# Design and implementation of speech recognition system integrated with internet of things

**Ademola Abdulkareem, Tobiloba E. Somefun, Oji K. Chinedum, Felix Agbetuyi**
Department of Electrical and Information Engineering, Covenant University, Canaan Land, Nigeria

**Article Info**

***Article history:***

Received Jul 14, 2020
Revised Aug 17, 2020
Accepted Nov 6, 2020

***Keywords:***

Digital signal processing
Internet of thing
Markov model
Programming interface

**ABSTRACT**

The process of speech recognition is such that a speech signal from a client or user is received by the system through a microphone, then the system analyses this signal and extracts useful information from the signal which is converted to text. This study focuses on the design and implementation of a speech recognition system integrated with internet of thing (IoT) to control electrical appliances and door with raspberry pi as a core element. To design the speech recognition system, digital signal processing (DSP) technique and hidden Markov model were fully considered for processing, extraction and high predictive accuracy of the system. The Google application programming interface (API) was used as a cloud server to store command and give the system to assess to the internet. With 150 speech samples on the system, a high level of accuracy of over 80% was obtained.

***Corresponding Author:***

Tobiloba Emmanuel Somefun
Department of Electrical and Information Engineering
Covenant University
Canaan Land, KM 10, Idiroko Rood, P. M. B. 1023, Ota, Ogun State, Nigeria
Email: tobi.shomefun@covenantuniversity.edu.ng

## 1. INTRODUCTION

Speaking is the major means of communication by a human. There are a lot of processes involved in the production of speech. Also, there are several body parts that aid in the production of speech, apart from the commonly known body parts such as tongue, mouth and lips. The lungs, trachea, larynx, vocal cord, oral cavity and nasal cavity are highly involved [1, 2]. Human speech is produced by the flow of air from the lungs through the larynx. It is produced by inhaling and exhaling through the nasal and oral cavity. Vowel sounds are produced by the flow of air from the lungs through the vocal cord, making them vibrate [3]. Consonants can be produced when the air is pressed through the closed vocal resulting in turbulent airflow. Due to the vibration of the vocal cords, sounds can be produced [4]. Each sound, word or speech vibrates differently. The frequency of the vibration is called pitch. In reference [5], the source-filter theory of speech production was introduced, which explains how speech is produced. According to [5], speech production is in two stages. In the first stage, air flows through the vocal cords to produce a basic signal. This basic signal is known as the signal source.

The recognition of the speaker is the process of recognising a speaker from the unique information, which is present in the wave of the word. This technique uses the speaker's voice to check the identity of the rapporteur and control access to services such as composition from voice, security, information service, remote access to a computer, purchases, etc. A lot of handicap (blind, lame) and aged persons in society have a limited capacity to perform certain tasks due to their physical and environmental conditions [6, 7]. Most often they require human help in several of their activities which usually cost a huge sum if the person is not their family member and persons who render such services are very minimal [8, 9]. This work seeks to help the physically challenged or disabled individual to perform the most basic tasks, such as opening doors, turning on/off electrical devices, calling a mobile line, automated activities and much more through the use of voice. It is like a telecommunication service that aids attention to the need of the disabled via automation [10, 11]. With the recent trend in automation as a means of control systems in different areas [12-16], this work deems it fit to integrate automation to meet some of the needs of the disabled individual. The proposed model in this study is limited to the sound or speech recognition mode of authentication. Although there is another authentication mode to gain access such RFID [17-19], biometrics [20-22], PIN [23, 24] and or a combination [25-27], this study focuses on the voice recognition.

## 2. MATERIALS AND METHODS

The bulk of this work hangs on the software part, although the hardware part is also important. Most of the hardware components used were ready-made; the technical aspect of the hardware lies on the correct ratings of all the components and right connection. In this work, the Raspberry Pi single board computer (SBC) software is deployed for the design of speech recognition system. The Raspberry Pi software is used to configure the hardware for the required action. As this work is based on internet of things (IoT), internet connectivity is required to be setup. In this work, internet connectivity is gotten through USB Wi-Fi adapter. For ease of identification, Figure 1 shows the block diagram of the proposed system.

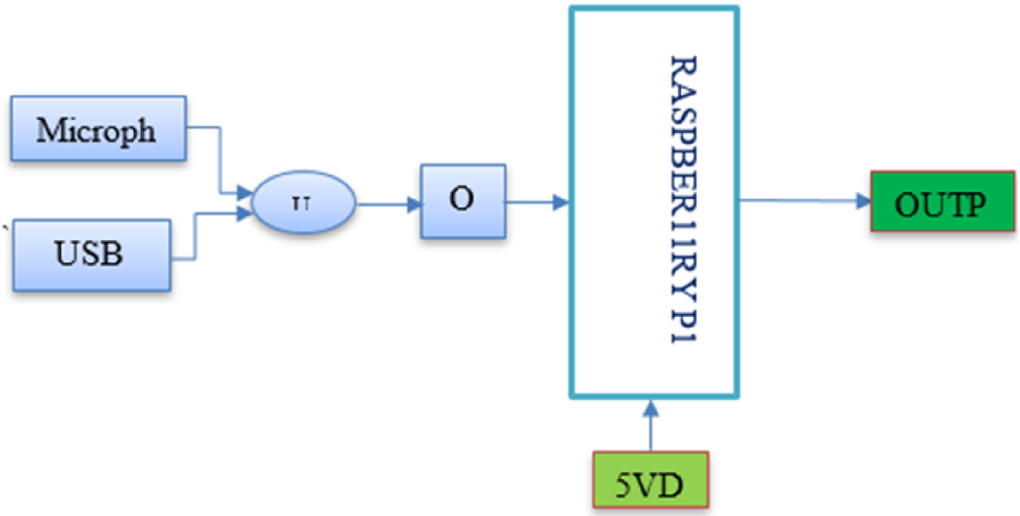

Figure 1. Block diagram for the speech recognition system

## 3. DESIGN SPECIFICATIONS

The design specifications of the speech recognition for access control module deals with the conditions necessary for the functionality of the module optimally. For this work, two types of design specifications would be considered namely: hardware and software specifications

### 3.1. Hardware specifications

The hardware specification deals with the optimal conditions necessary for the module to function. These conditions include:

a. Operating current: The current source needed to power up the raspberry pi, and all the peripheral devices attached to it should provide at least 2 A of current.
b. Operating voltage: The operating voltage by which the raspberry pi functions is 5 V. The current compensates to provide a required power of about 10 W. (5 V x 2 A = 10 W). It can be seen that this is a very low power device. Assuming a 7500 mAh battery is used to power up this device, the module can last for almost 4 hours before a battery recharge would be required.
c. Operating temperature: The official operating temperature range for the raspberry pi is from -40°C to 85°C. As the temperature begins to approach 82°C, the performance is thermally throttled.

### 3.2. Software specifications

The software specifications necessary for the module to run optimally include:

a. The RAM size: The RAM size required for this project is 512 MB and the raspberry pi zero meets this specification.

b. The ROM size: The minimum ROM size (storage space of the microSD card) is 8 GB. In this project, an 8 GB class 10 microSD card. This memory card has a very high read and write speed of 10 MB/s.

c. The processor: The processor that is necessary for running the software of this project work is the Broadcom BCM2835 system on chip (SoC) with an ARM 11 CPU running 1 GHz.

d. The operating system: The operating system, which the raspberry pi (the control unit of the module) runs is a Debian based distribution. It is mounted on an SD card, which takes at least 4 GB of space. The Debian distribution is known as Raspbian.

Figure 2 shows the circuit diagram of the proposed system and it is drawn from EasyEda application.

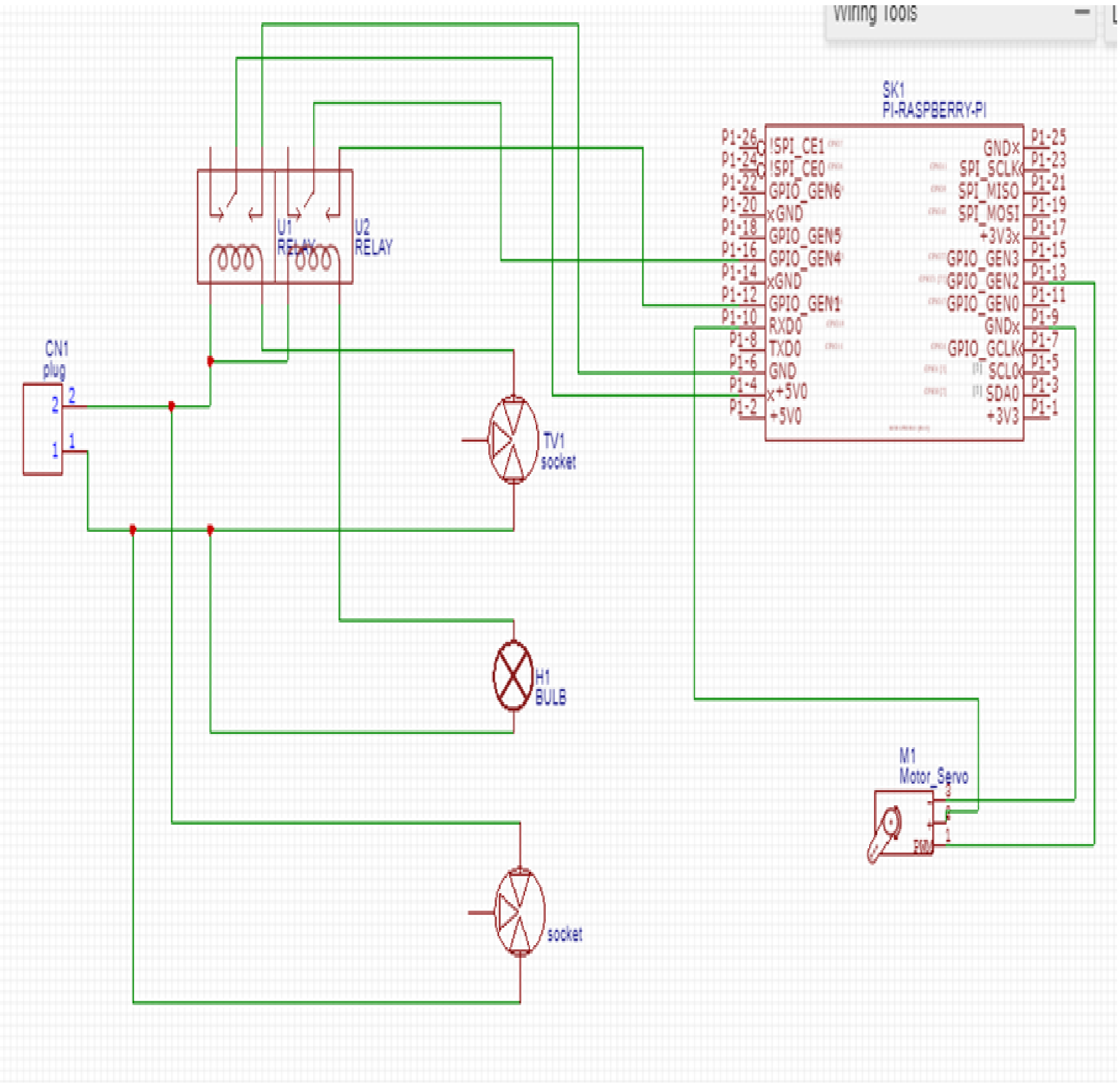

Figure 2. Circuit diagram

### 3.3. Design analysis

Let $x(t)$ represents speech signal which is a function of time. The time varying spectrum of the speech can be obtained from the time varying Fourier transform of the signal $Xn(\omega)$ as shown in Figure 3.

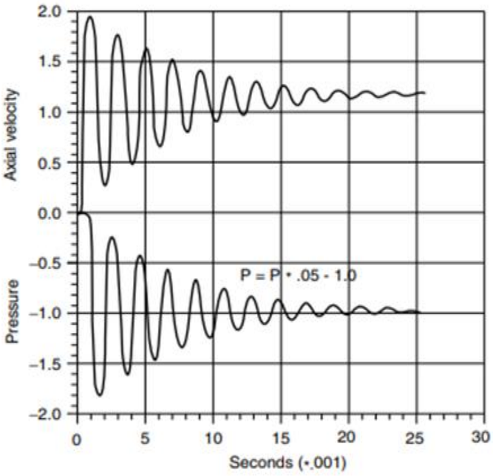

Figure 3. Time varying specturm of speech production

### 3.4. Fourier methods

The short-time Fourier transform of the speech signal $x(t)$ can be calculated using (1) as given in [28].

$$\text{Xn}(\omega)= \sum_{m=-\infty}^{\infty}(W_{n-m}X_m e^{-j\omega m}) \tag{1}$$

where index ‘n’ is referred to as time nT, which means that Xn(ω)=(nT, w). By inverse Fourier transform, $x(t)$ (speech signal) is recovered as shown in e (2):

$$X_n = \frac{1}{W_O 2\pi}\int_{-\pi}^{\pi} X_n(w)\, e^{j\omega n} dw \tag{2}$$

Since $Xn(\omega)$ is a function of time and it changes as time changes, it is sampled at a rate that allows the speech $x(t)$ to be reconstructed. With the bandwidth, Bx of the speech signal $x(t)$ being approximately equal to 5kHz, the sampling rate frequency (Fs) is therefore equals 10 kHz. For the Hamming window of length N=100, (wn), using (3), the bandwidth B is found to be

$$B = \frac{2F_S}{N} \tag{3}$$

$$B = \frac{20{,}000}{100} = 200Hz$$

The Nyquist rate for the short-time Fourier transform is twice the the bandwidth B, therefore, the Nyquist rate equals 400 Hz. Hence, at Fs equals 10,000, it requires a value of $Xn(\omega k)$ every 25 samples. Since N=100, the windows should overlap by 75% [28].

### 3.5. Gaussian mixture model (GMM)

For Gaussian mixture model, (4) gives a non-singular multivariate normal distribution of a d-dimensional random variable x [29]:

$$P(x) = \frac{1}{2\pi\frac{d}{2}\sqrt{|E|}} e^{(-\frac{1}{2}(x-\mu)^T(\Sigma -1)(x+1)} \tag{4}$$

Multivariate data are observation that are made on more than one variable, In (3) $P(x)$ is called probability density function formula, $\mu$ is the mean vector ($d \times 1\, matrix$) and $\sum is\ the\ covariance\ matrix$ [ $d \times d\, matrix$] of the normally distributed random variable X. In (4) the mean vector (expected vector) is as shown in (5):

$$\mu \triangleq F(x) \triangleq \int_{-\infty}^{\infty} xP(x)\, d(x) \tag{5}$$

The sample mean approximation as shown in (6)

$$\mu \approx \frac{1}{N}\sum_{i=1}^{N=1} X_i \qquad (6)$$

where the number of samples is $N$ and $X_i$ are the mel-cepstral feature vector.

The expression for variance-covariance matrix of a multi-dimensional random variance is described in (7):

$$\Sigma = \frac{1}{N-1}\sum_{i=1}^{N=1}(X_i - \mu)(X_i - \mu)^T = \frac{1}{N-1}[S_{xx} - N(\mu\mu)^T] \qquad (7)$$

where the sample mean $\mu$ is obtained from (5) and the second order sum matrix $S_{xx}$ as shown in (8) [28].

$$S_{xx} = \sum_{i=1}^{N-1} X_i X_i^{\ T} \qquad (8)$$

when the preparation information is prepared and the reason for an independent model that is saved as the previous statistics is assembled, many speakers are used to advance the Gaussian parameters and coefficients, using standard procedures, for example, maximum likelihood estimation (MLE), maximum posterior regulation (MPR) and maximum likelihood linear regression (MLLR). Now, the frame is ready to play the enlistment. Enrollment can be completed by taking an example of an objective voice sound and adjusting it so that it is ideal for adjusting this example. This guarantees that the probabilities returned when coordinating a similar example with the adapted model would be maximum.

### 3.6. Hidden Markov model (HMM)

Hidden Markov model is a model based on augmenting the Markov chain. A Markov chain is a non-linear model that is often used to represent a sequence of possible event such that the probability that even event would occur is dependent solely on the previous one. Figure 4 shows a Markov chain for assigning probability to a sequence of words, $w_{1,}w_{2\ldots\ldots\ldots}w_n$ i.e (the probability for a user to say “TO” after saying the word “GO” is 0.4) the speech recognition make decision based the most probably next state, which means from the Figure 3, the system is most likely to move to “BED” with “GO” as the initial state. The transition probability matrix is given as follows:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.5 & 0 & 0.5 \\ 0.6 & 0.2 & 0.2 \end{bmatrix}$$

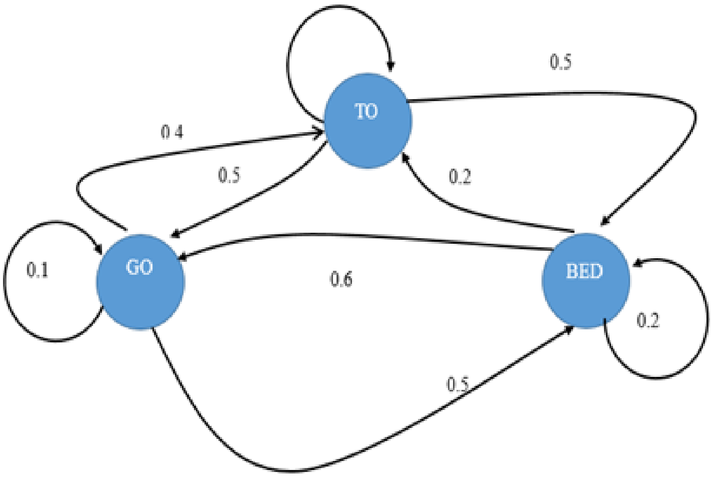

Figure 4. Markov chain [30]

The following components are required for a markov chain:

$*Q = q_1 q_2 \ldots \ldots . q_n$ a set of N state

$*A = \ a_{11,}a_{12,\dots\dots}a_{n1,\dots\dots}a_{nn,}$ a transition probability matrix

$$\sum_{j=1}^{n} a_{ij} \ = 1 \ \forall i$$

$*\pi \ = \ \pi_{1,}\pi_{2,}$ an initial probablity distrbution.

$$\sum_{i=1}^{n} \pi_{i} \ = 1 \ \forall i$$

The probability that a Markov chain will begin in state $i$ is $\pi i$. The flow chart for the procee is given in Figure 5.

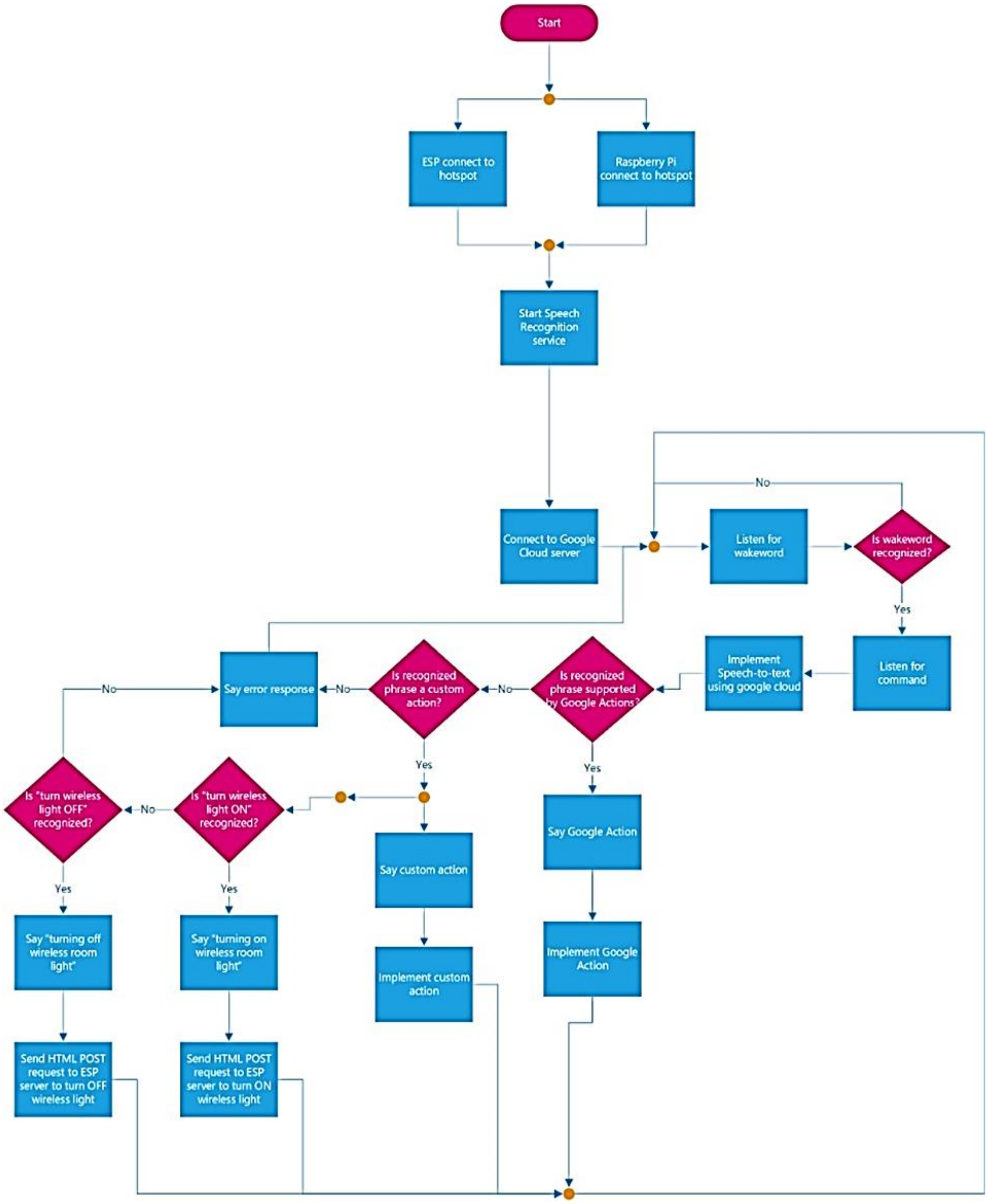
Start
ESP connect to hotspot
Raspberry Pi connect to hotspot
Start Speech Recognition service
Connect to Google Cloud server
No
Listen for wakeword
Is wakeword recognized?
Yes
Listen for command
Implement Speech-to-text using google cloud
Is recognized phrase supported by Google Actions?
No
Is recognized phrase a custom action?
No
Say error response
No
Yes
Yes
Say Google Action
Implement Google Action
Say custom action
Implement custom action
Is "turn wireless light ON" recognized?
No
Is "turn wireless light OFF" recognized?
Yes
Yes
Say "turning on wireless room light"
Say "turning off wireless room light"
Send HTML POST request to ESP server to turn ON wireless light
Send HTML POST request to ESP server to turn OFF wireless light

Figure 5. Flow chart

## 4. RESULTS AND DISCUSSIONS

The proposed system was tested severally to measure and ascertain its accuracy. The speaker recogniser was trained with various speech samples. Afterward, a candidate was enrolled for the system in order to confirm the accuracy of the system. The speech of the candidate was taken under different conditions to show the performance of the module with the test samples of the speaker's voice already in the database. The conditions under which the candidate voice was taken include:

- Crowded place with background noise;
- Silent place with little or no background noise;
- A condition such that the speaker's voice was low; and
- A condition such that the speaker's voice was loud.

By these four different conditions, recognition accuracy of the system is obtained as shown in Figure 6.

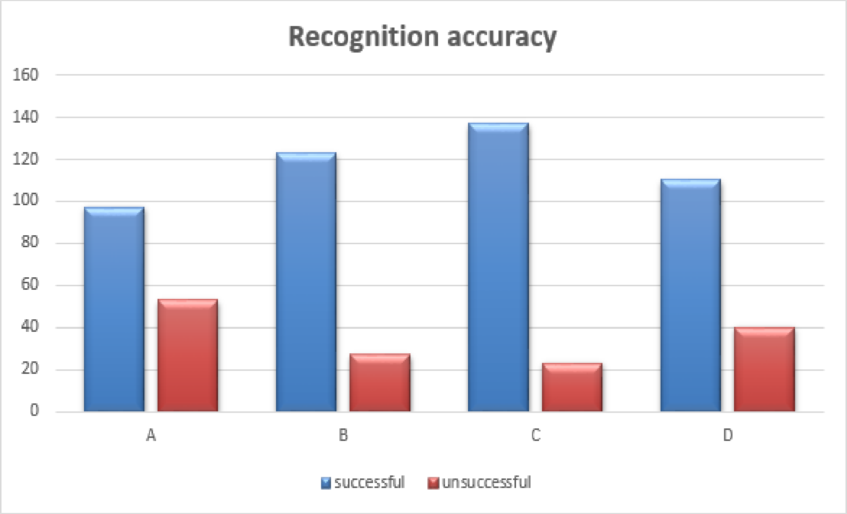

Figure 6. Result of tests carried out on same speaker under different conditions

where
A = A condition of crowded place with background noise;
B = A condition of silent place with little or no background noise;
C = A condition such that the speaker's voice was low; and
D = A condition such that the speaker's voice was loud

Table 1 shows the accuracy of samples taken.

Table 1. Accuracy of the sample taken

| S/No. | Test Condition | Number of Accuracy | Percentage (%) |
|---|---|---|---|
| A. | A condition of crowded place with background noise; | 97 | 64.67 |
| B. | A condition of silent place with little or no background noise; | 123 | 82.00 |
| C. | A condition such that the speaker's voice was low; and | 137 | 91.33 |
| D. | A condition such that the speaker's voice was loud. | 110 | 73.33 |

## 5. CONCLUSION AND RECOMMENDATION

This work has successfully constructed a prototype speech recognition system for home automation using Raspberry Pi Single Board Computer. The prototype worked well and gave a vary promising result for real production. With noise interference (i.e. worse scenario environmental condition in terms of noise), a very good result was obtained with approximately 65% accuracy and the highest accuracy recorded was 91%. The work can find application in different places and fields depending on the work to be carried out. The response time of this module is relatively fast. Further work can be done by adding more appliances, voice recognition module to ensure security for home automation and training of voice recognition module to adjust to diverse voice condition of the user. Moreso, other means of authentication can be added.

## ACKNOWLEDGEMENTS

The authors acknowledge Covenant University for her financial support.

## REFERENCES

[1] D. Blischak, et al., "Use of speech-generating devices: In support of natural speech," *Augmentative and alternative communication*, vol. 19, no. 1, pp. 29-35, 2003.
[2] M. Mills, "Aid for speech therapy and a method of making same," ed: *Google Patents*, 1984.
[3] K. N. Stevens and A. S. House, "An acoustical theory of vowel production and some of its implications," *Journal of Speech and Hearing Research*, vol. 4, pp. 303-320, 1961.
[4] K. Nishikawa, et al., "Speech planning of an anthropomorphic talking robot for consonant sounds production," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No. 02CH37292)*, Washington, DC, USA, vol. 2, 2002, pp. 1830-1835.
[5] G. Fant, "The source filter concept in voice production," *STL-QPSR*, vol. 1, pp. 21-37, 1981.
[6] A. Ismail, S. Abdlerazek, and I. M. El-Henawy, "Development of Smart Healthcare System Based on Speech Recognition Using Support Vector Machine and Dynamic Time Warping," *Sustainability*, vol. 12, no. 6, p. 2403, 2020.
[7] R. Gonzalez, et al,, "Voice Recognition System to Support Learning Platforms Oriented to People with Visual Disabilities," *in International Conference on Universal Access in Human-Computer Interaction*, 2016, pp. 65-72.
[8] T. Gomi and A. Griffith, "Developing intelligent wheelchairs for the handicapped," in Assistive Technology and Artificial Intelligence, *Springer*, pp. 150-178, 1998.
[9] R. C. Handel, "The role of the advocate in securing the handicapped child's right to an effective minimal education," *Ohio State University*, vol. 36, p. 349, 1975.
[10] T. E. Somefun, C. O. A. Awosope, and C. Sika, "Development of a research project repository," *TELKOMNIKA Telecommunication, Computing, Electronics and Control*, vol. 18, no. 1, pp. 156-165, 2020.
[11] A. Ademola, T. Somefun, A. Agbetuyi, and A. Olufayo, "Web based fingerprint roll call attendance management system," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, pp. 4364-4371, 2019.
[12] Y. Yamazaki and J. Maeda, "The SMART system: an integrated application of automation and information technology in production process," *Computers in Industry*, vol. 35, no. 1, pp. 87-99, 1998.
[13] L. Kocúrová, I. S. Balogh, and V. Andruch, "Solvent microextraction: a review of recent efforts at automation," *Microchemical Journal*, vol. 110, pp. 599-607, 2013.
[14] S. E. Shladover and C. Systematics, "Recent international activity in cooperative vehicle-highway automation systems," *United States. Federal Highway Administration. Office of Corporate Research*, pp. 1-95, 2012.
[15] C. von Altrock and J. Gebhardt, "Recent successful fuzzy logic applications in industrial automation," in *Proceedings of IEEE 5th International Fuzzy Systems*, New Orleans, LA, USA, vol. 3, pp. 1845-1851, 1996.
[16] L. Kamelia, S. A. Noorhassan, M. Sanjaya, and W. E. Mulyana, "Door-automation system using bluetooth-based android for mobile phone," *ARPN Journal of Engineering and Applied Sciences*, vol. 9, no. 10, pp. 1759-1762, 2014.
[17] A. Abdulkareem, I. U. Dike, and F. Olowononi, "Development of a radio frequency identification based attendance management application with a pictorial database framework," *International Journal of Research in Information Technology (IJRIT)*, vol. 2, no. 4, pp. 621-628, 2014.
[18] A. Juels, "RFID security and privacy: A research survey," *IEEE journal on selected areas in communications*, vol. 24, no. 2, pp. 381-394, 2006.
[19] A. Abdulkareem, C. Awosope, and A. Tope-Ojo, "Development and implementation of a miniature RFID system in a shopping mall environment," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 2, pp. 1374-1378, 2019.
[20] M. Lourde and D. Khosla, "Fingerprint Identification in Biometric SecuritySystems," *International Journal of Computer and Electrical Engineering*, vol. 2, no. 5, pp. 852-855, 2010.
[21] D. Bhattacharyya, R. Ranjan, F. Alisherov, and M. Choi, "Biometric authentication: A review," *International Journal of u-and e-Service, Science and Technology*, vol. 2, no. 3, pp. 13-28, 2009.
[22] N. L. Clarke, S. M. Furnell, and P. L. Reynolds, "Biometric authentication for mobile devices," *IEEE Security & Privacy*, vol. 13, pp. 70-73, 2015.
[23] T. Van Nguyen, N. Sae-Bae, and N. Memon, "DRAW-A-PIN: Authentication using finger-drawn PIN on touch devices," *computers & security*, vol. 66, pp. 115-128, 2017.
[24] J. Saville, "Authentication of PIN-Less Transactions," Google Patents, 2008.
[25] W. Shatford, "Biometric based authentication system with random generated PIN," Google Patents, 2006.
[26] F. Okumura, A. Kubota, Y. Hatori, K. Matsuo, M. Hashimoto, and A. Koike, "A study on biometric authentication based on arm sweep action with acceleration sensor," in *2006 International Symposium on Intelligent Signal Processing and Communications*, Tottori, 2006, pp. 219-222.
[27] Y. Li, J. Yang, M. Xie, D. Carlson, H. G. Jang, and J. Bian, "Comparison of PIN-and pattern-based behavioral biometric authentication on mobile devices," in *MILCOM 2015-2015 IEEE Military Communications Conference*, Tampa, FL, 2015, pp. 1317-1322.
[28] S. E. Levinson, "Mathematical models for speech technology," *Wiley Online Library*, 2005.
[29] S. K. Patel, J. M. Dhodiya, and D. C. Joshi, "Mathematical Model Based on Human Speech Recognition and Body Recognition," *International Journal of Engineering Research & Technology (IJERT)*, vol. 1, no. 4, pp. 1-5, 2012.
[30] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257-286, 1989.