Gesture Controlled Audio System

Design of a novel audio distributing & control system

By

Kehang Chang (kehangc2)
Ruofan Chen (ruofanc2)
Ruohua Li (ruohual2)

1. Introduction

1.1 Objective

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's also very common when you are trying to connect more speakers to enhance the listening experience, but you just don't have the right types of smart speakers to pair several speakers together. And 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 hasn't been an existing product in the market right now which would offer the convenience of both features.

1.1.2 Proposed Solution

Gesture Controlled Audio System is an audio sharing and coordination system. This system is aiming for providing 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.

1.2 Background

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 needs to be woken 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 too. The newest released version of Alexa: All-new Echo (4th Gen) costs 80 dollars [6]. And 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.

1.3 Physical Design

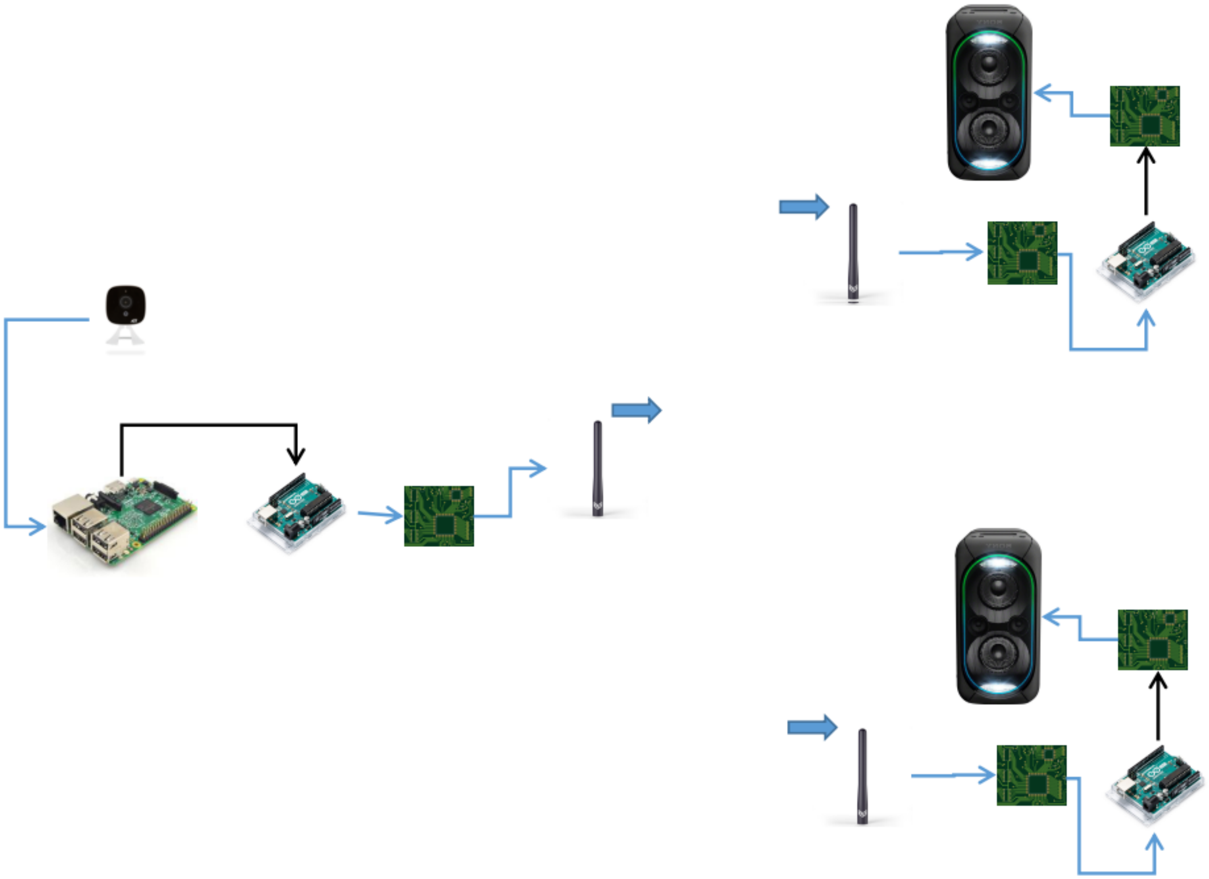

Fig1. Physical Design

1.4 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, previous track. We might let users create their own gesture profiles as well.
- **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 processing pipeline within 1 second: detect gesture, decode gesture, execute audio control.

2. Design

2.1 Block Diagram

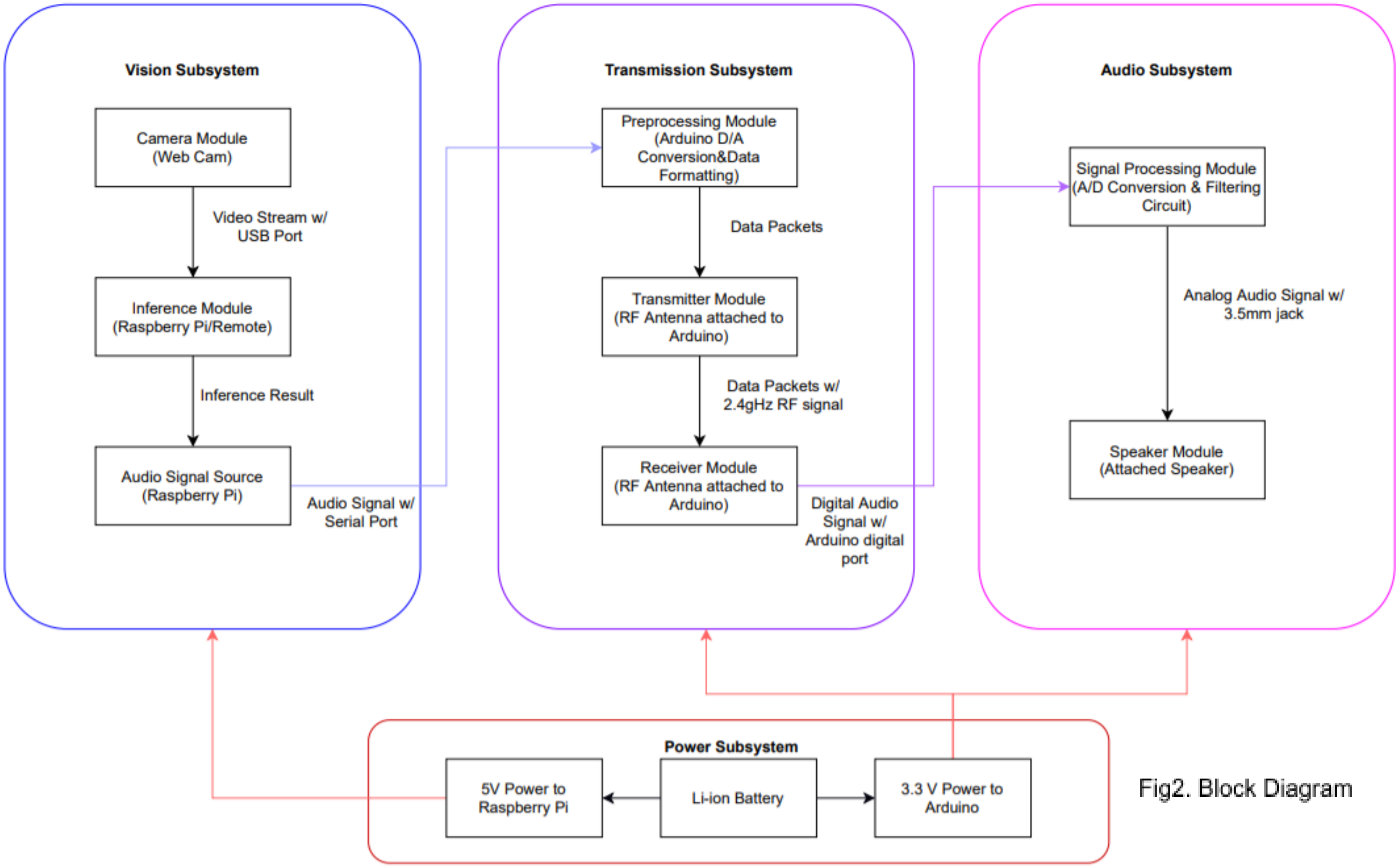
Vision Subsystem
Camera Module (Web Cam)
Video Stream w/ USB Port
Inference Module (Raspberry Pi/Remote)
Inference Result
Audio Signal Source (Raspberry Pi)
Audio Signal w/ Serial Port
Transmission Subsystem
Preprocessing Module (Arduino D/A Conversion&Data Formatting)
Data Packets
Transmitter Module (RF Antenna attached to Arduino)
Data Packets w/ 2.4gHz RF signal
Receiver Module (RF Antenna attached to Arduino)
Digital Audio Signal w/ Arduino digital port
Audio Subsystem
Signal Processing Module (A/D Conversion & Filtering Circuit)
Analog Audio Signal w/ 3.5mm jack
Speaker Module (Attached Speaker)
Power Subsystem
5V Power to Raspberry Pi
Li-ion Battery
3.3 V Power to Arduino

Fig2. Block Diagram

The vision subsystem tries 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 audio signal determined by the vision subsystem and transmit it to audio subsystems. Audio subsystems will then process the signals transmitted and output them with the attached speaker.

2.2 Functional Overview

- **Vision Subsystem**
  Our vision subsystem will try to 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 audio signals sent to the transmission subsystem. The signal will be sent by serial port (usb transmission) or 3.5mm audio jack.
- **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.
- **Audio Subsystem**
  Our audio subsystem, upon receiving audio signals, will process the data packets sent from the transmission subsystem to digital audio signals. 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.
- **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.

2.3 Block Requirements

- **Vision Subsystem**
  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.

- **Audio Subsystem**
  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.
- **Transmission Subsystem**
  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.
- **Power Subsystem**
  The voltage source is provided by a 4-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.

2.4 Risk Analysis

There are some problems that pose risk to the completion of our project. Firstly, the processing speed of the embedded system we are using , i.e. Raspberry Pi and AtMega, may be not enough to carry our proposed works given their computation-intensive nature. Secondly the RF module may experience unwanted noise like static or other 2.4GHz signals that will impair the integrity of the signal.

External cloud computing power might be used if Raspberry Pi or Atmega328P couldn't offer enough computing power to complete the required tasks in a certain time range(<= 1s).

Additional filters might be added to restore the RF signal caused by unwanted noises.

3. Ethics and Safety

3.1 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's important to protect the privacy of end users. We are planning 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].

3.2 Electric Hazards

The audio system needs to process the signal from its receiver module and output the signal into the speaker. There could be 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.

3.2 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.