An AI based Voice Controlled Humanoid Robot

Bisma Naeem (✉ mscs19018@itu.edu.pk)
Information Technology University https://orcid.org/0000-0002-7302-384X

Dr. Saeed-Ul-Hassan
Manchester Metropolitan University

Nadeem Yousuf
Information Technology University

Research Article

Keywords: Humanoid Robot, Artificial Intelligence, Speech Emotion Detection

DOI: https://doi.org/10.21203/rs.3.rs-2424215/v1

An AI based Voice Controlled Humanoid Robot

Bisma Naeem*, Nadeem Yousaf[1], Saeed-Ul-Hassan[2]

*[1]*Department of Computer Science, ITU, Lahore,* [1]*Department of Computer Science, ITU, Lahore,*
[2]*Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, United Kingdom*

**mscs19018@itu.edu.pk*

[1]*nadeem.yousaf@itu.edu.pk*

[2]*saeed-ul-hassan@itu.edu.pk*

Abstract

Humanoid robots may be utilized in daily life and are more efficient at performing tasks that humans would find unpleasant. Robots are getting more proficient and are capable of performing many tasks that humans can. In a world designed for people, creating robots that behave like humans is a significant problem for robotics. In this study, we introduce a Voice Controlled Humanoid Robot, a mobile robot that can be moved by the operator by issuing precise voice instructions. The Google Voice API is used to handle the voice command when it is picked up by an Android phone's microphone. The vocal signals are then translated into text by the app, which creates a variable against the text and sends it to the Arduino Node MCU in the form of a command. The Arduino Node MCU then examines the instruction and performs the necessary operations. The VCHR app and VCHR system are linked together using the Bluetooth module. The android app also has a camera for live video streaming, and the robot can utilize its SONAR sensors to identify any obstacles in its path and sound an alert as a result. VCHR can carry out around 20 distinct tasks in total. When given voice input through the supplied external mic, the VCHR system is also capable of speech-emotion recognition in addition to these characteristics. The IoT cloud service provider ThingSpeak receives the temperature sensor data from the VCHR system in order to analyze and interpret the sensor data at various time intervals. The performance obtained for movement, speech emotion recognition, and sensor data processing is demonstrated by experimental findings.

Keywords: Humanoid Robot; Artificial Intelligence; Speech Emotion Detection

1. Introduction

Robots are being used in several nations as the virus spreads around the globe. Some robots can relieve worn-out nurses in hospitals, while others can assist in warehouses. They can even perform simple cleaning tasks and deliveries. Manufacturing organizations can use industrial robots to maintain some production while their human coworkers are under quarantine [1,2]. Social robots can coexist with people and establish bonds that enhance their quality of life. The success of these interactions depends on people's perceptions of the capabilities of robots as a result of artificial intelligence, computational models, or robot exemplifications. This Extraordinary Issue compiles unique pledges showcasing technically sound, logical, and philosophical accomplishments in social robotics and artificial intelligence, or innovative concepts, revelations, and developments in social robotics [1,3]. In This paper, we propose a methodology addressing control movement, Context awareness, and emotion detection of Voice Controlled Humanoid Robot.

- Controls for Movement: The actions of the voice-controlled humanoid robot, which uses a total of 17 motors to aid in its operation and carry out the activities as directed, mimic those of a person due to the usage of servo motors on each joint. Various tasks are performed by the robot such as walking, dancing, playing football, warm-up, entertaining, hurdles detection, and live controlled camera view which is commanded by an android app. The robot can be controlled either by the buttons of the app or by voice recognition [4,5,6].

- Speech Emotion Detection: The speech recognition function independently forms the robot's intelligence. The voice recognition feature of the robot is achieved by connecting it to the google speech-to-text converter API that converts the speech-to-text commands and will pass it to the robot to perform the actions the robot provides a source of entertainment for the viewers and the operators. The majority of market-available robots lack a camera view and voice control, so you can only operate them with a gaming remote. The difficulty is in developing a robot that can recognize speech, perceive emotions, and then use neural networks to activate green and red LED lights. In order to accomplish the voice-controlled functionality in this example, we must write it, create an android application, and link it with this application. It's uncommon to find a robot that can distinguish between speech pitches and be given labels to identify different moods. The Voice Controlled Humanoid Robot (VCHR) technology for speech emotion detection lags when the robot is linked to the cloud for voice command communication between the user and the robot. The major goal was to create a humanoid robot that could feel and understand emotions, as well as transfer sensor data to the cloud for visualization. Although they make up one of the smallest subsets of service robots available now, humanoid robots have the most potential to replace other industrial tools in the near future. Robots that resemble people have been developed by a number of firms for use as teaching and medical aids. At the moment, humanoid robots—especially companion robots—are thriving in the medical field. The humanoid robot project aims to amuse people by carrying out behaviors and duties that will draw viewers' attention to itself. All of these motors have been controlled with a servo controller. The Bluetooth wireless board is also connected to the camera that is mounted on the robot head. This board transmits signals to the mobile device, which allows you to use an Android mobile application to see exactly what the robot can see. Within this software, there is a microphone button that you only need to press in order to command the robot. In this instance, we have also created an application for the Arduino board.

- Control for Context Awareness: The robot is getting additions like a microphone module integrated into the VCHR System and a temperature sensor to measure nearby objects' temperatures. The temperature sensor data from the VCHR System is transmitted via the cloud, ThingSpeak, and is displayed using Power BI. We used EdgeImpulse, a third-party supplier, to train and test the neural network model for speech emotion recognition using the provided speech dataset. The trained model enabled EdgeImpulse to create an Arduino library, which is now being used on the VCHR System. To eliminate the latency caused by cloud connections, a speech emotion recognition library was created utilizing EdgeImpulse rather than sending VCHR voice instructions directly to the cloud.

Section 2 includes a comprehensive review of the literature that covers the full development of a humanoid robot, the incorporation and improvement of Speech Emotion Recognition with effective speech processing, and includes studies of human-robot interaction that have a significant influence on the creation of helpful and sociable robots. Section 3 describes our methodology and the steps we took to create a voice-controlled humanoid robot that could recognize spoken emotions. We have described the stages involved in developing both software and hardware in this section. Results of VCHR are discussed in Section 4 along with successful results for speech emotion recognition and visualizations of sensor data. In Section 5, we came to a conclusion and offered suggestions for more VCHR-related research projects.

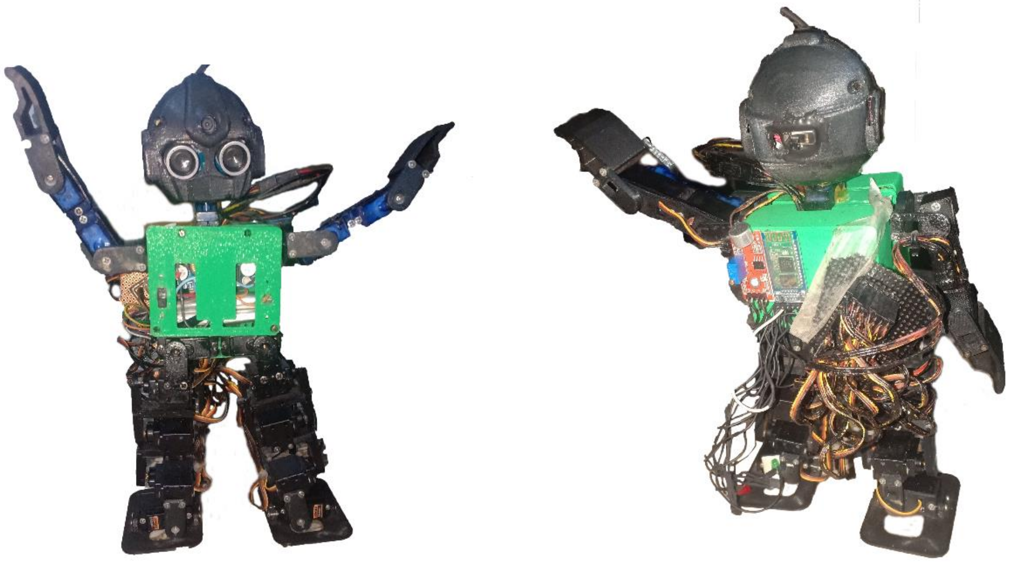


Figure 1 Voice-Controlled Humanoid Robot

2. Related Work

In this section we provide an introduction to the state-of-the-art movement of humanoid robots Section 2.1, Then we discussed the existing research on Speech Emotion recognition in Section 2.2. A brief review of the context awareness control is discussed in Section 2.3.

2.1. A brief review of the movements of *Humanoid Robots*

Robots that are social are made to live alongside people and interact with them in ways that improve everyone's quality of life. AI and computer models are essential to these interactions' success. Martín et al. [1] describe the development of social robotics and enlist its applications and advancements. Another recent study by Jeong et al. [2] proposed a method for interactive performance art, a method of interaction analysis that makes it simple to understand the intention of robots. Yuan et al. [3] Identify the participants' common worries, such as the robot's price, its robotic voice, and the lower acceptability of the robot by people with Alzheimer's disease-related dementia due to cognitive deficit. Jung et al. [4] development methodology used the Bluetooth connection for communication. The android application's voice instructions are used to control the robot, and the data is decoded before being sent to a different microcontroller to drive the DC motor. D'Mello et al. [5] research show how Human-robot interaction must be conducted in a natural way. They emphasize on precise and accurate commands should be delivered by the user for effective results. They have explained the methodology of how a system processes the commands and apply rules to match commands and control the systems.

A brief review on Speech Emotion Recognition

Robots are capable of identifying emotional cues in human voice signals for cordial interactions with people, leading to ultimately satisfying performance and effect. We followed the approach of Breazeal et al. [6] who studied robots' behavior to achieve better results, researchers looked into how robots responded to human speech and focused on various speech habits and patterns. They used a process in which human speech with many variants is used to test the robot's emotions once they have been integrated into it. Another study by Katzenmaier et al. [7] shows the efficiency of auditory and visual cues, as well as how they work together to identify the addressee in a human-human-robot interaction. Based on eighteen audiovisual records, they can tell one human from a robot when they are interacting with each other. This study compares the outcomes of three techniques. The first approach just uses auditory cues to identify the addressees. Investigated are both low-level, feature-based signals and higher-level cues. Their research shows that visual assessment of head posture during human-human-robot contact is a more trustworthy indicator for determining the addressee. James et al. [8] surveyed Human-Computer interaction by integrating emotion recognition in the robotic systems. Explained its research trends and different methods for its implementation. Zatarain-Cabada et al. [9] show that by observing and analyzing the learner's emotional response, the instructor can alter the lesson with the intention of improving the quality of the learning process in the traditional educational context. Intelligent Tutoring Systems (ITS) should consider the links between emotion, cognition, and action to improve interactions with students in a computer environment. As a result, ITS with emotional detection capabilities can provide useful tools for learning that is both effective and pleasurable. These tools are driven by behavioral and cognitive paradigms. To enable ITS to take into consideration and regulate the emotions of the learner, they created an architecture with a number of components for emotion recognition. Rifinski et al. [10] demonstrated a Web-based solution for teaching fundamental math. The web-based system consists of various parts, including an emotion recognizer, an intelligent tutoring system, and a social network for learning. The system was created with the intention of being accessed from any type of computer platform and a mobile device running the Android operating system. He developed a fuzzy system for tracking students' educational states and a neural system for identifying student emotions. They conducted numerous testing with the emotion recognizer and had a 96 percent success rate. Becker et al. [11] have investigated why emotions should be incorporated into conversational agents of acceptance regarding the useful outcomes of technological advancement in daily life, leading to the development of knowledgeable embodied conversation agents rather than unembodied expert systems. They have started to create systems with the abilities required to support social learning, such as discovering models based on facial identification and skin tone, merging saliency through a context-sensitive attention system, creating expressive displays, and managing social interactions. They contend that the challenges associated with integrating social learning systems into a robot push us to address issues that are relevant to biological systems but are not currently being researched.

A brief review of Contextual Awareness

Exploring the issues of developmental structure, physical embodiment, integration of multiple sensory and motor systems, and social interaction Brooks et al. [14], constructed an upper-torso humanoid robot called Cog. The robot has twenty-one degrees of freedom and a variety of sensory systems, including visual, auditory, vestibular, kinesthetic, and tactile. They have reported on a variety of implemented visual-motor routines, orientation behaviors, motor control techniques, and social behaviors. They further outlined a number of areas for future research that will be necessary to build a complete embodied system. Breazeal and Cynthia. [24] explore the expression of emotion in synthesized speech for an anthropomorphic robot. They have adopted several key emotional correlates of human speech to the robot’s speech synthesizer to allow the robot to speak in either an angry, calm, disgusted, fearful, happy, sad, or surprised manner. Breazeal and Cyntia. [25] proposed that the newly emerging robotics applications for domestic or entertainment purposes are slowly introducing autonomous robots into society at large. A critical capability of such robots is their ability to interact with humans, and in particular, untrained users.

3. System Design and Methodology:

A methodical approach is followed in the creation of humanoid robots as well as the justification of all robotic construction activities. The robotics implementation is interdisciplinary as a result of the input from different professional organizations and information from numerous fields. The major goal is to develop mathematical models with an emphasis on humanoid robots. The voice-controlled humanoid's hardware assembly and software-based components are both involved. It uses 17 servo motors to move in various directions and SONAR sensors to detect collisions. For data processing, visualization, and modeling in a manner similar to how people balance themselves while walking, the sensor data is uploaded to the cloud.

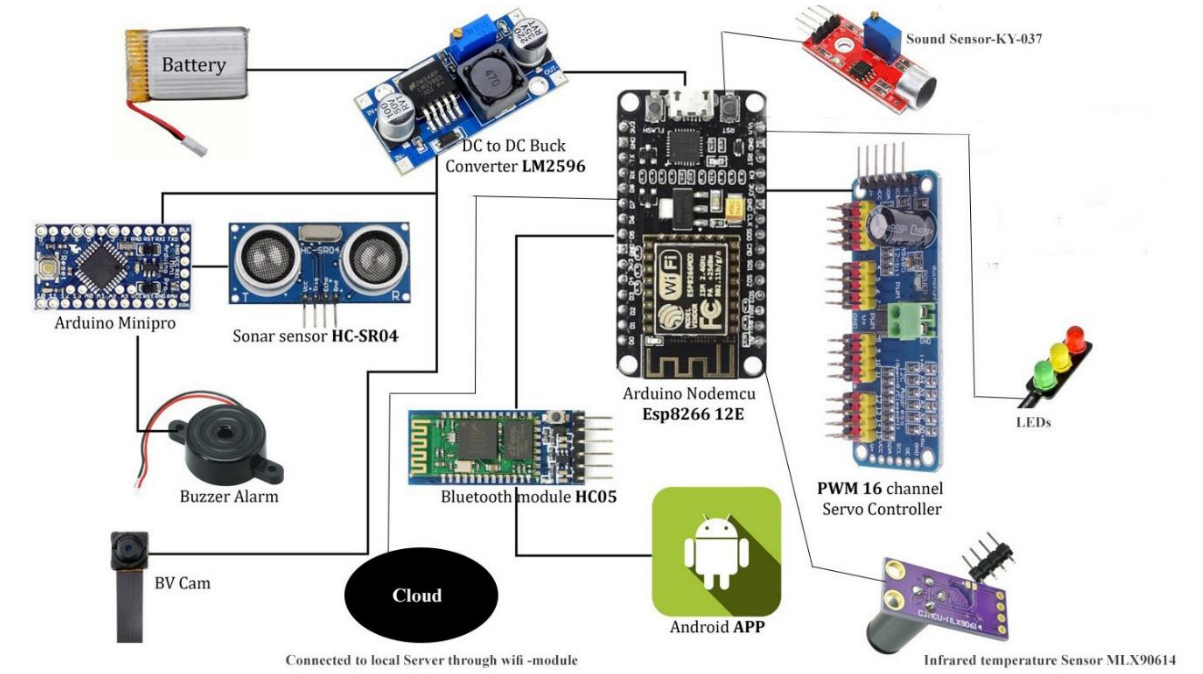
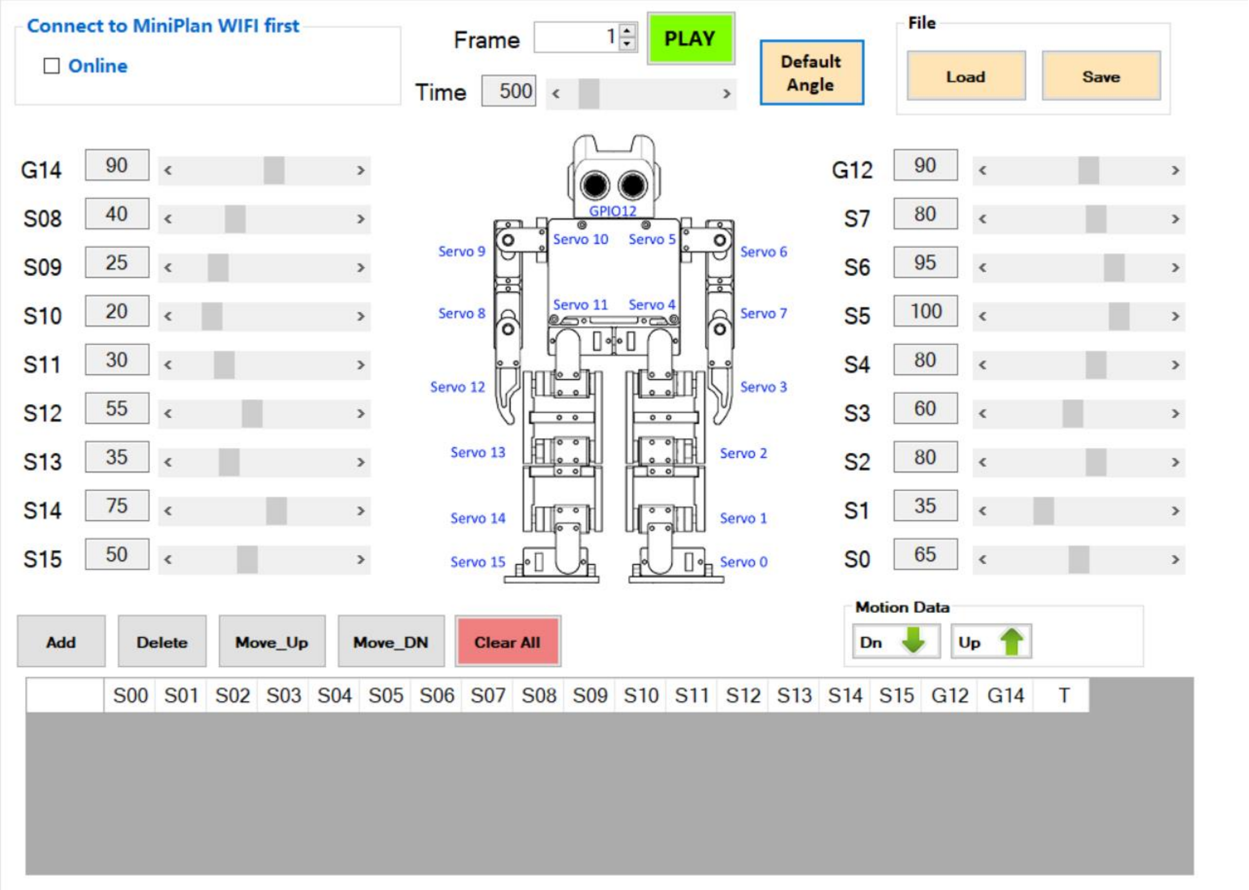

Battery
DC to DC Buck Converter LM2596
Sound Sensor-KY-037
Arduino Minipro
Sonar sensor HC-SR04
Arduino Nodemcu Esp8266 12E
PWM 16 channel Servo Controller
LEDs
Buzzer Alarm
Bluetooth module HC05
BV Cam
Cloud
Android APP
Connected to local Server through wifi -module
Infrared temperature Sensor MLX90614

Figure 2 VCHR System Architecture

3.1.Controls for Movement

The project Voice Controlled Humanoid Robot consist of a robot that has features and functions that are like humans. The movement features are being implemented to show the movement in all four directions forward, backward, left, and right directions. Through its voice recognition feature, the robot takes the input from the user’s voice and performs the function. The Android App which is developed to control the robot by simple button-pressed inputs and the voice inputs are made user-friendly and interactive. The app is based on Arduino Node MCU which is the main part that controls 17 servo motors of which each one is directed towards Arduino Node MCU, and the other motors are indirectly connected to the board. Fig 5 shows a miniplan designed interface for robots manipulation in different directions.



Frame 1

Time 500

G14	90
S08	40
S09	25
S10	20
S11	30
S12	55
S13	35
S14	75
S15	50

G12	90
S7	80
S6	95
S5	100
S4	80
S3	60
S2	80
S1	35
S0	65

	S00	S01	S02	S03	S04	S05	S06	S07	S08	S09	S10	S11	S12	S13	S14	S15	G12	G14	T

Figure 3 Miniplan View for Robot Motor Angle Configuration

3.2.Controls for Context Awareness

We have used a servo controller to control all these motors as well as there's a camera on the robot head and this camera is connected to the Bluetooth wireless board this board sends the signals to the mobile then you can see the view of what exactly the robot can see in your android mobile application as well as there will be a microphone button within this app you just need to click a microphone button and then you can give the instructions to the robot. In this case, we have programmed the Arduino board as well as an application too. The node MCU esp8266 has built-in Wi-Fi and can connect to the cloud with it. By using the cloud, ThingSpeak online utility, we have saved the temperature logs and Ultrasonic SONAR sensor readings from sensors used and exported it. For the Speech Emotion Recognition, we have used a third-party provider

EdgeImpulse, which made a speech emotion recognition Arduino library and is deployed on Arduino

3.3.Controls for Emotion Detection

The other functionality added is the voice emotion-sensing through a user-generated library by using Edge Impulse which is used to generate an Arduino library to deploy on the board. Breazeal et al. (2002), the board is also connected to the cloud through which the temperature sensor data is recorded.

The Temperature sensor used in the VCHR enables it to measure the temperature of its surrounding objects as it is necessary to measure the body temperature in the days of the Covid-19 pandemic. The Ultrasonic SONAR sensor used, work to avoid the obstacles in its range by triggering the alarm. The camera fitted in the center of the head works for its vision, inspired by human eyes to show the live streaming view and the ability to respond to human speech i.e., sensing and detecting emotions made VCHR unique and humanoid.

Algorithm 1 VCHR Algorithm

```
Input: V : Voice Commands
Input: T: Voice pitch
Output: Mf: Movement
Output: Ee: Emotion Detection response
ZeroCondition():
Mf ← GETS NO INPUT AND IS INITIALIZED TO DEFAULT ANGLES
ForwardCondition():
 Mf ← GETS INITIALIZED TO 16 FORWARD MOVEMENT STEPS
 BackwardCondition():
 Mf ← GETS INITIALIZED TO 10 BACKWARD MOVEMENT STEPS
 TurnLeftCondition():
 Mf ← GETS INITIALIZED TO 27 TURN MOVEMENT STEPS
 TurnRightCondition():
 Mf ← GETS INITIALIZED TO 27 RIGHT MOVEMENT STEPS
 m← MICROPHONE INFERENCE RECORD
```

```
for each $V \in T$ do
    if m has some pitch then continue
        if T.classification[2]0 > 0.7 then ledpin, HIGH
            if T.classification[2] < 0.7 then LEDpin, LOW
```

Speech is one of the most natural ways for humans to express themselves. The detection and analysis of the same region are crucial in today's digital world of remote communication since emotions are so important in communication. Due to the subjectivity of emotions, emotion identification is a difficult task. On how to evaluate or classify them, there is no widespread agreement. To model them after conventional CNN, features gleaned from audio clip extraction are presented in a matrix manner. When time signals are analysed traditionally, any periodic component, such as echoes, is seen as strong peaks in the accompanying frequency spectrum, also known as the Fourier spectrum. This is what happens when a Fourier transform is applied to the time signal. A spectrogram can be used to obtain any cepstrum feature by using the Fourier Transform. The unique feature of MFCC is that it is measured using the Mel scale, which connects a tone's perceived frequency to its actual measured frequency. It scales the frequency to more precisely match the range of human hearing. The vocal tract is represented by the envelope of the temporal power spectrum of the speech signal, and MFCC accurately depicts this envelope. We used Google Speech. The commands dataset used the first 50 samples with different inflections of the pitch. Edge impulse automatically split the data between training and test sets, and it also let it infer the label from the filename. Any periodic component, such as echoes, is detected in the traditional analysis of time signals as sharp peaks in the associated frequency spectrum, or Fourier spectrum. Applying a Fourier transform to the time signal yields this result. Used CNN is used to classify MFCC blocks (Mel Frequency Cepstral Coefficients) and then the trained model is deployed as a C++ library to use as an Arduino library.

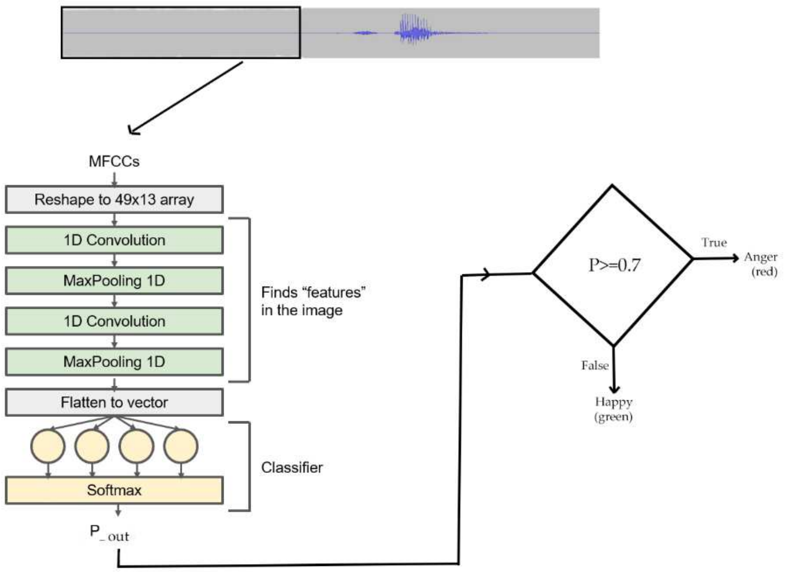


Figure 4 Data Flow Diagram for Speech Emotion Recognition

4. Results and Discussion

The Result section is based on the experimental results performed on the VCHR. We divide it into sub-sections according to the features performed by the robot.

4.1. Movement Control

Robots' movement is performed by 17 DOF servo motors that are fitted on different joints of the robot, with the help of these servo motors the robot performed different movements i.e., forward, backward, left, and right directions. A legged robot's activity is greatly controlled by the interaction of balance as well as contact restraints. In this scenario, there are more means to disperse contact that requires a larger series of stable configurations. Nevertheless, each contact imposes added closed-loop kinematic restraints, which limit the robot's series of activities. Angles for each movement of the motor is specified using miniplan that is used to adjust the servo motor angles for the movement. Here the figure below shows the movement control of different motors by using mini plan. Figure 5 shows motor alignment and configuration by using Miniplan.

A configuration q of the robot can be represented in a 6+N dimensional configuration area Q, with 3 levels of flexibility each for the translation as well as turning off a picked origin as well as N levels of flexibility for the joints. The following graph shows the motion of the robot in the forward direction with the steps taken by different motors for the accomplishment of the result. Figure 6 shows the motor configurations at different angles to accomplish Forward movement. The x-axis shows the number of steps and the y-axis shows angle configurations

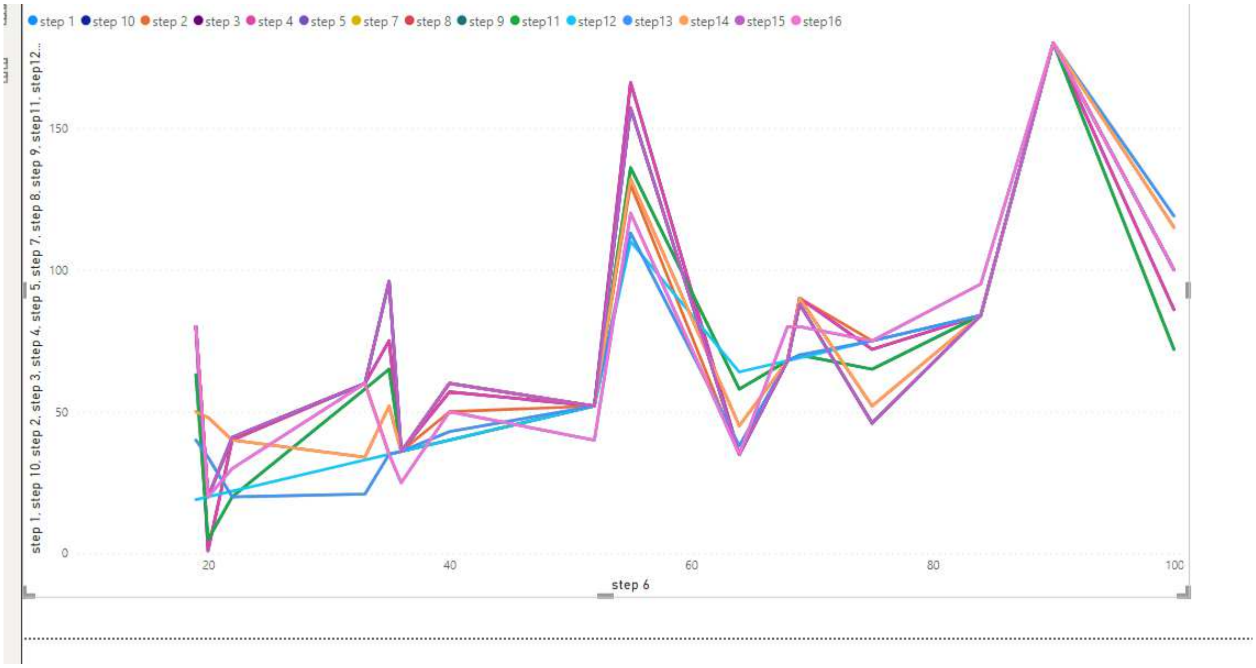


Figure 5 Motors Configurations for Forward Movement

The movement of the robot is accomplished according to the given user commands by the app made in the MIT app inventor. Voice commands are given and processed by the google speech tool and converted into text and then conveyed to the VCHR system. The app interface is made user-friendly so that the user can control the robot through his voice and through the buttons of the app that are made for robot movement control. A Bluetooth connection is required to connect the robot with the mobile app and control it.

4.2.Sensor Results

We export the IR temperature sensor and the SONAR sensor data values by connecting it with a third-party cloud provider ThingSpeak that sends data to the cloud where it is stored in a private channel. Figure 8 shows the IR Temperature sensor readings and Figure 9 shows the SONAR sensor readings used in collision detection.

4.2.1. Infrared Temperature Sensor

It is Used To measure the surrounding temperature and to detect the temperature of an item without any touch. Body thermometers, fire alarms, water heaters, and room warmers are just a few examples of real-world applications that employ an infrared temperature sensor as user assistance or for autonomous monitoring. We have used an IR sensor to monitor body temperature. Figure 8 shows the temperature measurement in Celsius °C at different time intervals. We build the x-axis and y-axis lists programmatically. Each second, the temperature was recorded from the MLX90614 IR temperature sensor and appended to the y-axis list. The x-axis lists the time stamp for each reading of the temperature sensor. The x-axis and y-axis lists are then used to create a plot with ax.plot(x-axis, y-axis). The red line in the mid of the graph shows the average recorded temperature which is 37.7°C.

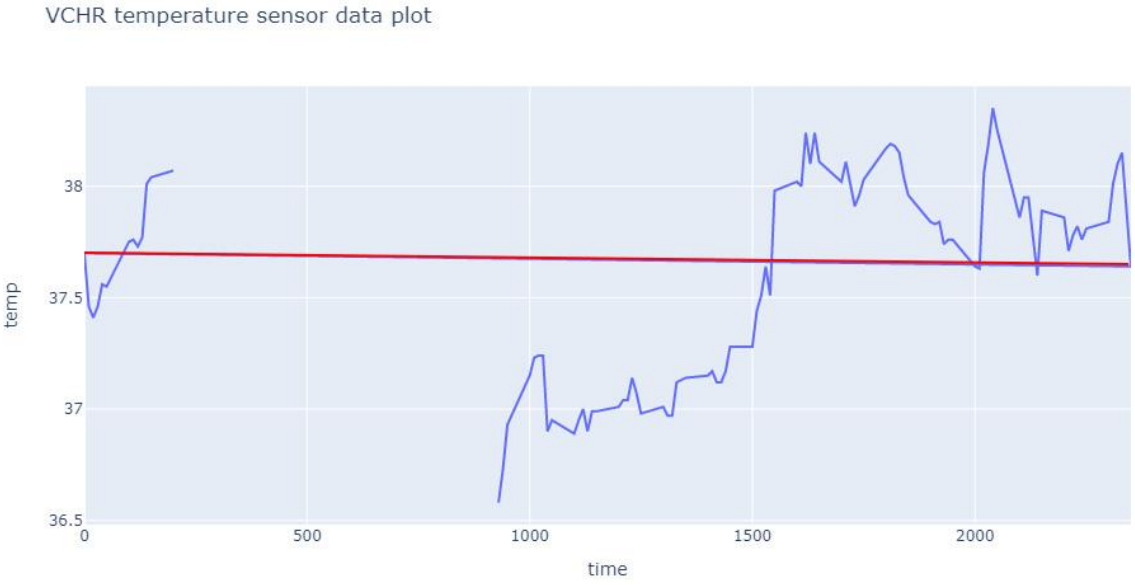


Figure 6 Temperature Sensor Graph VCHR

4.2.2. *Ultrasonic SONAR Sensor*

It is used to estimate distance and operates on the principle of reflected sound waves. The ultrasonic sensor emits sound waves, which are reflected if an item is in front of it. The sensor detects these waves and calculates the time it takes for them to be transmitted and received. We used short-range ultrasonic SONAR sensors for collision detection and obstacle avoidance. The graph below shows the readings of the SONAR sensor tested in an indoor and outdoor environment. Figure 9 shows Ultrasonic SONAR sensor graph. The x-axis lists the number of readings recorded by the ultrasonic SONAR sensor and the y-axis lists the range recorded in meters (m) by the SONAR sensor.

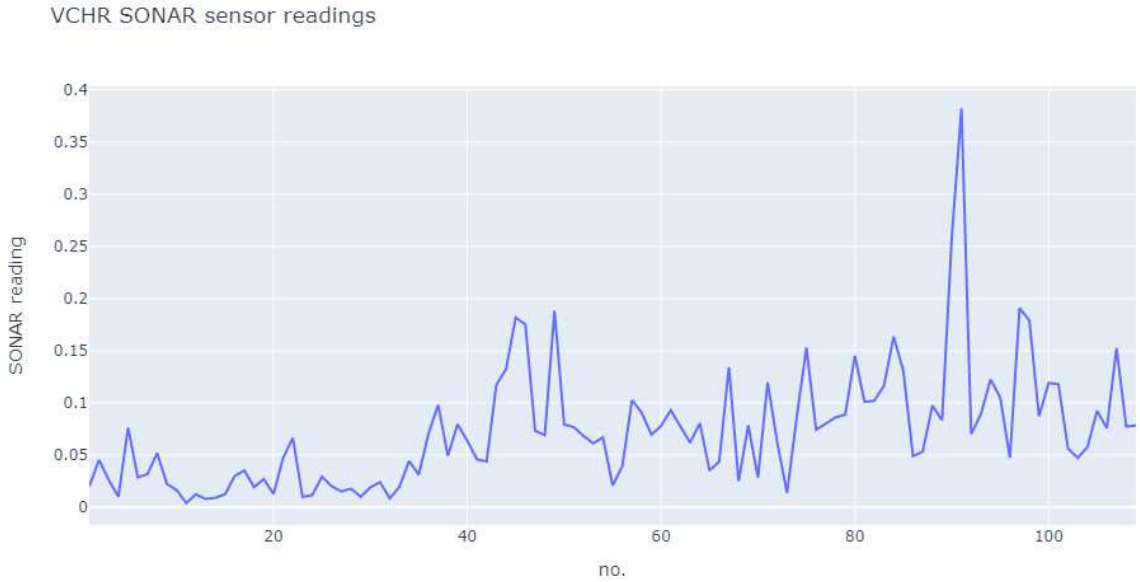


Figure 7 Ultrasonic SONAR Sensor Graph

4.3.Speech Emotion Recognition

The model is not able to choose a single median duration at which to clip all audio files and pad any shorter files with zeroes to maintain dimensions because the audio clips in our datasets range in length from under 2 seconds to over 6 seconds. Additionally, we've included a few unique terms. Our data consists of 50 samples with various pitches and inflections. To match the microphone on our VCHR system, the audio files are resampled to 16 kHz. developed a pipeline for these samples, produced Mel Frequency Cepstral Coefficients (MFCC) for the audio samples, and sent the audio file data to Edge Impulse. The most important thing to understand about communication is that the vocal system, which includes the tongue, teeth, and other parts, filters the sounds that a person makes. This form specifies the audio that will play. If the form can be established accurately, this must provide us a clear picture of the voice that is being generated. The brief time power range envelope is where the vocal system takes shape, and MFCCs' job is to correctly represent this range. A pure tone's perceived frequency, or pitch, and its actual measured frequency are connected by the Mel scale range. Low frequencies are much easier for people to hear little changes in pitch than high frequencies. By including this spectrum, we can better align our operations with what individuals listen to. Although only the power spectrum envelope of a single frame is defined by the MFCC feature vector, it appears that speech would also contain actual information in the characteristics. In the end, calculating the MFCC trajectories and adding them to the initial attribute vector increases the efficiency of ASR (Automatic Speech Recognition) by a small amount. We utilized a common image classifier (CNN) to identify the keyword in each sample, after which we inserted a neural network block to train the model. For each sample, EdgeImpulse calculates the MFCC for training and testing. With a loss of 0.30 percent and a validation accuracy of 0.88 (88%), the model was implemented into an Arduino voice recognition library designed for the esp8266 board. Following the library's installation on the board, we tested it using different speech patterns and obtained the following findings. All speech input above 0.7 is deemed harsh and is shown by the VCHR as a red signal. An average range of pitch 0.7 is noted as the mid-point. According to Breazeal and Cyntia [6] research, which states that the speakers' autonomic nervous system causes the most significant changes in the arousal of different types of emotions, our experimental results heavily depend upon pitch variations of the speakers' effective state. For example, voice pitches greater than 0.7 show anger, and voice pitches lower than 0.7 show happiness and politeness. Results are presented in Section 4.2 for various voice pitch levels. When working with machine learning on such a resource, the processor is fully loaded performing these calculations for 274 milliseconds (or 82 percent of the time), which leaves little time for other tasks. We added some background noise (to help the model recognize the keyword in a variety of environments) and curated the data so that we end up with about 1500 samples in each of our categories. The classifier's output, where each number effectively represents the likelihood for each label, is useful. We conducted testing using specific keywords from the dataset. The keyword is seen in the last label shot at 0.98. Compared to polysyllabic words and phrases, short, monosyllabic words perform significantly worse. The anticipated model is not particularly accurate for this keyword because the model overfits the training data in the previous section. Because the micro-controller has a limited amount of resources and the CPU is mostly utilized to listen (record audio) and do inference. Always keep that in mind when designing the finished product. Code is included carefully to prevent the microcontroller from acting as a co-processor that just listens for keywords or overflowing the audio buffer. The co-processor might then transmit a message or toggle a GPIO line to alert other electronics.

4.3.1. Emotional Feedback for Human-Robot Interaction

The features are used by the deep learning model (CNN). The y-axis of the feature matrices generated will depend on the n mfcc or n mels chosen during data extraction. The x-axis is determined by the audio length and sample rate we choose during feature extraction. The audio clips in our datasets ranged in length from under 2 seconds to over 6 seconds, thus it would not be possible to select a single median duration at which to clip all audio files and pad any shorter files with zeroes to maintain dimensions. This is due to the fact that longer recordings would have lost information as a result, while shorter clips would have had only quiet for the second half of their audio length. In order to check this problem, we used different sampling rates in extraction in accordance with their audio lengths. In our approach, any, audio file greater or equal to 5 seconds was clipped at 5 seconds and sampled at 16000 Hz and the shorter clips were sampled such that the audio duration and sampling rate multiple remains 80000. A power, which is energy per unit of time, was found to be a more accurate metric to assure uniformity in our investigation of energy variation because the audio clips in our dataset were of varying lengths. Plots of this measure were made in relation to various emotions. It is clear that higher energy delivery is the main way people show their anger or fear. We also notice that, with certain exceptions, disgust and grief have an energy that is closer to neutral. Results show that the model conflated happiness with low-energy emotions. In the context of socially intelligent systems, the goal of this study was to provide the best classification of emotions present in a speech signal that could be used to feed the decision support system of a synthetic agent capable of supporting the societal participation of individuals lacking traditional communication channels. For all of the studies involving the classification of emotions, a CNN model was applied to the speech dataset. Our test results demonstrate that the robot expresses "happiness" at pitch intensities of 0.1, 0.25, and 0.5 and displays anger at 1.72, 2.1, 3.0, 1.5, and 2.62. Confusion is shown in our model at pitch levels 0.84 and 0.98. It is evident that polysyllabic words and phrases work better for evoking emotions. Results from experiments were accurate to within 68 percent. The best results are obtained when polysyllabic words are used and the voice is recorded without any background noise.

5. Conclusion

The Voice Controlled Humanoid Robot is constructed to resemble a human by using 17 DOF motors on each joint, which is in charge of the robot's movement just like the joints that control the movement of the human body. A Voice-controlled Humanoid Android App that is connected to the Google API controls the robot. It converts speech to text and sends commands to the robot, which response to the input by moving**.** The Robot also gives live streaming by the camera attached to its head and it can also detect the objects in its path with the help of the SONAR sensors. Additionally, the speech emotion detection Arduino library is used to introduce voice emotion detection, which detects emotions. Our experimental findings demonstrate how the VCHR reacts to various voice commands and how it interacts with human speech as an input, displaying emotional response through LED blinking. The robot was designed with kindergarten students' education, entertainment, and learning in mind during the COVID19 epidemic. Numerous countries have declared states of emergency because to the virus's ability to transmit through direct human contact, and the general public has been advised to keep a strict social distance. Robots are capable of carrying out human-like functions and can be financially built to replace some human interactions. In order to help control the spread of COVID-19, robotics, artificial intelligence, and human-robot interactions have grown more common in hospitals, airports, transit systems, recreation, scenic locations, hotels, restaurants, and communities as a whole. Humanoid robots, autonomous vehicles, drones, and other intelligent robots are used to lessen human contact and the potential spread of the SARS-CoV-2 virus. Examples include delivering materials, cleaning and sterilizing public spaces, detecting or measuring body temperature, providing safety or security, and comforting and amusing patients.

1. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

2. Author's Contribution

All persons who meet authorship criteria are listed as authors, and all authors certify that they have participated sufficiently in the work to take public responsibility for the content, including participation in the concept, design, analysis, writing, or revision of the manuscript.

3. CRediT authorship contribution statement

Bisma Naeem: Methodology, Software, Investigation, Data curation, Hardware- assembling, Design

Nadeem Yousaf: Conceptualization, Validation, Writing

Saeed-ul-Hassan: Supervision, Data curation, Writing – original draft, Writing – review & editing, Investigation, Validation

References

[1] Martín, F. A., Castillo, J. C., Malfáz, M., & Castro-González, Á. (2022). Applications and Trends in Social Robotics. *Electronics*, *11*(2), 212.

[2] Jeong, J., Yang, J., & Baltes, J. (2022). Robot magic show as testbed for humanoid robot interaction. *Entertainment Computing*, *40*, 100456.

[3] Yuan, F., Anderson, J. G., Wyatt, T. H., Lopez, R. P., Crane, M., Montgomery, A., & Zhao, X. (2022). Assessing the acceptability of a humanoid robot for alzheimer's disease and related dementia care using an online survey. *International Journal of Social Robotics*, 1-15.

[4] Yousefi-Koma, A., Maleki, B., Maleki, H., Amani, A., Bazrafshani, M. A., Keshavarz, H., ... & Ashtiani, M. S. (2021, July). SURENAIV: Towards A Cost-effective Full-size Humanoid Robot for Real-world Scenarios. In *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)* (pp. 142-148). IEEE.

[5] D'Mello, S., McCauley, L., & Markham, J. (2005, August). A mechanism for human-robot interaction through informal voice commands. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* (pp. 184-189). IEEE.

[6] Breazeal, C., & Aryananda, L. (2002). Recognition of affective communicative intent in robot-directed speech. *Autonomous robots*, *12*(1), 83-104.

[7] Katzenmaier, M., Stiefelhagen, R., & Schultz, T. (2004, October). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 144-151).

[8] Jaimes, A., & Sebe, N. (2007). Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, *108*(1-2), 116-134.

[9] Zatarain-Cabada, R., Barrón-Estrada, M. L., Alor-Hernández, G., & Reyes-García, C. A. (2014, November). Emotion recognition in intelligent tutoring systems for android-based mobile devices. In *Mexican International Conference on Artificial Intelligence* (pp. 494-504). Springer, Cham.

[10] Rifinski, D., Erel, H., Feiner, A., Hoffman, G., & Zuckerman, O. (2021). Human-human-robot interaction: robotic object's responsive gestures improve interpersonal evaluation in human interaction. *Human–Computer Interaction*, *36*(4), 333-359.

[11] Becker, C., Kopp, S., & Wachsmuth, I. (2007). Why emotions should be integrated into conversational agents. *Conversational informatics: an engineering approach*, 49-68.

[12] Peter, C., Beale, R., Crane, E., & Axelrod, L. (2007, September). Emotion in HCI. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference University of Lancaster, UK 21* (pp. 1-2).

[13] Breazeal, C., & Scassellati, B. (2002). f 4 challenges in building robots that imitate people. *Imitation in animals and artifacts*, *363*, 9.

[14] Brooks, R. A., Breazeal, C., Marjanović, M., Scassellati, B., & Williamson, M. M. (1998, April). The Cog project: Building a humanoid robot. In *International workshop on computation for metaphors, analogy, and agents* (pp. 52-87). Springer, Berlin, Heidelberg.

[15] Sasagawa, A., Fujimoto, K., Sakaino, S., & Tsuji, T. (2019). Imitation learning for human-robot cooperation using bilateral control. *arXiv preprint arXiv:1909.13018*.

[16] Wang, S., Braaksma, J., Babuska, R., & Hobbelen, D. (2006, July). Reinforcement learning control for biped robot walking on uneven surfaces. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings* (pp. 4173-4178). IEEE.

[17] Scassellati, B. (1998, April). Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In *International Workshop on Computation for Metaphors, Analogy, and Agents* (pp. 176-195). Springer, Berlin, Heidelberg.

[18] Konidaris, G., Kuindersma, S., Grupen, R., & Barto, A. (2012). Robot learning from demonstration by constructing skill trees. *The International Journal of Robotics Research*, *31*(3), 360-375.

[19] Slaney, M., & McRoberts, G. (1998, May). Baby ears: a recognition system for affective vocalizations. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)* (Vol. 2, pp. 985-988). IEEE.

[20] Fernald, A. (1989). Intonation and communicative intent in mothers' speech to infants: is the melody the message?. *Child development*, 1497-1510.

[21] Aryananda, L. (2002, September). Recognizing and remembering individuals: Online and unsupervised face recognition for humanoid robot. In *IEEE/RSJ International Conference on Intelligent Robots and Systems* (Vol. 2, pp. 1202-1207). IEEE.

[22] Fitzpatrick, P. (2002). *Role Transfer for Robot Tasking*. MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE.

[23] Varshavskaya, P. (2002). *Behavior-based early language development on a humanoid robot*. MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB.

[24] Varchavskaia, P. (2002). *Early pragmatic language development for an infant robot*. MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE.

[25] Breazeal, C. (2001, October). Emotive qualities in robot speech. In *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)* (Vol. 3, pp. 1388-1394). IEEE.

Figures

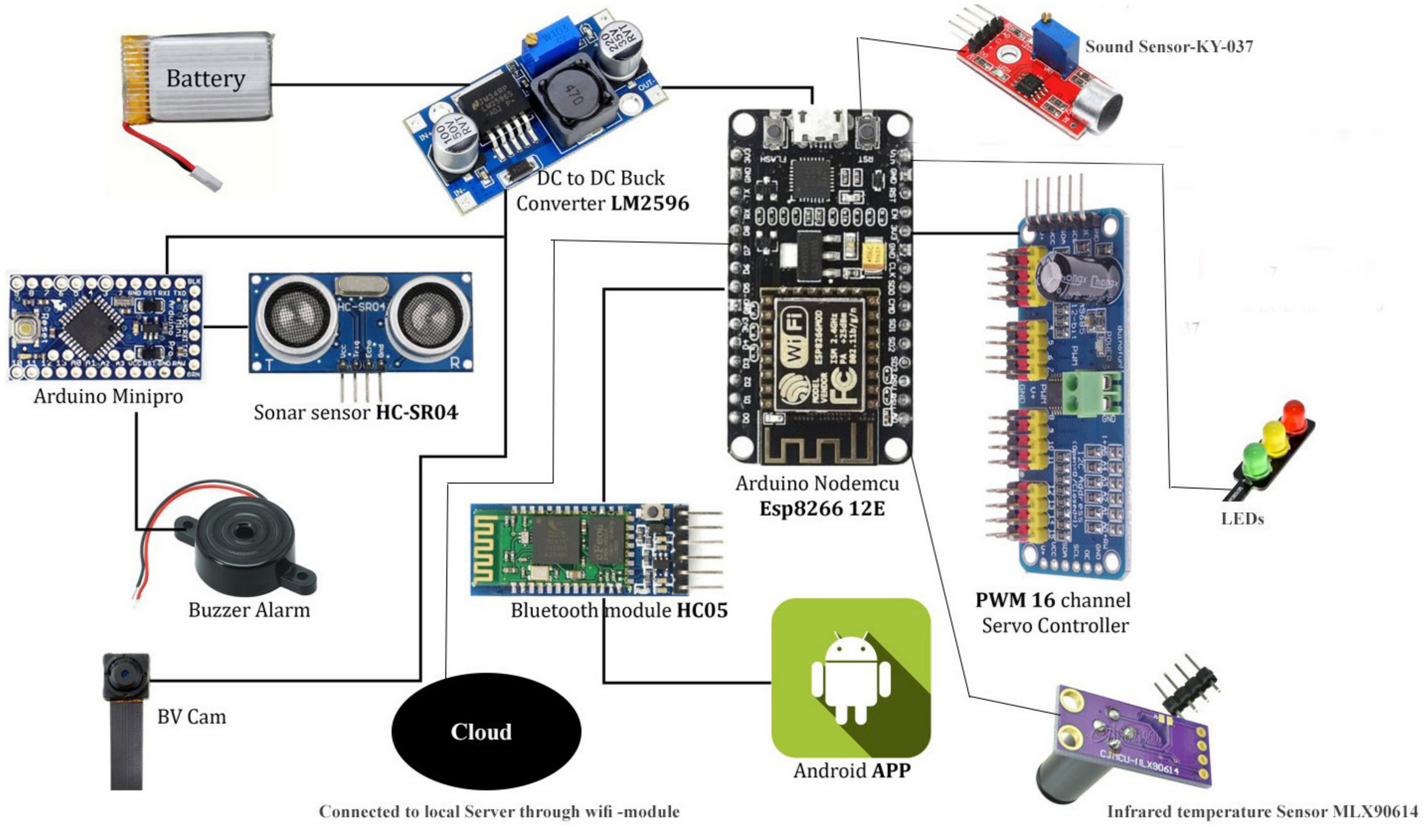

Battery
DC to DC Buck Converter LM2596
Sound Sensor-KY-037
Arduino Minipro
Sonar sensor HC-SR04
Arduino Nodemcu Esp8266 12E
PWM 16 channel Servo Controller
LEDs
Buzzer Alarm
Bluetooth module HC05
BV Cam
Cloud
Android APP
Connected to local Server through wifi -module
Infrared temperature Sensor MLX90614

Figure 1

System Architecture Diagram

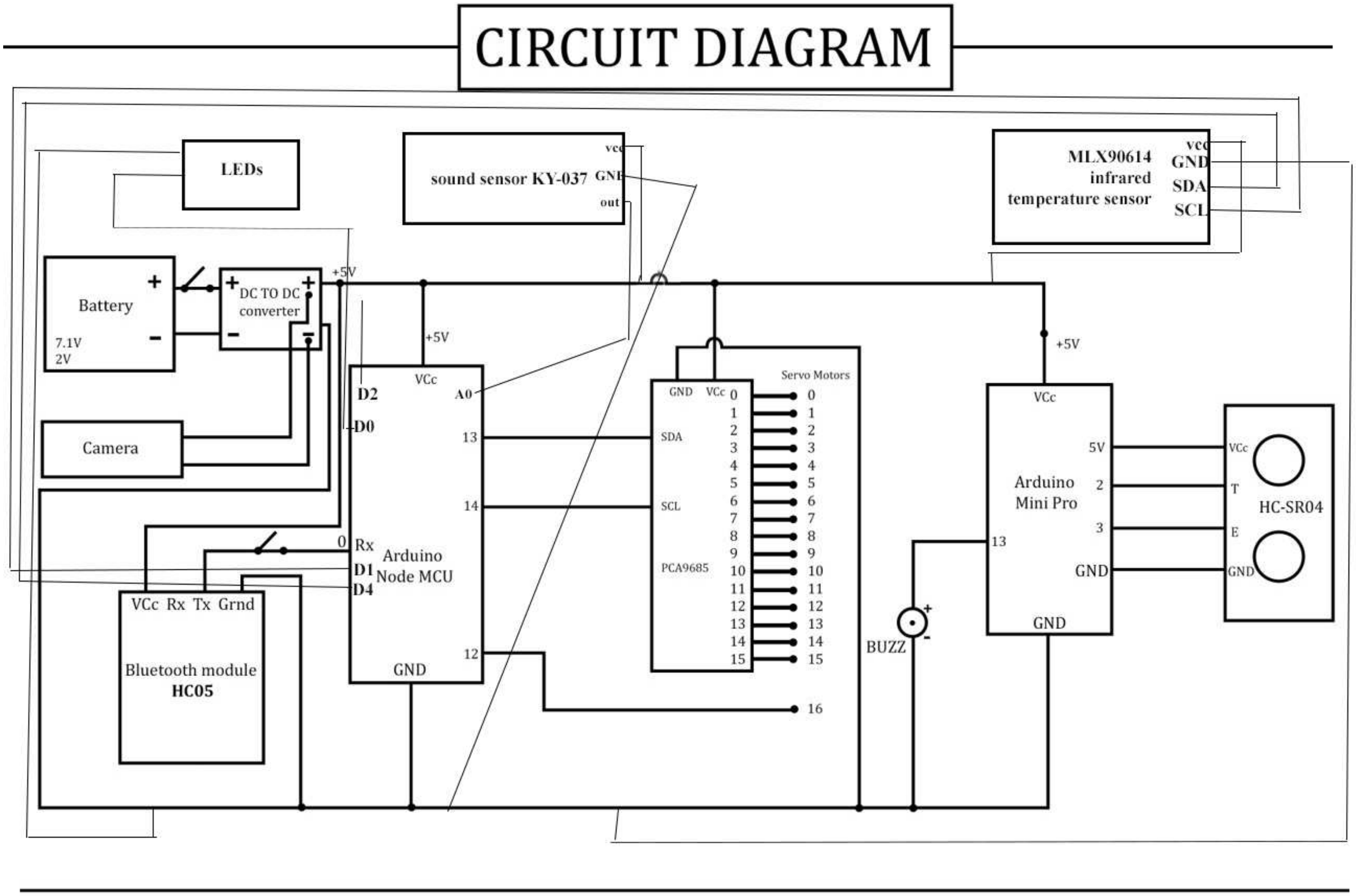


Figure 2

Circuit Diagram

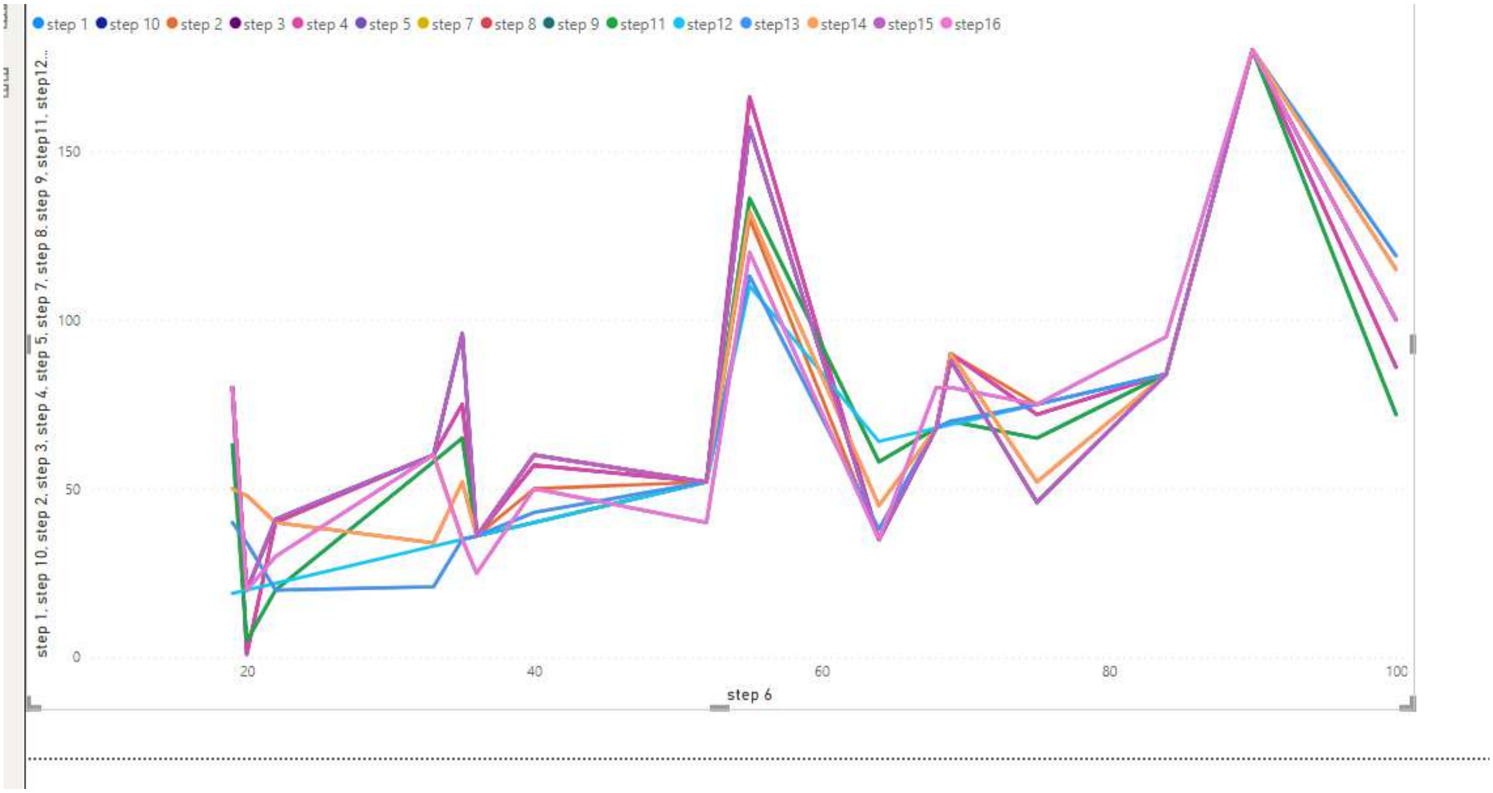


Figure 3

Forward Movement graph

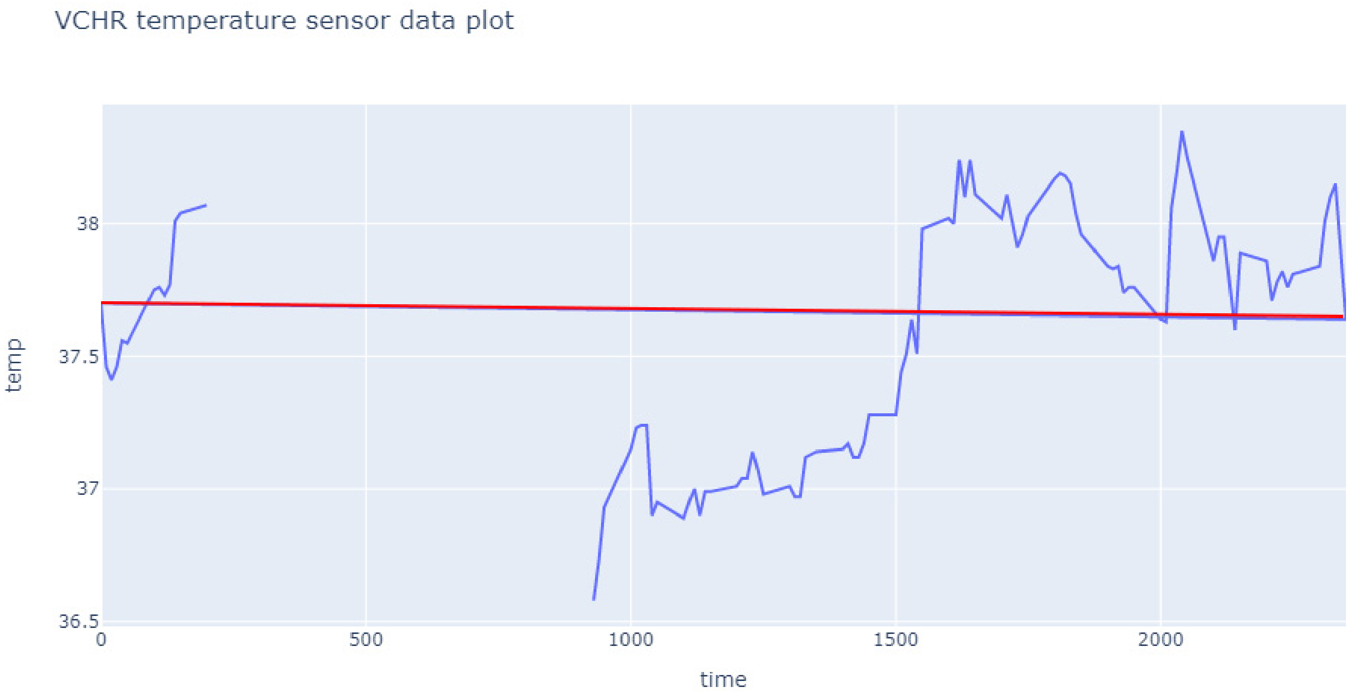


Figure 4

Temperature Sensor Reading Graph

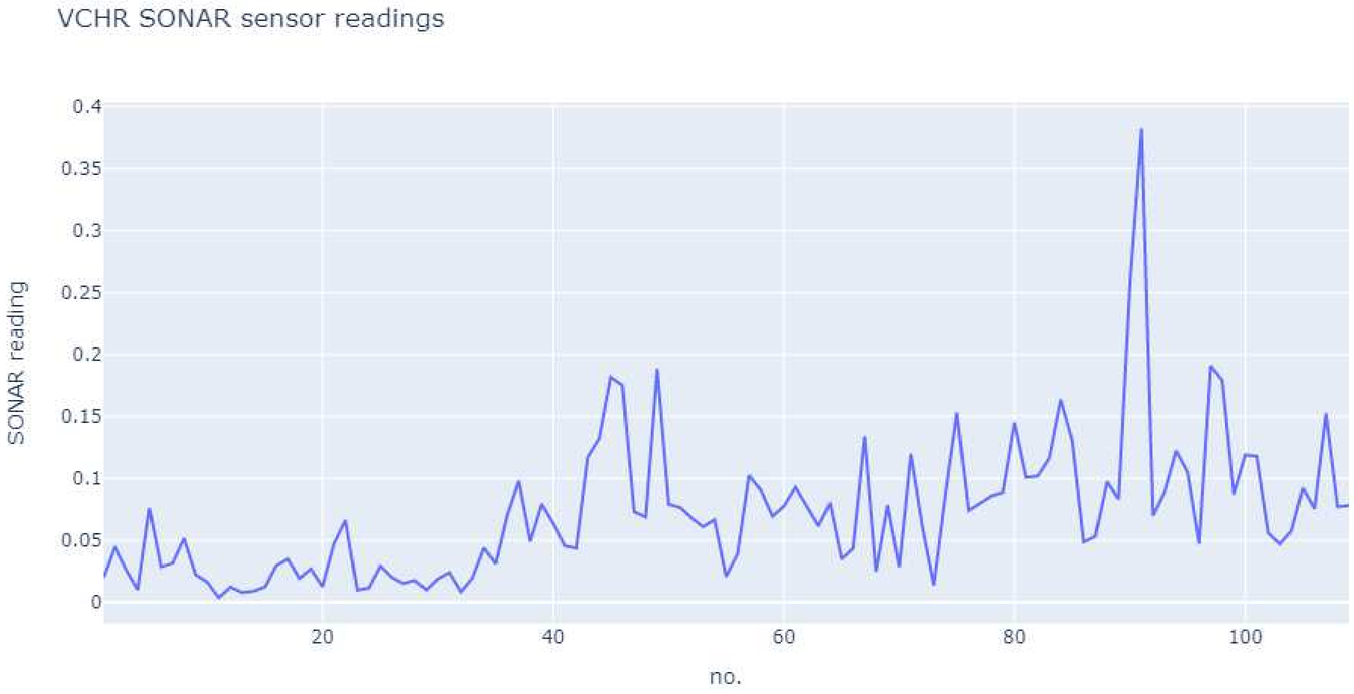


Figure 5

SONAR Sensor Reading Graph