Voiceprint Authentication System

Palestine Polytechnic University

Team: Mohamad Ateyyah Salah, Mohamad Shalodi, Mahmoud Skafi

Supervisors: Dr. Radwan Tahboub, Dr. Mousa Farajallah

January, 2021

Acknowledgement

It is with great pride that we hereby present our bachelor thesis. During this 4 years long process, we have gained a lot of knowledge about computer science and it has been beneficial to our development in furthering our career, our projects and studies in the future.

We would like to express our special thanks of gratitude to our supervisor Dr. Radwan Tahboub for his able guidance and support. By giving his feedback from his major experience and knowledge of information security, he made a great effect on our project.

Secondly, we would also like to express our appreciation to our parents and friends who helped us a lot in finalizing this project within the limited time frame.

Above all, to the great almighty, Allah S.W.T, the author of knowledge and wisdom, for his countless blessings. This project was made from the support and contribution of our group members. So we will thank each one of us.

Dedication

This study is wholeheartedly dedicated to our beloved parents, who have been our source of inspiration and gave us strength when we thought of giving up, who continually provide their moral, spiritual, emotional, and finncial support.

To our brothers, sisters, relatives, mentors, friends and classmates who shared their words of advice and encouragement to finish this study.

And lastly, we dedicated this book to the Almighty God, thank you for the guidance, strength, power of mind, protection and skills and for giving us a healthy life. All of these, we offer to you.

Abstract

Voiceprint authentication system aims to authenticate users using their voiceprints in an easy and fast way.

Using voiceprint detection and speech recognition algorithms(SRS), we designed a system on a Raspberry Pi which will receive audio from the user using a microphone installed on the Raspberry Pi. Then, process it and check if the recorded voiceprint matches the corresponding voiceprint stored in the database for that specific user, and check if the spelled words are correct. This process of authentication is processed by the authentication system to perform an operation for that user, for instance, a small door sample system, which is controlled by a microcontroller to open.

This project combined the concepts of Artificial Intelligence(AI) and Internet of Things(IoT) in a new authentication method.

The system can succeed in authenticating users in a fast way and is able to be embedded in a subsystem that needs user authentication.

Keywords: *Voiceprints, Speech Recognition, SRS, Internet of Things(IoT), Artificial Intelligence(AI).*

Contents

List of Figures

1 Introduction

1.1 Overview

User Authentication (UserAuth) is the process of verifying a human to a computer system in order to do an activity like login to a website, making financial transactions, or simply opening a door.

UserAuth can depend on different types of things in our life, the factors of UserAuth are something you know, something you have, something you are, somewhere you are and something you do.

Passwords can be seen as the first solution for UserAuth, which are knowledge-based, simple and easy to implement, cards like VISA card and ATM card which are physical cards depends on physical possession for UserAuth, also human biometrics like fingerprint and voiceprint can be used for UserAuth.

The reliability of the UserAuth system depends not only on the number of factors involved in the process but also on the implementation and the technology used.

In this project, we implemented one of the "something you are" factors, which is the human voiceprint. It has pros, it has cons, but mainly, it is a very reliable security method to be used.

1.2 Motivation

Passwords, the most commonly used approach for UserAuth, although it's simple and easy, it has a lot of weaknesses in security perspective, like the ability to crack passwords using a brute-force attack or dictionary attacks, sometimes passwords are hard to remember, also repeated passwords for multiple systems can be a fatal problem.

All these security problems can be solved when using biometric-based UserAuth techniques like voiceprint identification, which is way more secure than passwords, and the authentication process can be (based on the implementation) faster and more convenient.

For such systems like a small door system, a voiceprint UserAuth system will be the most proper way of authenticating, so instead of entering a password that can be detected or cracked, or inserting a card that can be stolen, use your unique voiceprint to authenticate.

It's worth mentioning that we used speech recognition and voiceprint technologies. These technologies are very common among the top companies' products such as Alexa, Google, Siri, and others, so we will mimic their recognition systems to be used in an efficient security scheme.

1.3 Importance

The voiceprint UserAuth is a very strong type of authentication because it depends on something you are, not something you know or something you have, so no room for stealing identity or losing availability.

Unlike other biometric-based authentication methods, voiceprint does not require special hardware like fingerprint sensor or iris-scanning equipment, it just requires a microphone which is cheap and easy to get.

This system can be used as a subsystem almost in every IoT system, just with a little modification on the system software, i.e: authentication protocol.

1.4 Objectives

- To design and implement a small door system controlled by an Arduino microcontroller which can open the door.
- To design and implement the voiceprint authentication system on a Raspberry Pi microcontroller.
- To design a master-slave authentication protocol to be implemented in the system.
- To connect the two systems and to make them function together as one whole system.
- To make the ability of implementing this voiceprint authentication protocol in any IoT device.
- To test and validate the functionalities of the system.

1.5 Problem Analysis

Many computer-based systems require authentication, which assures that unauthorized people won't have access to critical data or special services, some doors are considered included. Let's say that we have a room, which has critical information, so we want it secured by not allowing everyone to enter it, but we will allow for a set of known people to be able to enter. To make this work, we need to make sure the door is the only way to enter the room, and it should be controlled by a computer.

The first solution that came in mind is that we put a secret password which only that known group of people will know, so anyone that wants to enter the room should enter the correct password to the computer, but what if someone somehow got to know the password? Then he would be able to enter the room although he is not authorized to.

Another idea is to create special cards that only that group of people will have, so anyone that wants to enter, he can not enter unless he has the card, but what if the card is stolen!

Another solution is to have a fingerprint scanner attached to the computer, so fingerprints are unique and no two persons have the exact same fingerprint, but what if someone gets to have a copy of an authorized person's fingerprint, like copying the fingerprint on a paper tape just like in movies!

After all, here comes the idea of using voiceprint for authentication, so anyone that wants to enter the room, he has to say some words, the computer checks if his voiceprint matches a voiceprint of an authorized person, if there is a match, then we can be sure that this is an authorized person since voiceprint is unique and cannot be duplicated. But what if someone records an authorized person while he is talking, then he will have his voiceprint and now he can have access, right? here comes the idea of using speech recognition technology, so every time someone wants to enter the room, the computer shows a random set of words, which that person has to say, then the computer checks if the spelled words match the words that the person asked to say and checks for the voiceprint, so if a person has a recorded sound of an authorized person, he cannot do anything.

In our project, we have a small door system, and the voiceprint authentication system which is separated from the door system in order to make it integratable with other systems, so we have integrated the two systems together by connecting them online, also we have used a Speech-To-Text system in order to initiate the authentication process using speech commands.

1.6 List of Requirements

- The system should respond to the user in less than 3 seconds(preferably).
- The system should turn on a green light for the user after the success of the authentication process.
- The system should output random words on a screen for the user when he initiates the authentication process.
- The randomly generated words that are displayed for the user must not be redundant.
- The voiceprint detection algorithm should be executed locally on the Raspberry Pi.
- The system should have a Speech-To-Text system.
- The system will have the ability to receive audio signals from the user to command the system to do specific functions.
- The system should contain the function of adding a new user, a function to reset the system and to initiate the authentication process itself.
- The system should execute a function if there is a permission for that user, so not every user can create new users.
- The system should prevent the user from attempting new authentication tries for 3 minutes after 5 consecutive false attempts(configurable).
- The system positive acceptance ratio should be more than 80%.

2 Background

2.1 Overview

Having a reliable access control scheme is one of the main security concerns for any system, as well as choosing suitable authentication methods.

The most common way of authentication used worldwide is the text-typed passwords, they have the advantage of being generally secure and reliable, though it might not be the best option in the matter of accessibility.

There are also biometric authentication methods, these include fingerprint, iris scanners, and voice pattern. These methods give high security because they depend on something you are, not something you have.

But when comparing each one of them we find that both iris scanners and fingerprints require a dedicated hardware components wherever they are used, while the voiceprint recognition requires only a microphone in addition to the processing equipment (which is often available).

This is where voiceprint recognition has the advantages of high security and accessibility, as well as low cost to be used.

Voice is also used commonly to send messages and commands on mobile devices or even computers by using a Speech-To-Text system.

Using voiceprint as the method of the authentication and combine it with voice commands would result in a powerful tool that is able to control sensitive systems.

Feature	Fingerprint	Palmprint	Retina	Iris	Face	Vein	Voiceprint
Ease of Use	High	High	Low	Medium	Medium	Medium	High
Accuracy	High	High	High	High	High	High	High
Cost	High	Very High	Very High	Very High	High	Very High	Low
User Acceptance	Medium	Medium	Medium	Medium	Medium	Medium	High
Remote Authentication	Available	Available	Available	Available	Available	Available	Yes
Mobile Phone Collection	Partly Available	Yes	Available	Available	Yes	Available	Yes

Figure 1: Comparison Types of Biometrics Signatures Based on Various Factors [1].

2.2 Theoretical Background

Any security project has to maintain the balance of the CIA triad, which consists of Confidentiality, Integrity, and Availability. Also, it needs good access control techniques by determining who or what can view or use the available resources in our computing environment.

In short, confidentiality is the set of rules that are limiting access to the information, whereas integrity is to be sure that the information we have is accurate and trusted, and availability is the reliability for authorized users of accessing this information.

We can define the process of speech recognition as receiving, interpreting and understanding the spoken sentences by a machine or a program, this advanced and got its fame due to the high development in Artificial Intelligence.

This allowed people around the world to interact with technology with the maximum availability by only speaking without using their hands.

In 1976, computers were only able to understand about 1000 words, which developed to be 20,000 in the 1980s, so the growth was exponential over the past decades in speech recognition technologies.

In 1996, a continuous speech was able to be recognized as IBM introduced the first speech recognition product that can do so.

And now it is clear how speech recognition became such a huge part in this technology-depending world, companies make their own products such as Amazon's Alexa, Apple's Siri, Google Assistant, and Microsoft's Cortana.

By using machine learning and other algorithms, speech recognition gave people the ability to turn their spoken words into written text.

But on the other side, all speech recognition systems and programs make mistakes and errors even with the high accuracy rates that are improving. Another factor is the background noise that could generate false output for the system to handle. In addition to the great challenge of having words that sound like each other with different spelling, for example, the words 'where' and 'were'.

Before we go into voiceprint authentication details, we would like to define the authentication as the process of verifying the identity of someone who attempts to connect to the digital resource or -like in our case- accessing a place physically. The speech information in a human voice that is carried by the acoustic frequency spectrum is defined as the Voiceprint. It is similar to the fingerprints in the way that both have a unique biometric signature, so it can be used as an identification method.

The process of recognizing this voiceprint (voiceprint recognition) is also a biometrical identification technology that works as a feature extractor for the voice signals of the speaker, in order to know his identity.

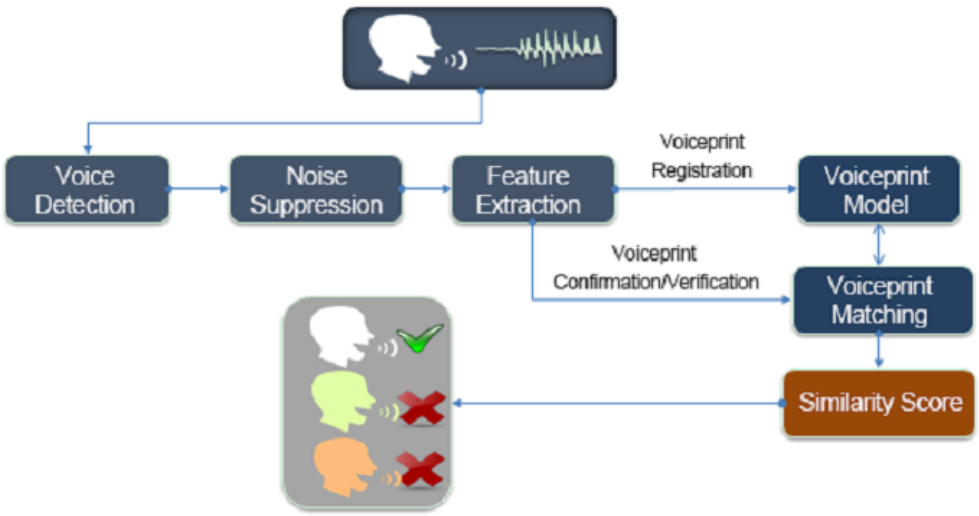


Figure 2: Working Process of a Voiceprint Recognition System [1].

The voiceprint is unique that everyone has his own. It was formed gradually throughout the development of his vocal organs, so it's worth mentioning that no matter how similar the voice can be to the original input voice, the voiceprint of them will keep different from each other.

The performance of biometric systems is expressed on the basis of the False Acceptance Rate (FAR), which is the percentage of identification instances in which unauthorised persons are incorrectly accepted, and the False Rejection Rate (FRR) which is the percentage of identification instances in which authorised persons are incorrectly rejected [10].

2.3 Literature Review

Most natural form of human communication depends on speech. In order to understand human speech by machines, computers can act as an intermediate for human experts, so that it can respond accurately and reliably to human voices. This can be achieved by a speech recognition system, which permits a data processor to identify the words a person speaks in a microphone or telephone, and converts them into written text. Speech Recognition (SR) and its application is the thrust area in research during the past three decades which is carried out on various aspects, especially in the field of Information and Communication Technology (ICT) for speeding up scientific advancements. Also, people interact with the system to utilize the technology in greater amounts without the acquaintance of operating a computer keyboard. At present, Automatic Speech Recognition (ASR) is effectively utilized for communication between human and machines [11].

Voiceprint Recognition System also known as a Speaker Recognition System (SRS) is the best-known commercialized form of voice Biometrics. Automated speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices. In contrast to other biometric technologies which are mostly image-based and require expensive proprietary hardware such as vendor's fingerprint sensor or iris-scanning equipment, the speaker recognition systems are designed for use with virtually any standard telephone or on public telephone networks. The ability to work with standard telephone equipment makes it possible to support broad-based deployments of voice biometrics applications in a variety of settings. In automated speaker recognition, the speech signal is processed to extract speaker-specific information. These speaker specific information are used to generate voiceprints which cannot be replicated by any source except the original speaker. This makes speaker recognition a secure method for authenticating an individual since unlike passwords or tokens; it cannot be stolen, duplicated, or forgotten [2].

To understand how voiceprint detection works, we need to study how humans produce voices. The origin of differences in the voice of different speakers layers in the construction of their articulatory organs, such as the length of the vocal tract, characteristics of the vocal chord, and the differences in their speaking habits.

Teeth, lips, tongue, oral cavity, jaw, nostril, nasal cavity, larynx, lung, and diaphragm, they all participate together in the voice production mechanism, and also they differ between people and that's what makes voiceprints unique.

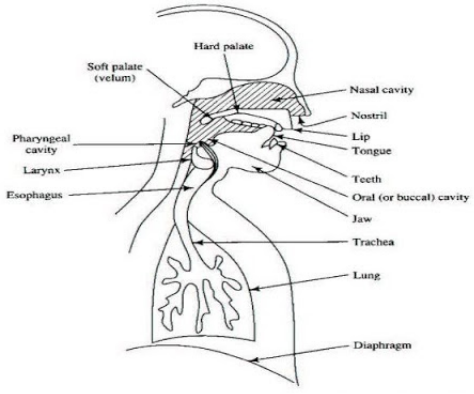


Figure 3: The Speech Production Mechanism [2].

2.4 Speech to Text

There are multiple steps involved in the process of converting speech into text. When you're talking you create a series of vibrations. These are translated into digital language by the analogue-to-digital converter or the ADC.

The ADC is able to complete this conversion by sampling sounds from an audio file and taking frequent, very detailed measurements of the waves. The system has a filter to distinguish the sounds that are relevant and differentiate frequencies. The speed of the speech is also modified and the volume set at a control level.

The next stage involves segmenting the signal into hundredths or thousandths of seconds and matching these parts to phonemes (a phoneme is a unit of sound that distinguishes one word from another in a particular language). There are over 40 phonemes within the English language. Each phoneme is then examined and evaluated in relation to other phonemes around them, and the system then runs the network of phonemes through a complicated mathematical model to compare them to well-known sentences, individual words and phrases. The system using machine learning then creates text based on what is most probable that the person said. This is either presented as a chunk of text (text file) or as a final computer-based command [12].

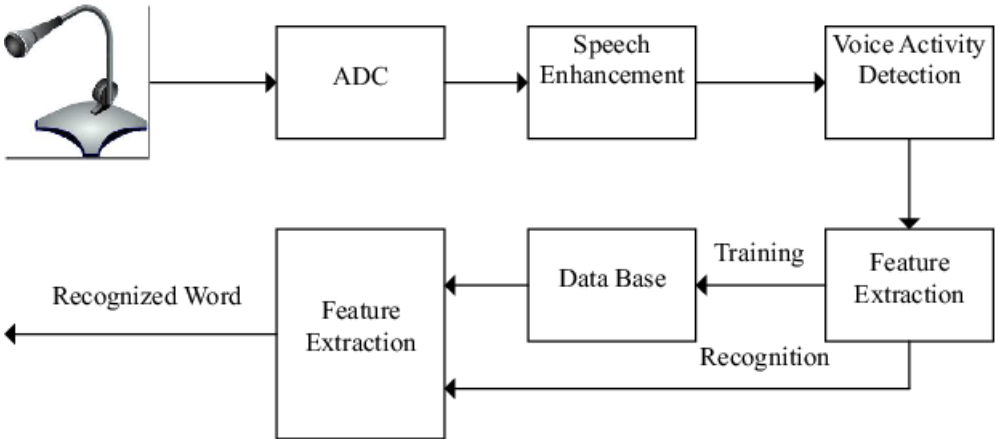


Figure 4: General Model of a Speech Speaker Recognition [3].

2.5 Mel-Frequency Cepstrum Coefficients

MFCCs [4] are the most widely used acoustic feature for speech recognition, speaker recognition, and audio classification.

Calculation of MFCC coefficients:

Step 1: Cut the signal into several windows that intersect each other. This is called sliding windows.

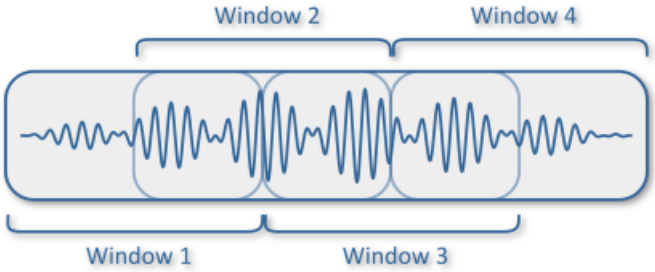


Figure 5: Signal Sliding Windows [4].

Step 2: In order to reduce the spectral distortion, it is necessary to apply a specific window to the signal. There are many types of windows but the most known are Hamming, Hann (also called Hanning), and Blackman.

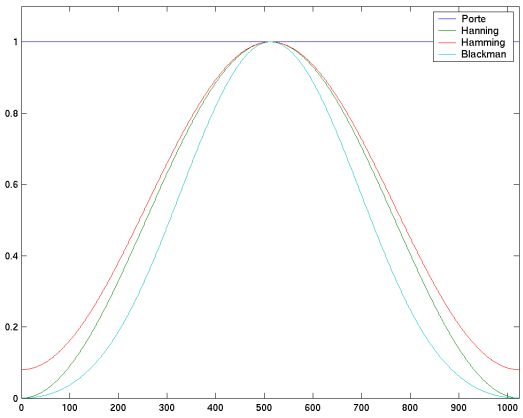


Figure 6: Temporal Shapes of Some Windows [4].

Step 3: Apply the FFT (Fast Fourier Transform) to the window to bring out the magnitude, so we get the spectrum. A fast Fourier transform (FFT) is an algorithm that samples a signal over a period of time (or space) and divides it into its frequency components. The frequency spectrum of a signal is the distribution of the amplitudes and phases of each frequency component against frequency.

Step 4: Then we move on to Mel's scale. This is a psychoacoustic scale of pitches of sounds, in the sense of their identification between bass and treble, which unity is Mel. Mel is related to Hertz (Hz), the unit of measurement of the International System for Frequencies, by a relationship based on human hearing. It is a frequency scale closer to what the human ear is actually able to capture.The transfer formula is rather simple: $m = 2595log(1 + f/700)$.

To simulate the human ear, it is necessary to go through a Filter Bank, a filter for each frequency. These filters have a triangular bandwidth response.

Step 5: Finally, we work on the cepstrum. We convert the logarithmic spectrum of Mel's scale into time's scale by using the DCT (Discrete Cosine Transform). A cepstrum is the result of taking the inverse transform of the logarithm of the estimated spectrum of a signal. The name "cepstrum" was derived by reversing the first four letters of "spectrum".

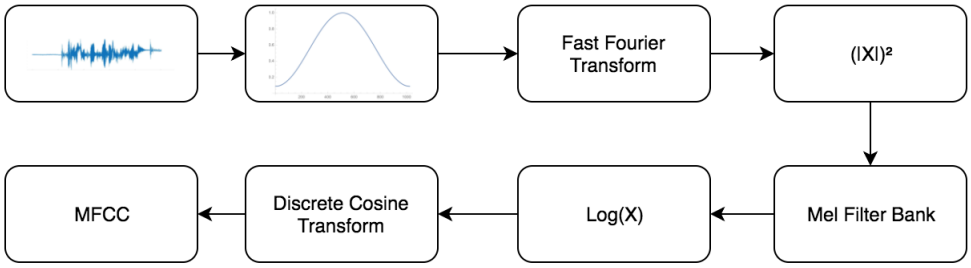


Figure 7: MFCCs Extraction Algorithm [4].

2.6 Linear Predictive Coding

LPC methods are the most widely used in speech coding, speech synthesis, speech recognition, speaker recognition and verification and for speech storage.

LPC methods provide extremely accurate estimates of speech parameters, and does it extremely efficiently, basic idea of Linear Prediction: current speech sample can be closely approximated as a linear combination of past samples.

For periodic signals with period Np, it is obvious that $s(n)s(n - Np)$ but that is not what LP is doing; it is estimating $s(n)$ from the $p(p << Np)$ most recent values of $s(n)$ by linearly predicting its value [5].

For LP, the predictor coefficients are determined (computed) by minimizing the sum of squared differences (over a finite interval) between the actual speech samples and the linearly predicted ones.

LP is based on speech production and synthesis model:

- Speech can be modeled as the output of a linear, time-varying system, excited by either quasi-periodic pulses or noise.
- Assume that the model parameters remain constant over speech analysis interval.

LP provides a robust, reliable and accurate method for estimating the parameters of the linear system (the combined vocal tract, glottal pulse, and radiation characteristics for voiced speech).

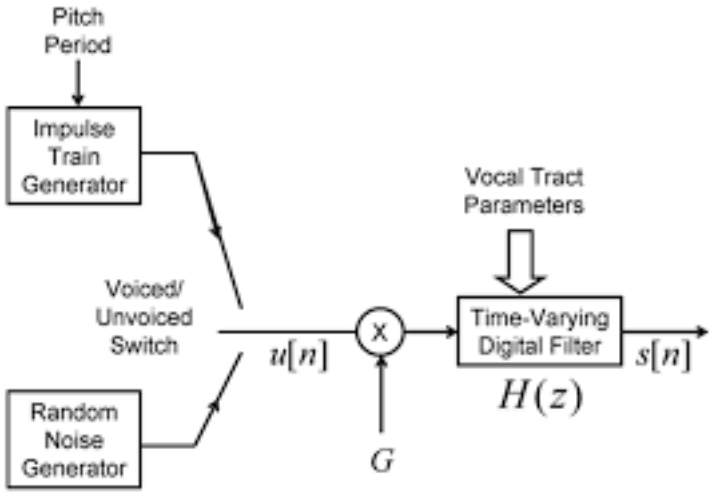


Figure 8: LPC Diagram [5].

2.7 Design Options For Hardware Components

Basically, we have two microcontrollers, one for the authentication system which will process authentication requests, and the second one for the sample door controller system, which will control opening or closing the door, also we will have a small door and mechanical-electronic hardware to do open/close operations.

As a microcontroller, we used a Raspberry Pi, which has high specifications in order to process audio and feature extraction in little time.

Microphone for detecting speech installed on the authentication system microcontroller, also a screen for displaying the words that the user required to say.

For the door controller system, we used Arduino microcontroller.

2.8 System Software Component

Raspbian OS: which is an operating system with the set of programs and utilities that makes the Raspberry Pi run. It is a free OS based on Debian which uses the command line to manage the operation and is optimized to be compatible with the Raspberry pi hardware.

Text Editor to be used in programming Raspberry pi in Java language.

Windows 10, the operating system to run the text editor to program the Raspberry Pi.

2.9 Programming Languages

We used open source libraries for voiceprint detection and speech recognition, so it depends on the programming languages used in those libraries, but we used java.

Java has the advantage of being easily understandable, it also has many ready-to-use libraries that would help us build the interfaces between the two systems.

For the Arduino controller, we wrote the code in C++ language.

The server that connects the controllers is built using NodeExpress Javascript programming language, and for the database we use mongodb.

3 Design

3.1 Methodology

In this chapter, we talked about how we built the systems, and design a block diagram for each system and explain each block and how it works, also we talked about the authentication protocol, and finally, we talked about the Speech-To-Text system and how it works.

3.2 General Block Diagram

We defined the general diagram of the system. The user will interact directly with the user authentication system. After he was authenticated (by inner processing within the chip itself), an "Enter the system" command can be said by the user, which will initialize the authentication process.

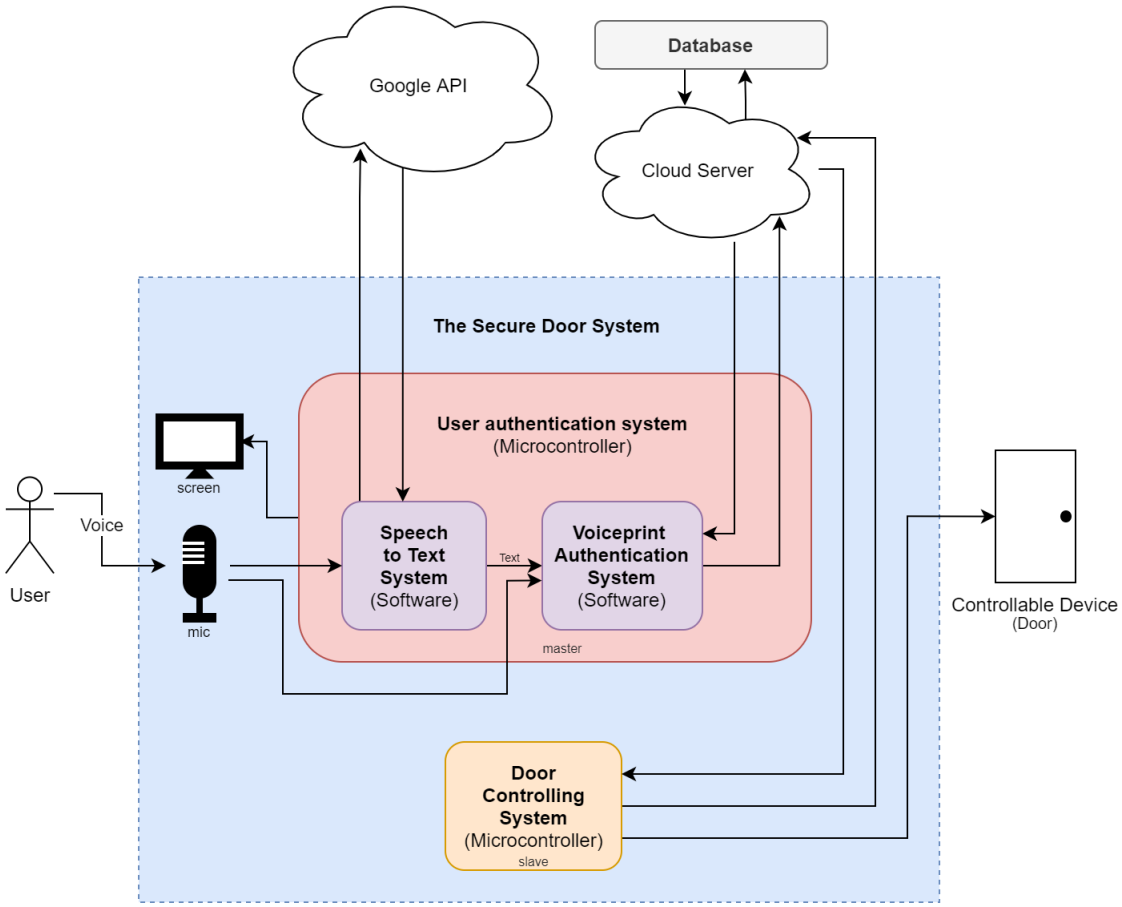


Figure 9: General Block Diagram.

As you can see, the authentication system is the master system, and the door controller system is the slave system.

As mentioned before, the master system(User authentication system) is the system that will take care of all the voice checking and deciding, this system contains two main subsystems that will work in collaboration with each other to make a decision. The first subsystem of the two is the Speech-To-Text system, which will try to convert the audible speech to string texts that are ready to be processed by the second subsystem.

The Voiceprint Authentication System is where the checking happens. At its stage, the system will translate the meaning of the words the user said for the system, and process them by comparing the words in addition to the voiceprint to get to a conclusion. that conclusion will be either granting the user the access to say the command to open the door, or the authentication attempt fails.

The Database is needed to store the users' information, containing their role, and their voiceprint extracted features, also it contains the phrases that will be generated randomly.

3.3 Voiceprint Authentication System

This system is the master system, it's on a Raspberry Pi microcontroller. Mainly, this system contains the user authentication process, adding a new user process, request to open the door process, and show authentication history.

3.3.1 User Authentication Process

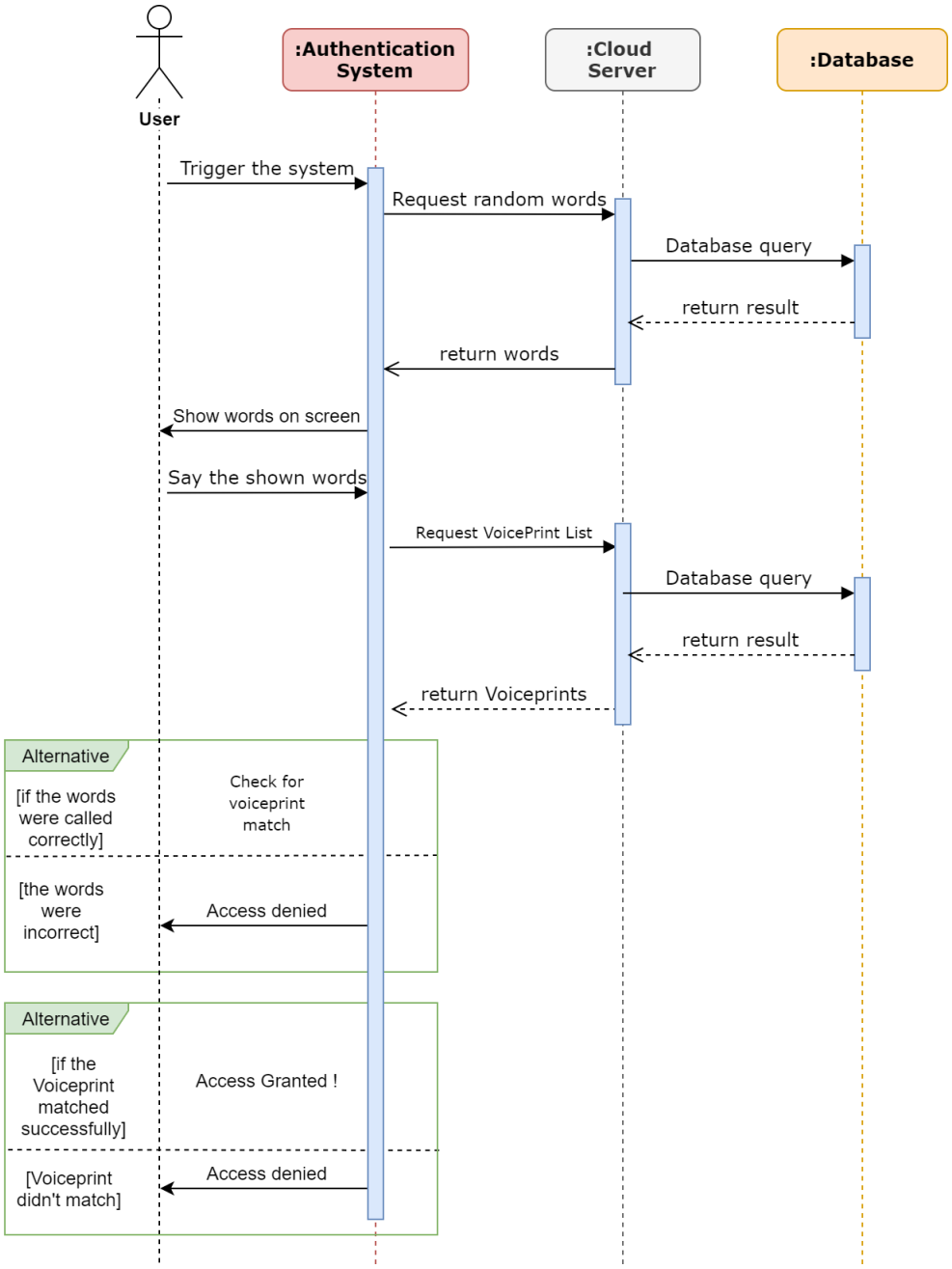


Figure 10: User Authentication Process Sequence Diagram.

As shown in the figure, the process starts with a user triggering the authentication system, the system will initiate the authentication process by requesting random words that are already stored in the database. Then, the words are going to be shown on the screen that is connected to the microcontroller. At that moment, the Speech-To-Text system will be in a listening state (it's not shown in the diagram because it acts as an interface). It will wait for the user to say the shown words, and when he does, the authentication will first check for words matching, then the voiceprint matching.

3.3.2 Add New User Process

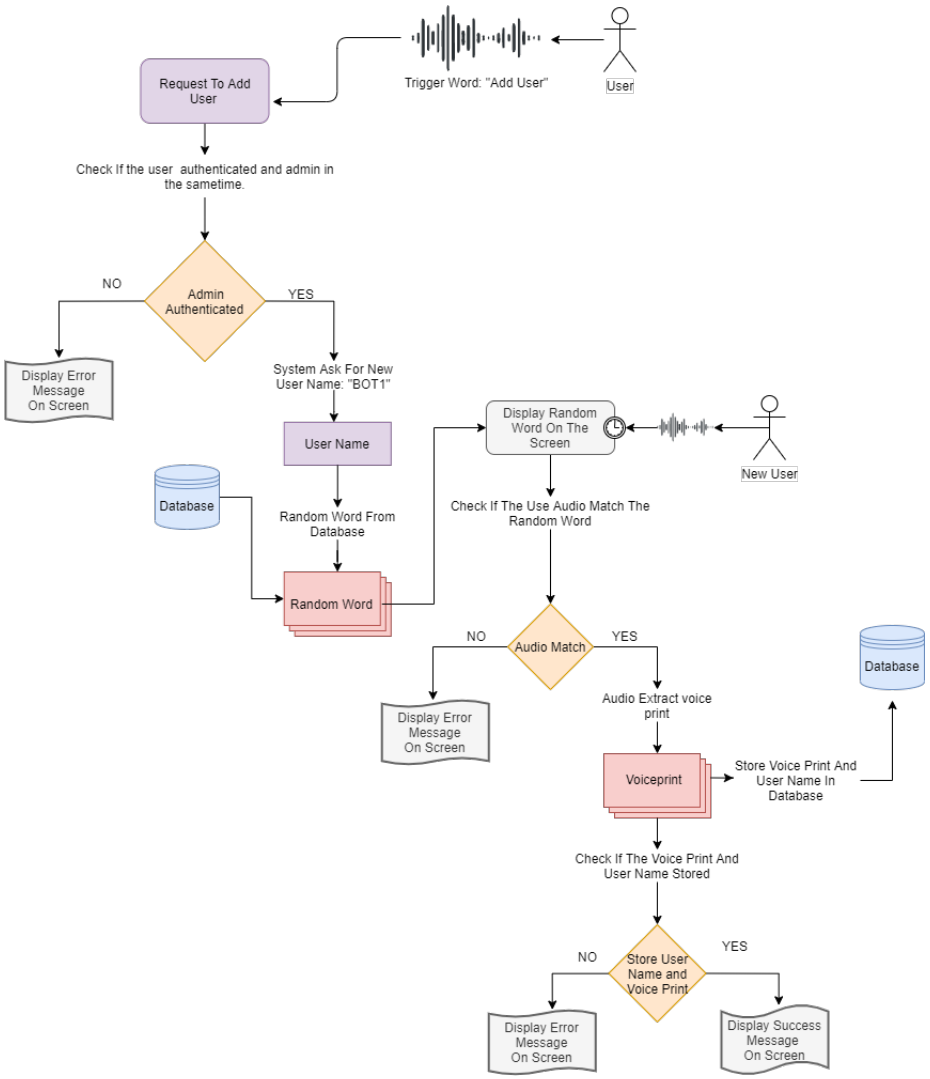


Figure 11: Add User Flow Chart.

The process of adding a new user to the system's database requires the current session to be an admin session, in which the last authorized person was the admin.

So it starts with the admin saying the keywords "add user". The Speech-To-Text system will then inform the Authentication System which will initiate the process of adding a new user.

To add the new user, the Authentication System generates random words from the database to be shown on the screen for the new user, who is asked to read them all. After he does that, the Authentication System will be running the process of voiceprint feature extraction. Then, the system checks whether they were called correctly or not.

If they were correct, the Authentication System will store the new user's information that contains his Voiceprint (his privileges are determined as a normal user, and can be modified later), Otherwise, he is asked to retry the process.

3.3.3 Show Authentication History

For every authentication attempt, it will be logged on the server, only the admin can show the authentication history.

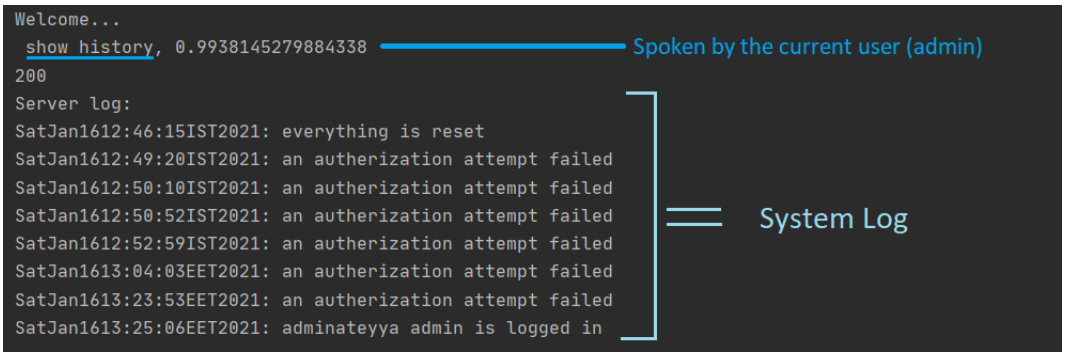


Figure 12: Authentication History Log.

3.4 Speech-To-Text System

Basically, this is the system that the user will be interacting with, it's main function is to listen to the user and convert the sound into words, so it's participating in the authentication process, and it's responsible for triggering sequences of words to call a function.

The algorithms used in this form of technology include PLP features, Viterbi search, deep neural networks, discrimination training, WFST framework, etc.

For Speech-To-Text service, we used Google Cloud Speech-To-Text Api[6].

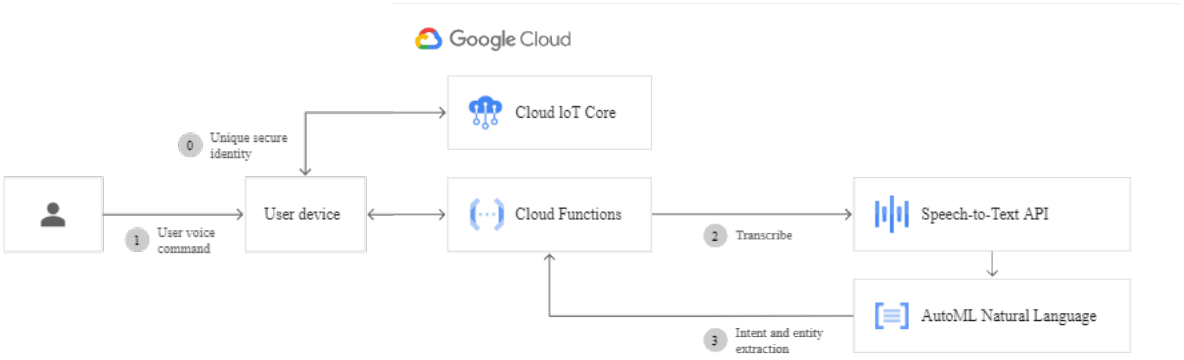


Figure 13: Google Cloud Speech-To-Text Api [6].

3.5 Door Controlling System

This system's main functionality is only limited to opening or not opening the door. It will receive its commands from the system's online server.

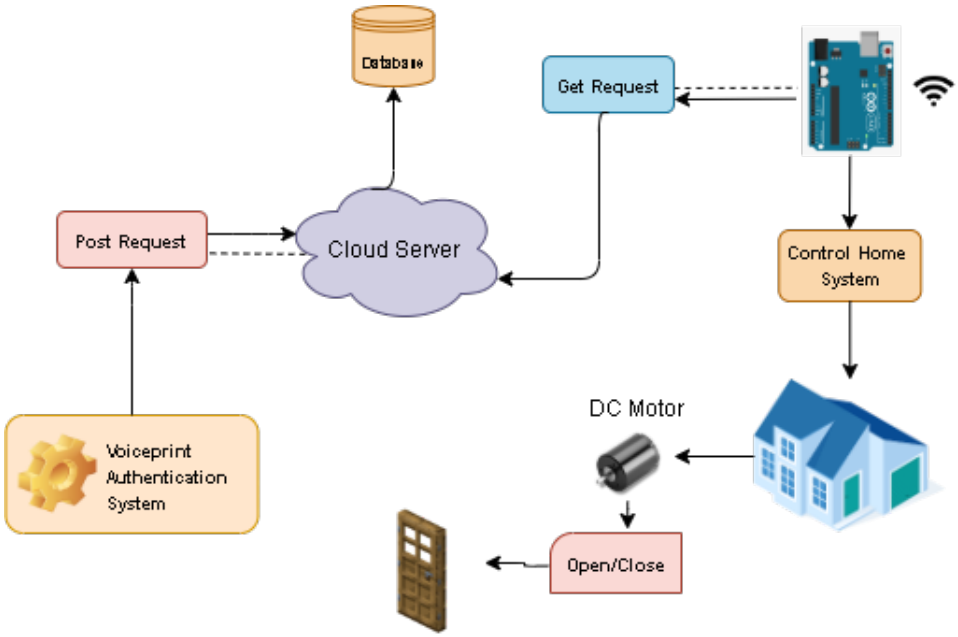


Figure 14: Door Controlling System Diagram.

3.6 Voiceprint Detection

Mel Frequency Cepstral Coefficient (MFCC) [4] is considered a key factor in performing Speaker Identification. But, there are other features lists available as an alternate to MFCC, like Linear Predictor Coding(LPC) [13], Spectrum Sub-band Centroid (SSC) [14], Rhythm, Turbulence, Line Spectral Frequency (LPF), Chroma Factor, etc.

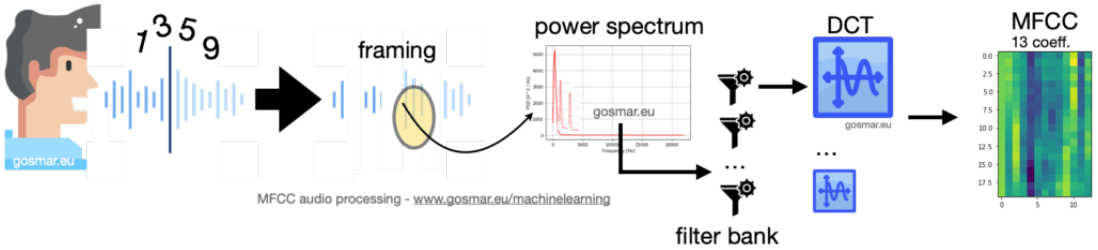


Figure 15: Voiceprint Extracting Using MFCC [7].

For voiceprint detection, we used amaurycrickx/recognito github repository [15].

4 Implementation

4.1 Description of the Implementation

Mainly, the voiceprint authentication system is implemented on a RaspberryPi microcontroller, connected with a microphone and a screen. An Arduino connected with a dc motor through h-bridge driver moving a CD ROM driver representing the door in the controlled slave system (door controller system). All contained in a small wooden house.

Another part is the server on the cloud, which will be the interface between the Raspberry Pi and the Arduino(Wemos D1). Though the two devices are working together on the system, we didn't make them directly connected to each other for the following reasons:

- The database of the system is on the cloud, not locally on one of the two devices.
- This method makes the system more secure, every action that happens on one of the components is logged on the server.
- The online server would be easily manipulated to send slave-controlling commands not only from the Raspberry Pi device, but also from any connected apps/devices in the future. In that case, there would be multiple Master systems connected.

4.2 Software

The project contains four main software components.

4.2.1 Speech-To-Text System

Which runs on the Raspberry Pi, it's responsible about handling user interactions and participating in the authentication process as described in the design chapter, an open source java code is imported to the project from Github, which uses Google Cloud api for speech-recognition, a new project in google cloud is created and added the implementation for handling the previously defined operations.

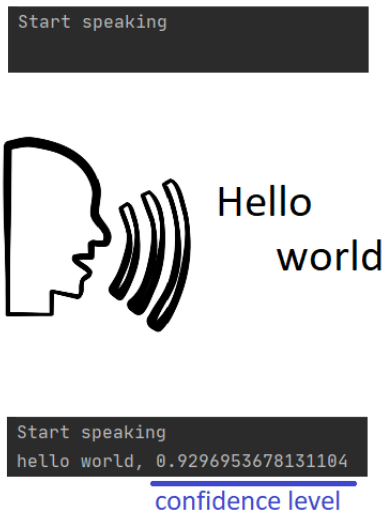


Figure 16: Speech-To-Text System Test.

4.2.2 Voiceprint Detection System

An open source java library from github, which is amaurycrickx/recognito github repository [15] is used, which uses LPC feature extractor and Windowed feature extractor, to create a voiceprint for a given wav file or to match a wav file with other stored voiceprints.

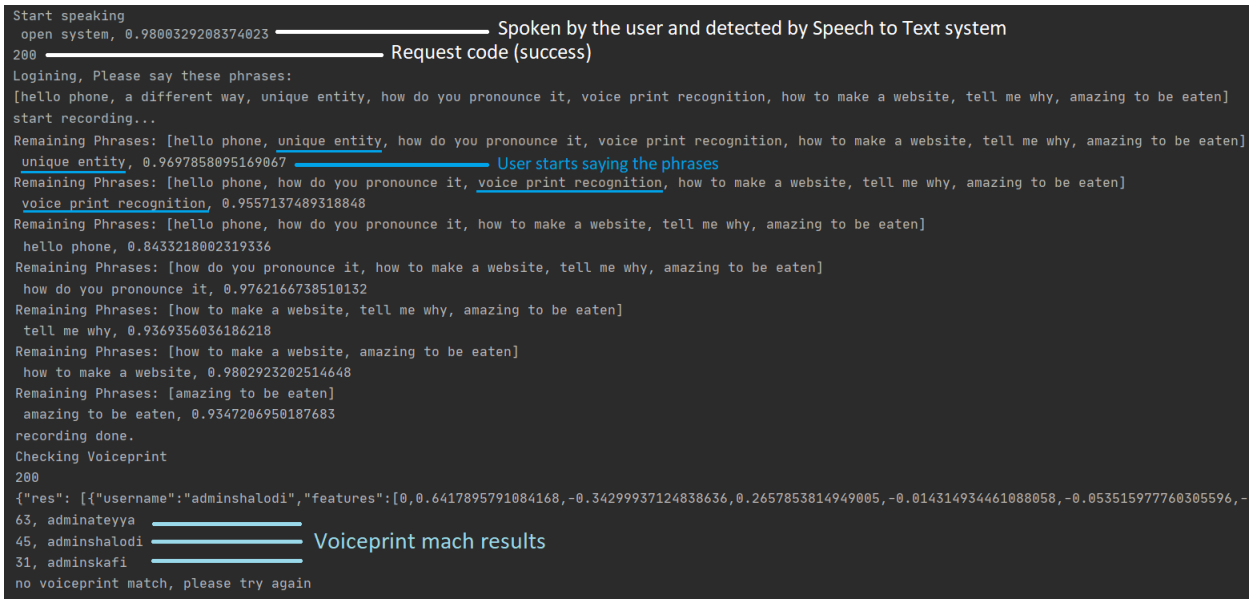


Figure 17: Voiceprint Detection Process.

4.2.3 Door Controller System

Which runs on an Arduino Wemos microcontroller, it's responsible for opening/closing the door, implemented in C++ using Arduino IDE.

4.2.4 Online Server

implemented using Node ExpressJS as a RESTfull API, responsible for handling the interaction between the authentication system and the door controller system, it contain API endpoints for the door controller to read commands sent from the authentication system, it reads the command and execute it, also it contain endpoints to edit a command for the authentication system, also it's responsible for reading/editing/deleting from the database.

Part of the software used is the main voice recorder at the first stage of the program, which is manipulated to get its output file to be the proper input file for the next stage, in terms of file extension, Sample Rate and Encoding. It is a Java code on the Raspberry Pi that uses the Speech-To-Text microphone input buffer data and stores it in a wav file when recording is running.

So the software components are working as multiple blocks connected together by communicating through requests sent to the online server, and some blocks are represented as a Master system, which will be the commander for the connected slave system.

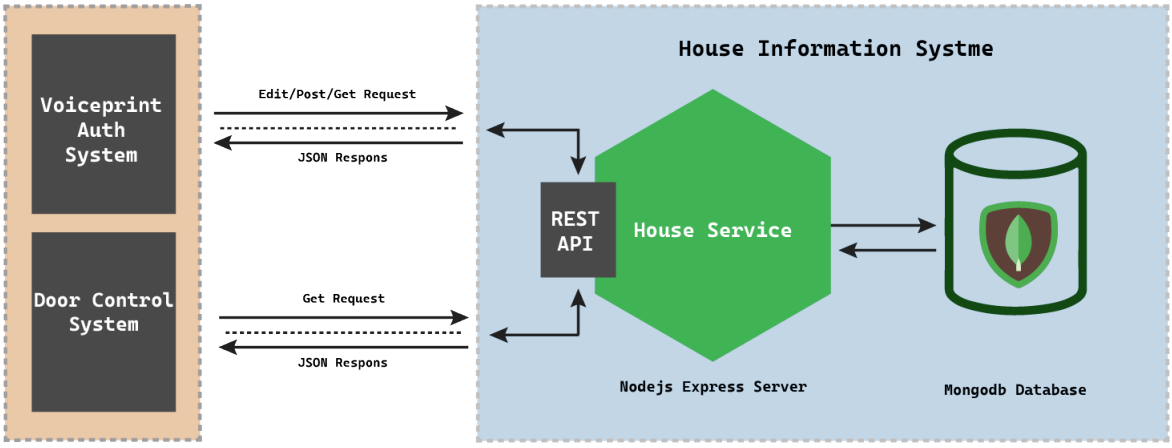


Figure 18: NodeJs Restfull Api System.

4.3 Hardware

There are two main hardware components in our system:

4.3.1 Raspberry Pi Microcontroller

Which the authentication system is running on, It is a microcontroller which can host and run Linux operating systems(Raspbian OS). It has the processor Broadcom BCM2711: a Quad core Cortex-A72 (ARM v8) 64-bit SoC @ with a clock speed of 1.5GHz, also it contains a Random Access Memory with the size of 2GB LPDDR4-3200 SDRAM.

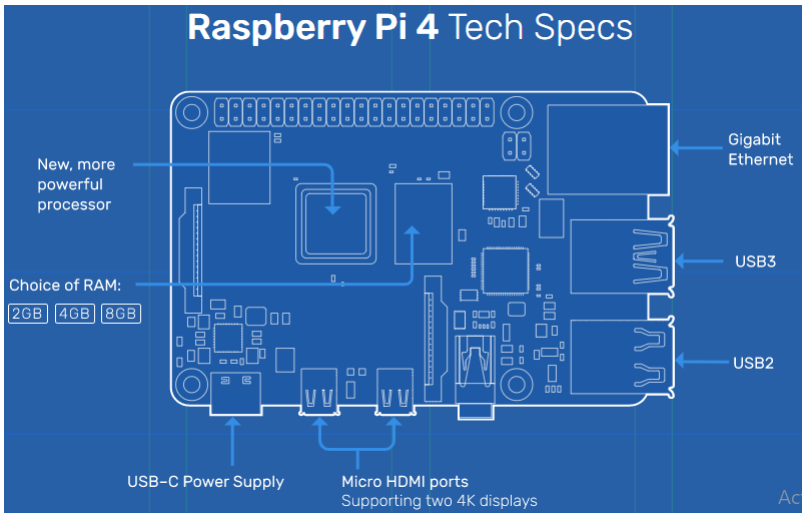


Figure 19: Raspberry Pi 4 Tech Specs [8].

It supports Gigabit Ethernet and 2.4 GHz and 5.0 GHz IEEE 802.11ac wireless connection, and will be connected wirelessly to the internet in our project through wifi. Also it's connected to a computer screen via HDMI cable and a microphone via USB Sound Card.

Figure 20: USB Sound Card.

4.3.2 Arduino Microcontroller

Where the door controller is running, it is an Arduino Wemos D1 Microcontroller. It's specifications are shown in the table below.

Operating Voltage	3.3V
Digital I/O Pins	11
Analog Input Pins	1
Clock Speed	80MHz/160MHz
FLash	4M bytes
Length	68.6mm
Width	53.4mm
Operating Voltage	3.3V

Figure 21: Arduino Wemos D1 Microcontroller.

It is connected to a DC motor using Dual H-Bridge Motor Driver L298N, which is connected to a CD ROM driver representing the door in the wooden house, and a 9V battery to supply the Arduino.

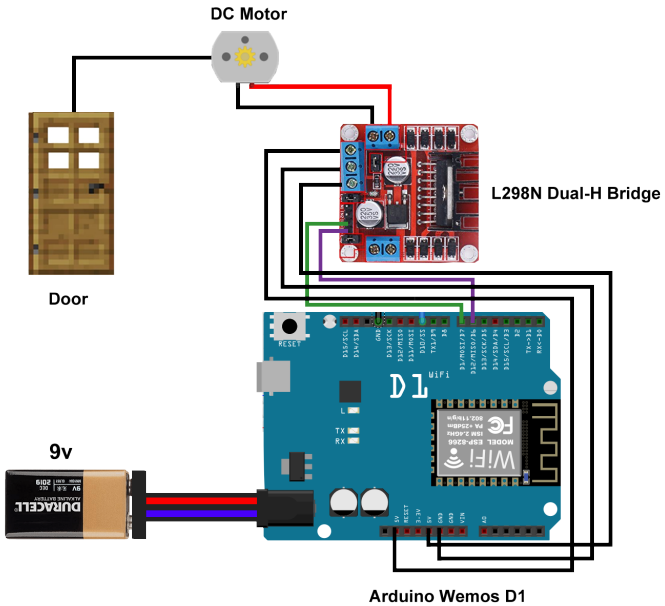


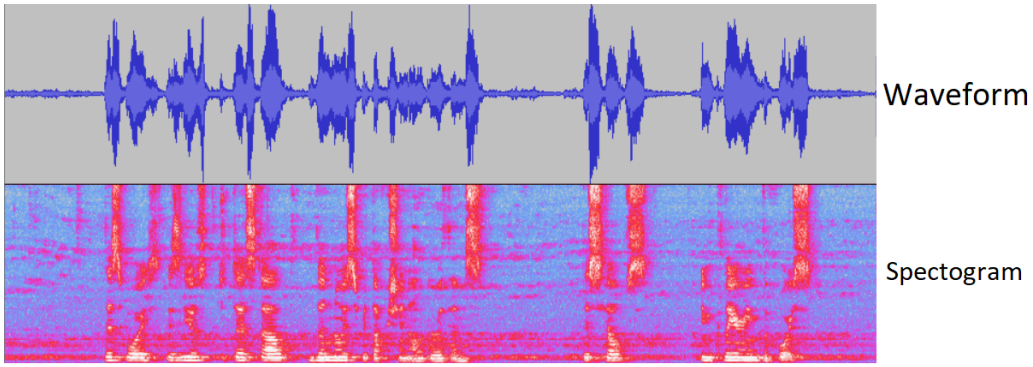
Figure 22: Door Controller Circuit Diagram.

5 Testing, Validation And Discussion

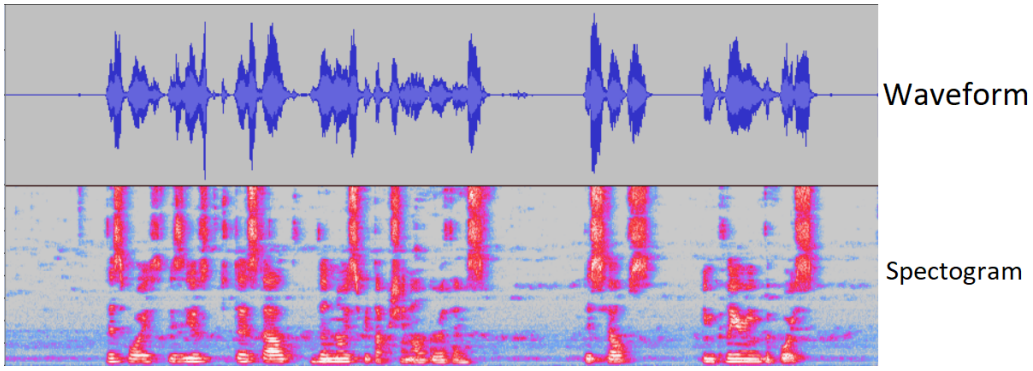
5.1 Challenges

5.1.1 Mic Sound Clearance

When the recording was using the internal laptop mic, the resulting sound had very low quality when compared to the external mic. This was not the main problem here; if we created a voiceprint using a low-quality mic, this would affect all the confidence levels of all the other voiceprints.



(a) Voice Signal With Noise.



(b) Voice Signal Without Noise.

Figure 23: Voice Signal With/Without Noise.

5.1.2 File Format Compatiblity

The system won't perform well with any voice file you provide, there are constraints on the sound files for the system to work, including the file extension, The encoding of the file, and the file's Sample rate, so recording the voices using any default recording software won't work.

5.2 Validation Methods

Since the project is concerned about an authentication method, we did multiple experiments trying to sabotage or gain unauthorized access from the system.

Let's define the confidence percentage as the insurance percentage of the voiceprint detection algorithm that there is a match for the given voiceprint in the database.

5.2.1 Voiceprint and Speech-To-Text Algorithms Unit Testing

Experiment #1, We registered 10 different people's voiceprints and saw what the confidence percentages would be like. We think that it performed very well in identifying every single person in average. We think that the more different voiceprints registered in the system, the more the positive acceptance rate would become.

Experiment #1 Results					
Voiceprint	#1	#2	#3	#4	#5
Confidence Percentages	89%	88%	93%	89%	90%

Experiment #1 Results					
Voiceprint	#6	#7	#8	#9	#10
Confidence Percentages	90%	92%	86%	89%	91%

Experiment #2, We tried to get a recording from an authorized person to be run in the authentication process, and at the same time, an unauthorized person would say the displayed words correctly, this trial would show whether the system would know that this access is invalid. And Indeed, the system didn't accept this authentication attempt. Our theory is that the input here contains two persons' voiceprints, so the output voiceprint is completely different.

Experiment #3, We tried to run the Speech-To-Text Algorithm in an environment which has noise sources including multiple persons talking around. The system failed in this experiment because the Speech-To-Text system returned much less accurate results.

Experiment #3 Results			
Environment	Quiet Environment	One Person Speaking	Four People Speaking
Speech-To-Text Average Accuracy	95%	82%	64%
Ability To Authenticate	Very Easy	Easy - Medium	Hard

This was solved when using a microphone with built-in noise reduction.

Experiment #4, We tried to generate a fake voiceprint using resemble.ai website for an already registered user, to test the voiceprint algorithm and how much the match would be between the real and the fake voiceprint. The result showed that the confidence percentage was close to the threshold but not exceeding it.

Experiment #4 Results			
Person	Shalodi	Skafi	Ateyyah
Real Voice Confidence Percentage	90%	89%	93%
Fake Voice Confidence Percentage	72%	77%	65%

The previous experiments are done seperately for each algorithm, we asked ten persons to record themselves while speaking for about 10 seconds, they recorded themselves using different microphones and different noise environments, at this stage the results for the voiceprint detection algorithm were very good and as expected, but later on, we did an integration testing for the whole system in a stable environment using high quality microphone, here the voiceprint detection algorithm showed very bad results compared with the results we got before in the unit testing for some reason.

5.2.2 Integration Testing

The hardware and software components are now working together, after doing testing for the functionalities for the system, it was all working fine and as expected, except for the voiceprint detection algorithm, we registered 5 users in the system using 'add new user' process, then started 'authentication process' for each user.

Experiment #5 Results					
Voiceprint	#1	#2	#3	#4	#5
Confidence Percentages	55%	69%	53%	72%	41%

To understand why the voiceprint detection algorithms gave bad results in experiment #5, we need to analyze all the experiments we did, since the voiceprint detection algorithm we used is using LPC as feature extraction algorithm, it seems the LPC is very susceptible to the recording environment and the noise in the audio signal, so in experiment #1, the algorithm succeeded in identifying each person successfully because the noise contained in the audio files is same for the registeration and verification files.

Another thing to notice, is that in the registeration and verification of voiceprints, the recorded files consists of spoken phrases and small silence intervals between them, so maybe this affecting the voiceprint detection algorithm accuracy, because the continuous speech causes to generate more accurate voiceprint features.

5.3 Results

The system is supposed to make sure that the door will not open for unauthenticated users, as well as to provide fast and reliable processing and response from the authentication system, but failed at that due to the weaknesses of the voiceprint detection algorithm.

The authentication system is performing and acting as a flexible master system and can be modified easily to be attached to any secondary slave system in the future.

The system have the ability to be implemented in a real door easily and without much change in the software.

The above experiments covered the main functionalities of the system, including the Speech-To-Text system and the Voiceprint authentication system.

Based on the results, we recommend implementing a voiceprint detection algorithm that uses MFCC as feature extractor because it showed better accuracy results in practical results [9].

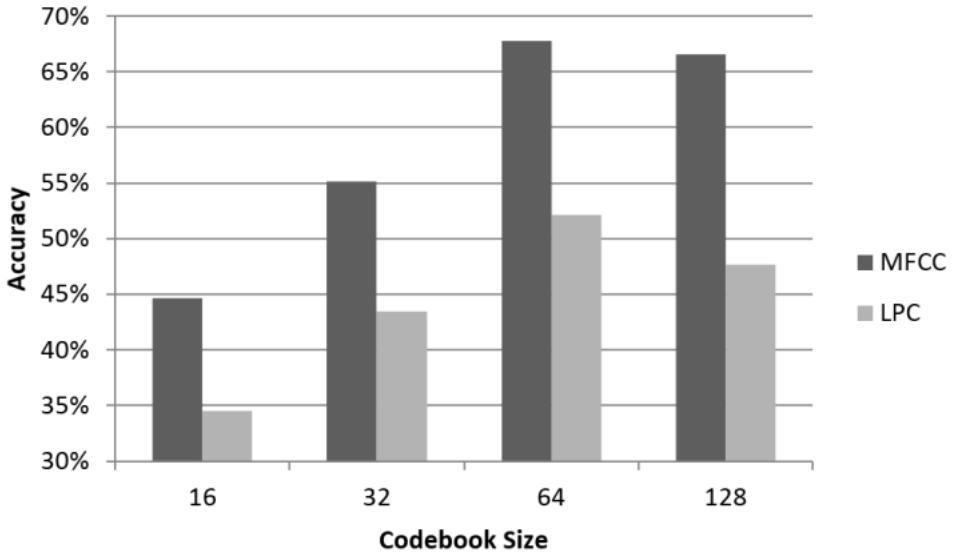


Figure 24: Accuracy Comparison Between MFCC and LPC [9].

On the other hand, we think that the system contains some drawbacks such as, in case there is no internet, the system will just stop working, so an offline secondary method of authentication is required to be implemented such as passwords.

The problem that causes not finalizing the project is that our focus was on implementing the software and hardware components in the system as a ready functioning blocks without much research about each component and how it works, so we think we selected the improper library for voiceprint detection. The very next step is to include another working voiceprint detection library that uses MFCC feature extractor due to it's higher accuracy compared to LPC.

6 Conclusion

6.1 Summary

A unique way of authentication is implemented by using human voiceprint and speech recognition. in order to control a physical door.

By using the speech recognition service provided by Google Cloud API for speech recognition, the feature of voice commands to open/close the door is implemented, also the Speech-To-Text system participates in the authentication process. And this happens in cooperation with the online server that works as the command-director for the overall system which receives it's commands from a master system and moves it towards the designated slave system.

As we can see, we can implement this authentication method for almost any other IoT device that requires authentication.

6.2 Future Work

First of all, configuring a solution for the voiceprint detection algorithm used to make it gives more accurate results, or to use another library that uses MFCC feature extractor. Also we recommend implementing a noise reduction algorithm to increase the overall accuracy of the system.

The second point would be the implementation of a secondary way of authentication like entering a password, which will work as a backdoor in case the system goes offline.

Finally, having the authentication system to be implemented completely online, and to be used on website logins, or any sensitive online activity.

References

[1] A. Cloud, "Voiceprint recognition system - not just a powerful authentication tool," Jul 2019, accessed: 2020-3-27. [Online]. Available: https://alibaba-cloud.medium.com/voiceprint-recognition-system-not-just-a-powerful-authentication-tool-6b3702b5c5a

[2] "Voiceprint-recognition-systems-for-remote-authentication-a-survey," April 2011, accessed: 2020-2-10. [Online]. Available: http://www.gvpress.com/journals/IJHIT/vol4_no2/6.pdf

[3] Y. Shi, Q. Huang, and T. Hain, "Robust speaker recognition using speech enhancement and attention model," *Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020.

[4] A. CHONÉ, "Computing mfccs voice recognition features on arm systems," Oct 2018, accessed: 2020-3-11. [Online]. Available: https://medium.com/linagoralabs/computing-mfccs-voice-recognition-features-on-arm-systems-dae45f016eb6

[5] V. Jain, "Accurate pole estimation by modified linear prediction," *ICASSP 84. IEEE International Conference on Acoustics, Speech, and Signal Processing.*

[6] "automatic speech recognition — google cloud," https://cloud.google.com/speech-to-text/, accessed: 2021-1-8.

[7] Admin, "Neural networks and speech recognition," Aug 2020. [Online]. Available: https://www.gosmar.eu/machinelearning/2020/05/25/neural-networks-and-speech-recognition

[8] R. Pi, "Raspberry pi 4 model b specifications," accessed: 2020-3-11. [Online]. Available: https://www.raspberrypi.org/products/raspberry-pi-4-model-b/specifications/

[9] R. M. Fauzi, "The recognition of hijaiyah letter pronunciation," https://www.shorturl.at/irNU7, accessed: 2020-4-21.

[10] A. B. n. BV, "Far and frr: Security level versus user convenience," https://www.recogtech.com/en/knowledge-base/security-level-versus-user-convenience, accessed: 2020-4-21.

[11] Y. Kumar and N. Singh, "A comprehensive view of automatic speech recognition system - a systematic literature review," *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, 2019, accessed: 2020-3-11.

[12] Muhamad, "What is speech to text software?" https://takenote.co/what-is-speech-to-text-software, accessed: 2020-4-21.

[13] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, 1975, accessed: 2020-3-11.

[14] N. P. H. Thian, C. Sanderson, and S. Bengio, "Spectral subband centroids as complementary features for speaker authentication," *Biometric Authentication Lecture Notes in Computer Science*, p. 631–639, 2004, accessed: 2020-3-11.

[15] Amaurycrickx, "Kitt-ai/snowboy," https://github.com/Kitt-AI/snowboy/, accessed: 2010-4-21.