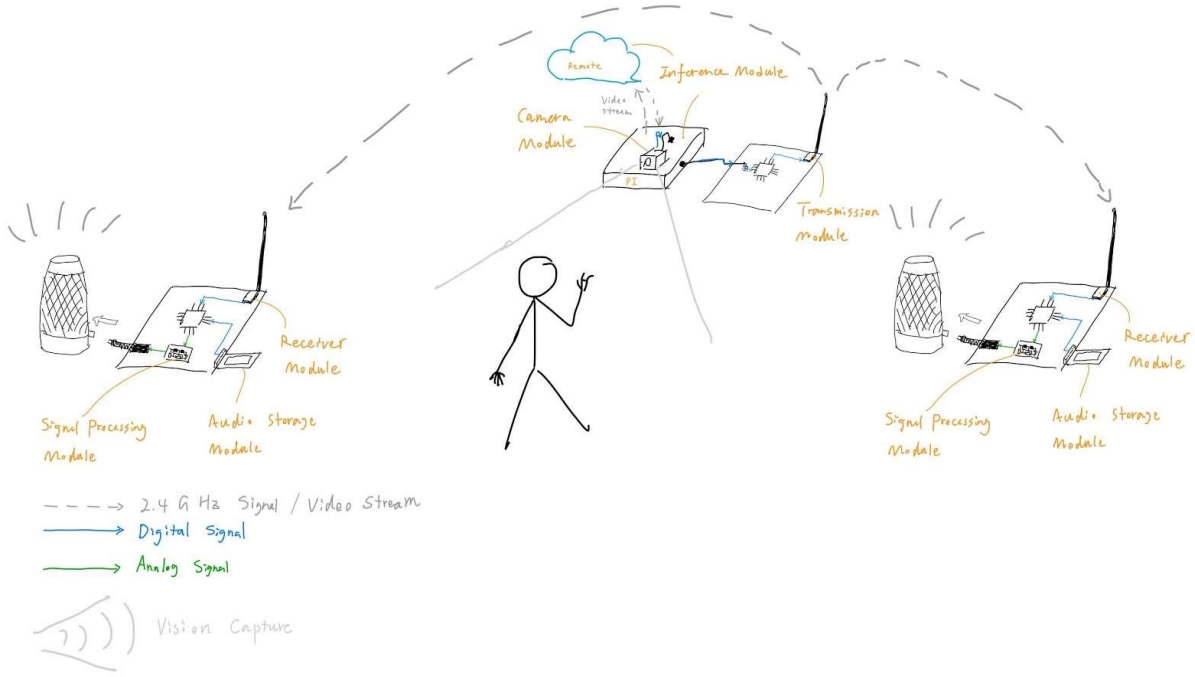

Figure 1 Physical Diagram

2.3 Block Diagram

The following is a diagram shows the system's functionality modularly and the path of signal generation, transmission, and output is demostrated.

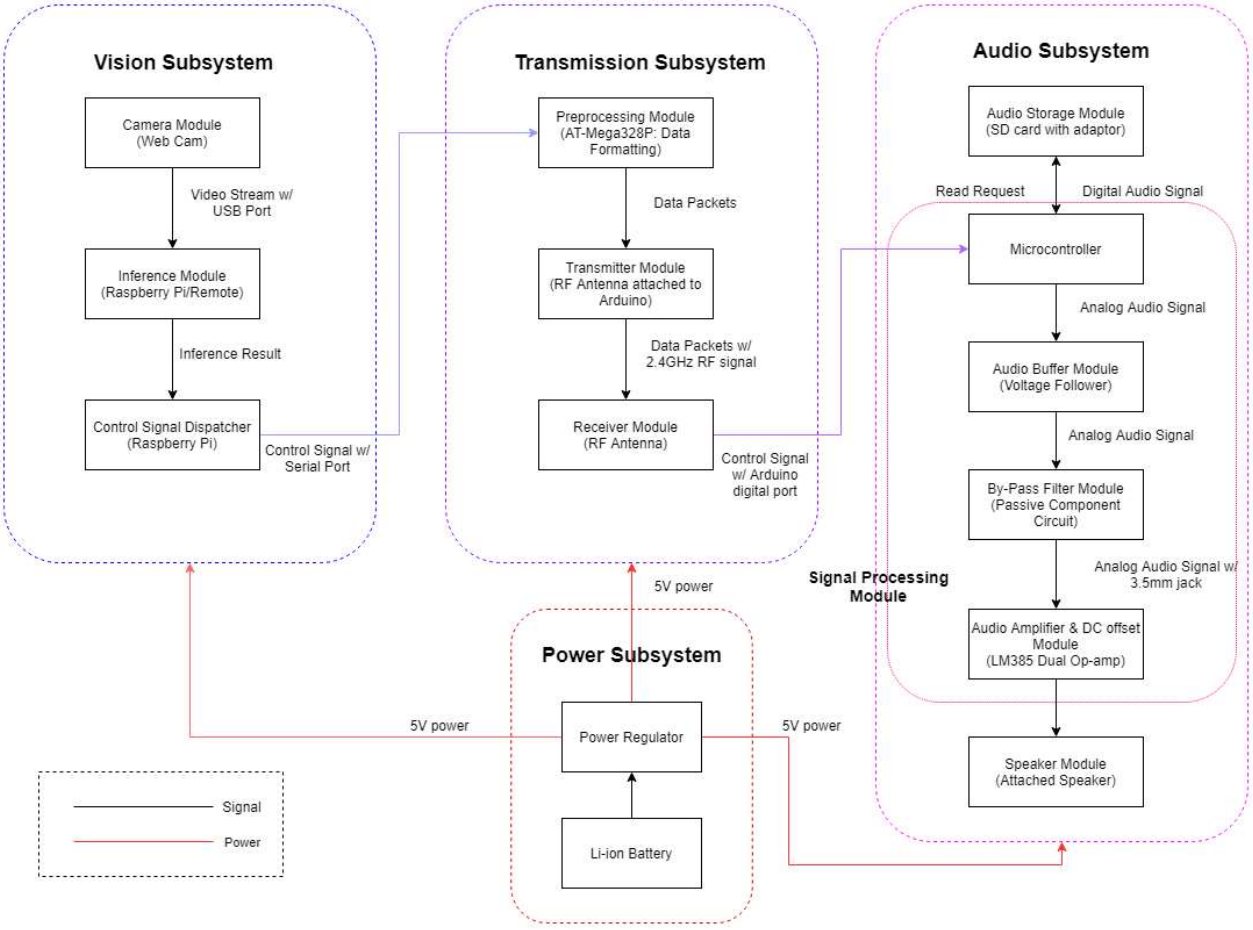

Figure 2 Block Diagram

2.3 Subsystem Descriptions

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.

2.3.1.a Camera Module

The camera module will capture a video stream at 25fps 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).

Figure 3 The ELP megapixel Super Mini 720p USB Camera web camera used in our design.

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.

Figure 4 RaspberryPi 3b+ board used in our design.

We adopted keypoint-based inference paradigm in our design, that is, we first use a machine learning model to inference, based on a frame, keypoints on hand. Then with this information, we use an algorithm to further inference the specific gesture the hand is posing based on each keypoint's relative position.

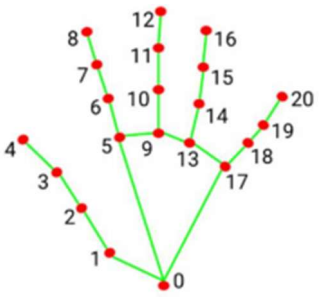

Figure 5 Keypoints our model tries to locate.

Algorithm 1: Keypoints2Gesture

```
Result: G
Input: K
Initialize F := empty list;
for finger in fingers do
    if K[finger][tip].y < K[finger][midjoint].y then
        F[finger] := True
    end
    else
        F[finger] := False
    end
end
if F[okay fingers] = True and F[not okay fingers] = False then
    G := Okay
end
else if F[pistol fingers] = True and F[not pistol fingers] = False
 then
    G := Pistol
end
else if F[fist fingers] = True and F[not fist fingers] = False then
    G := Fist
end
else
    G[finger] := None
end
return G;
```

Figure 6 Pseudocode of the algorithm we use to determine hand gestures based on keypoints' locations.

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.

Figure 7 ATMega328p-pu microcontroller used in our design, which is the standard chip used in Arduino Uno.

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.

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.

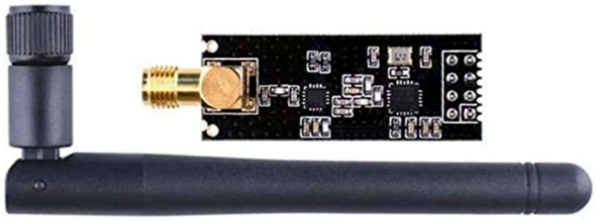

Figure 8 NRF24L01+PA+LNA RF Transceiver used in our design which support 2.4G Hz bluetooth signal two-way communication.

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.

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 amplification and discharge signal through the speaker.

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.

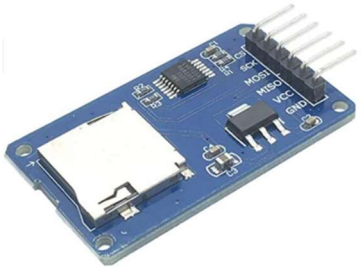

Figure 9 The SD card adaptor used in our design.

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.

2.3.3.c Signal Processing Module

The signal processing module consists of an audio amplifier, and a DC offset circuit. If an audio signal is used to drive the speaker and then it will distort the signal. 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.

Figure 10 The LM-386N op-amp used in our design, which performs 20 times amplification by default.

2.3.3.d Speaker Module

This module is a speaker output signal generated by signal processing module.

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 6V~14V and output a steady 5V voltage for any module on the PCB board.

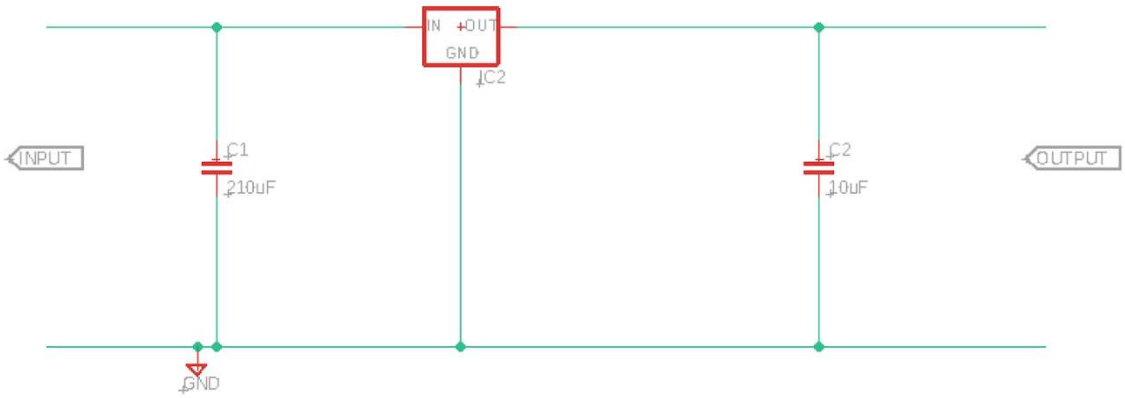

Figure 11 Power regulator diagram schematics for our design.

2.3.4.b Li-ion Battery

This module is a battery case with 4x1.5V batteries that supply power to all modules.

Figure 12 The 4x1.5V AA battery case used in our design.

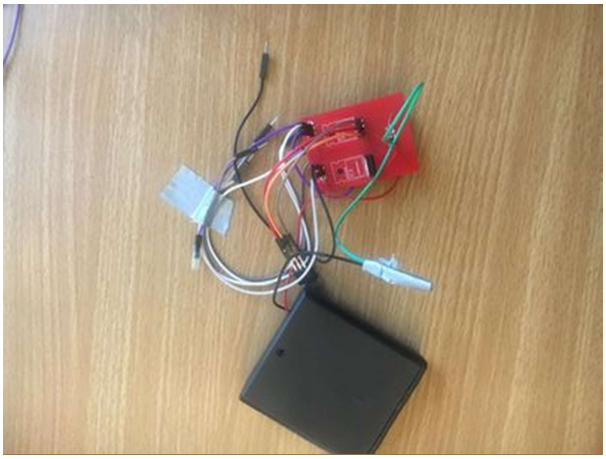

Figure 13 Actual soldering with the battery.

2.3.5 Integration Scheme

In reality, just like what depicted in the physical diagram, vision subsystem and transmitter are integrated together while audio subsystem and receiver are integrated together mechanically.

We drafted our PCB design to make PCB board of transmission subsystem and audio subsystem to include a serial port that make it possible for us to check, in real time, the data being processed in each subsystem. In the PCB board of power subsystem, there are both 3V and 5V outlet to make the usage more convenient for other subsystems with different power consumption.

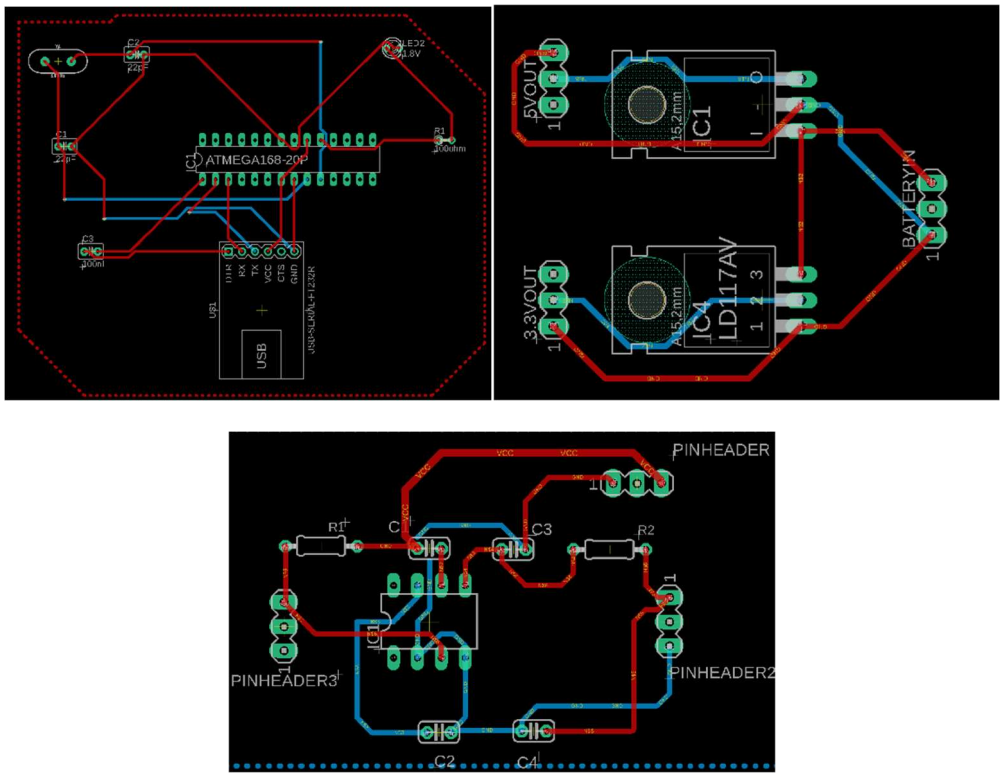

Figure 14 The PCB designs of our subsystems, including transmission subsystem (left), power subsystem (right), and audio subsystem (bottom).

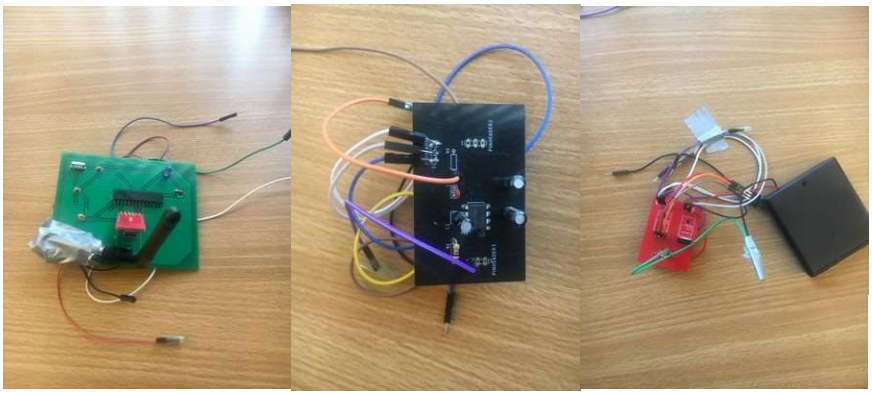

Figure 15 The soldered PCB boards, including transmission subsystem (left), power subsystem (right), and audio subsystem (middle).

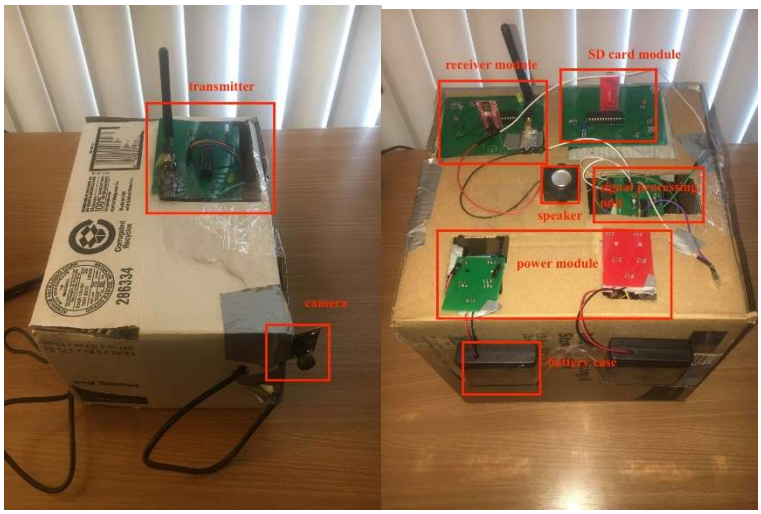

Figure 16 The actual integration of our subsystems, including transmitter side (left) and receiver side (right).

3. Design Verification

During the course of implementing the system, we unit tested modules individually to test if our design was viable in reality.

3.1 Vision Subsystem

We designed our vision subsystem to capture human hand gestures visually in real time, so we tested the accuracy of the model we adopted to accomplish this [1].

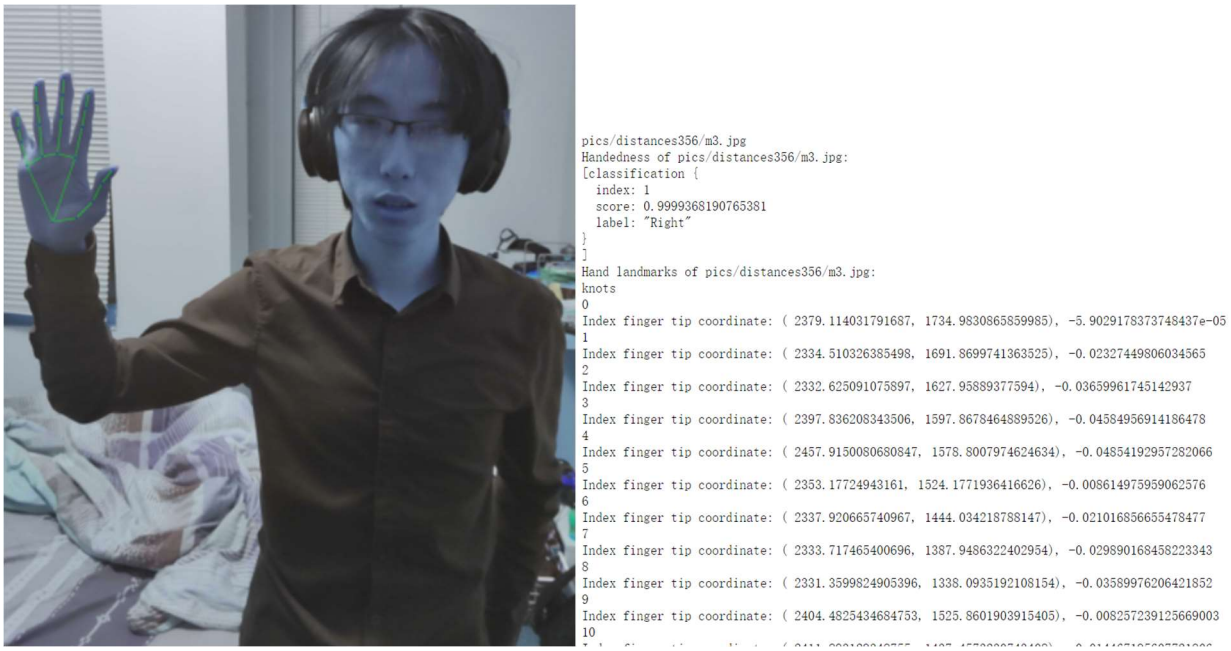

Figure 17 An example test for the accuracy of hand gesture recognition. Visualization on the left and keypoints' locations on the right.

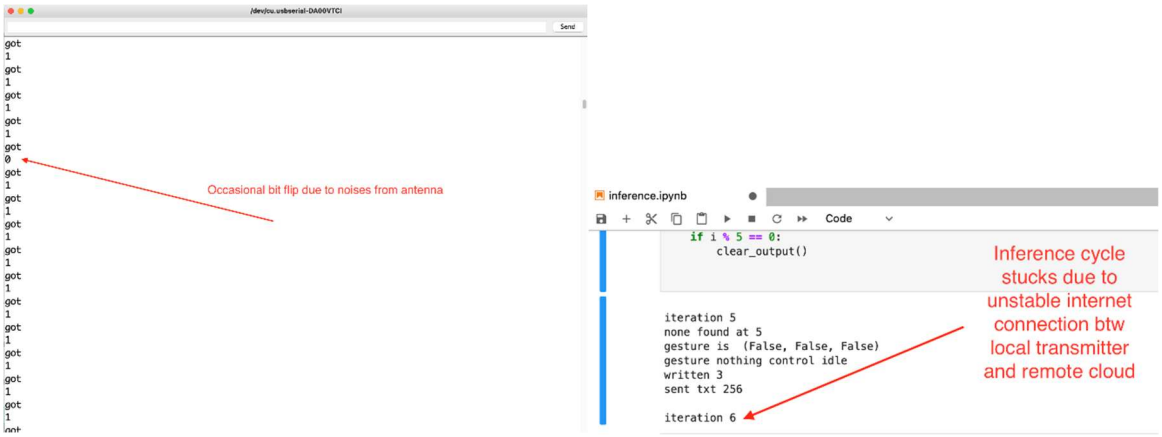

Figure 18 Examples of trying to debug software based on the result of inference and transmission.

3.2 Transmission Subsystem

The transmission system is designed to transmit and receive control signal generated by the inference module, so we thoroughly test the transmitting and receiving accuracy to see if there is any information distortion or loss in the process of data transmission.

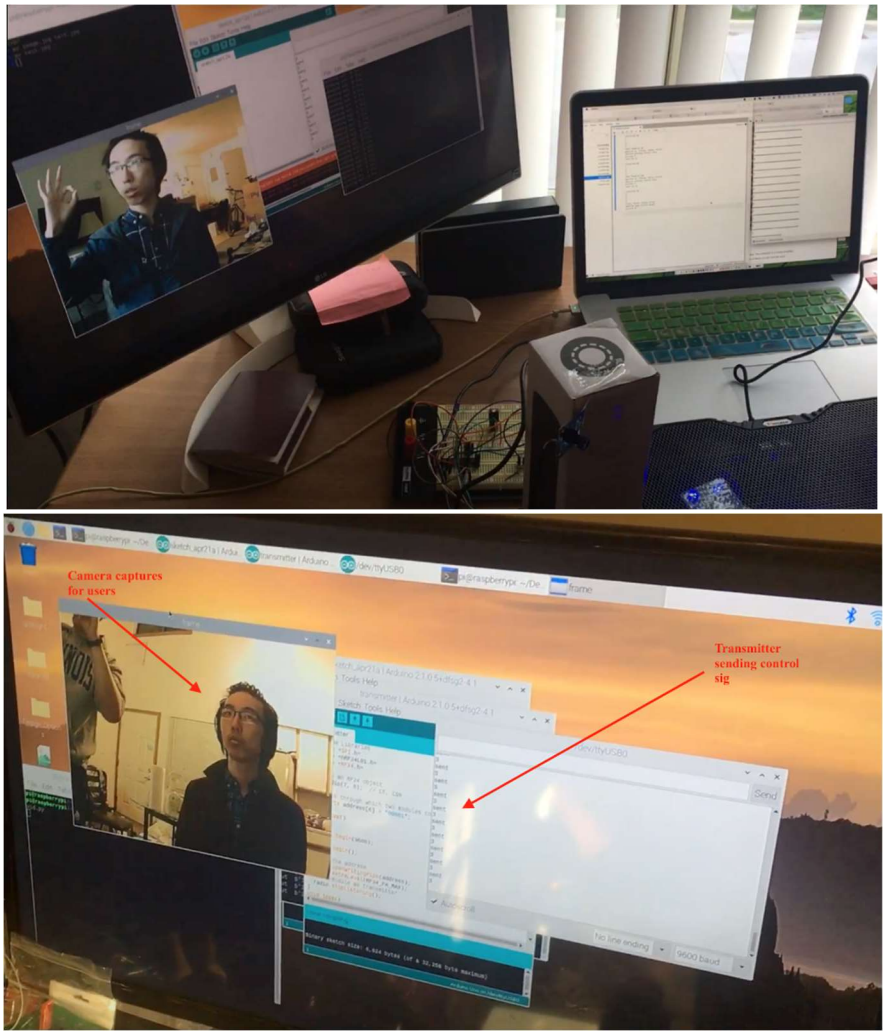

Figure 19 Examples test to verify if the correct signal is sent and received by printing signal flow on both sides while a valid gesture is posed.

4. Costs

4.1 Parts

Table 1 shows the costs for each individual part.

Table 1 Parts Costs

Part	Manufacturer	Retail Cost ($)	Bulk Purchase Cost ($)	Actual Cost ($)
ATMega328p-pu microcontroller	Microchip	$20.39	$20.39	$20.39
HiLetgo 5pcs Micro SD TF Card Adater Reader Module	HiLetgo	$6.99	$6.99	$6.99
SanDisk 32GB Ultra SDHC UHS-I Memory Card	SanDisk	$8.99	$8.99	$8.99
ELP megapixel Super Mini 720p USB Camera Module	ELP	$29.9	$29.9	$29.9
E-Projects B-0004-H15 Ceramic Disc Capacitor, 50V, 0.01uF, 103 (Pack of 25)	E-Projects	$5.69	$5.69	$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)	National Semiconductor	$8.50	$8.50	$8.50
OCR 24Value 500pcs Electrolytic Capacitor Assortment Box Kit	OCR	$15.99	$15.99	$15.99
Total		$109.45	$109.45	$109.45

4.2 Labor

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

5. Conclusion

5.1 Accomplishments

After the design, implementation, and testing of the whole system, we have accomplished 1) coding a keypoint-based hand gesture recognition system that is capable of capturing and classifying hand gestures in real time, 2) assembling a series of circuits that is able to read in data through serial port and communicate via RF signal, and 3) designing a series of circuits that manage audio signal storage and amplification.

5.2 Uncertainties

The main uncertainty we identified during testing and demonstration is that the performance of the system relies upon high-speed Intranet connection. Since the remote inference server and local video capturing system is connected via local SCP protocol, low-speed connection may cause the system to have undesirable lagging. Also, sometimes the positions of antenna may cause data packets fail to be transmitted through RF signal.

5.3 Ethical considerations

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].

5.4 Future work

Many future improvements are possible based on the current implementation. Firstly, a more control gestures can be supported such as volume-changing gestures and previous song gesture. Secondly, the audio signal storage can be centralized to the main node instead of being assign to each receiver, making updating the song list more convenient. Thirdly, a better user interface can be developed to make initiating the system easier.

References

[1] MediaPipe documentation, web page. Available at: https://google.github.io/mediapipe/solutions/hands.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.