researchvision.us

Smartify: Journal of Smart Education and Pedagogy
Vol.1 | No. 1 | 2024 pp. 28–46 |ISSN: XXXX-XXXX

Speech Recognition Model in Yoruba Language

Habeeb A. Rahmon,[1]

Tope G. Jimoh,[2]

Fatimoh O. Madaiyese,[3]

[1]Federal Polytechnic Ukana Department of Computer Science

[2]Federal Polytechnic Ukana Department of Computer Science

[3]Federal University LokojaDepartment of Computer Science

Submitted: 21-Jul, 2024
Accepted: 28-Aug, 2024
Published: 13-Sep, 2024

Vol. 1, **No.** 1, 2024.
10.62762/SJSEP.2024.000000

***Corresponding author:**
Niyi Jacob Ogunode
ogunodejacob@gmail.com

Copyright © 2024 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Speech has been widely known as the primary mode of communication among individuals and computers. The existence of technology brought about human-computer interfaces to allow human-computer interaction. While speech recognition systems have been developed rapidly over the last few years, many challenges have been encountered, such as language barrier, speech rate, environmental conditions, channel, and context of utterance. To enhance the power of speech recognition systems, it is required to design speech recognizers in local languages. There is a lot of development and research in the field of foreign languages but to enhance its power and utility for native people, it's essential to use this technology in native languages. In this paper, we employ acoustic modeling and language modeling for the task of recognizing speech and also converting the speech to text by constructing sequential MFCC features from the frames of the audio sample. The result shows that concatenating Mel-Frequency Cepstral Coefficients (MFCC) features sequentially and applying acoustic modeling and language modeling provide a solution to the problem of recognizing speech and also converting the speech to text with an accuracy of 83%.

Key words: Supervised Learning, Support Vector Machine (SVM), Unsupervised Learning, Utterances, Vocabularies, Speaker Dependence, Training, Accuracy.

Introduction

Language can be seen as a basic medium for communication, so it's important to expect human-computer interfaces in natural languages. Communication has played a major role in the development of human civilization. Different methods or styles of communication have been used throughout history to make understanding (each other) easier and more efficient (Ashok and Vikas, 2019). We also need to understand that there have been a lot of communication methods among humans since the start of time, verbal, non-verbal, and written forms are the major forms of communication.

Verbal communication is the basic and most efficient means of communication between humans but advancement in both society and technology has made even voice communication not just between humans but with machines (systems) a possibility (Akshata, Mahima, Laxmi, and Satish, 2020). More so, Speech is the most crucial, widespread, and proficient means of communication method for people to communicate with one another. Humans are comfortable with speech hence persons would also like to interact with computers through speech, rather than through primitive interfaces such as keyboards and pointing devices (Ayushi, Krina, Yesha, and Nirali, 2017).

Machine Learning has played a key role in the advancement of human civilization. Advancement in areas such as computing, engineering, robotics, medicine, and agriculture has made day-to-day activities easier and even more efficient but there is still a basic need in humans for perfection which has given rise to the need to provide even more efficient machines and systems. Machine learning (ML) and Automatic Speech Recognition(ASR) communities have had significant growth in each other (Li and Xiao, 2013).

Speech Recognition or Automatic Speech Recognition (ASR)is the ability of a machine or program to recognize words spoken verbally and successfully convert them to an understandable signal to provide a viable response or execute instructions (Neha and Somya, 2019).

Communication with machines (systems) verbally (through voice) has eliminated the need for manual and cumbersome operations of such systems which increases the efficiency and reduces complications involved in operating the said machine (system) as instructions can be given through voice commands without the knowledge of technical know-how of the said machine (system). Current application areas include web search, intelligent personal assistants, home automation, automotive, healthcare, financial transactions, and consumer electronics, just to name a few. The growing interest in speech recognition is also witnessed by the remarkable investments of the most important tech companies, that have recently committed huge resources in the field (Adekola et al, 2019).

While speech recognition models have been developed at a rapid pace over the last few years a lot of challenges have also been encountered during this development such as language barrier, speech rate, environmental conditions, channel, and context of utterance. To enhance the power of speech recognition systems, it is required to design speech recognizers in local languages. Multilingualism is a new evolving field in the area of speech recognition. There is a lot of development and research in the field of foreign languages but to enhance itspower and utility for native people, it's essential to use this technology in native languages (Ashok and Vikas, 2019). This study aims to develop a model that will convert speech in English to

the equivalent text in the Yoruba Language.

Literature Review

Overview of Speech Recognition

The early computer systems were limited in scope and power. However, the revolution in computer technology has evolved the field of automatic speech recognition. Currently, it's easy to store huge databases for speech recognition as a result of advancements in computer technology (Ashok and Vikas, 2019). More so, in the earlier part of the previous century, there was curiosity in making computers do what only humans could recognize, such as recognizing speech, understanding natural language, processing images, etc. Speech being the primary, most significant mode of communication between human beings, research work in speech recognition has received much enthusiasm for the past five decades right from the advent of artificial intelligence (Jayashree and Melvin, 2015). Over the years, sophisticated skills were developed to recognize patterns such as speech, handwriting, facial features, etc. The pursuit of computer programs that make computers learn the above skills from experience gave birth to machine learning. Depending on how the machine gains knowledge to respond correctly, learning can be categorized into four basic methods. This includes Supervised learning, Unsupervised Learning, Semi-Supervised Learning, and Active Learning.

Deep learning algorithms have been mostly used to further enhance the capabilities of computers so that they understand what humans can do, which includes speech recognition. Speech in particular, being the main method of communication among human beings, received much interest for the past five decades, right from the introduction of Artificial Intelligence (AI). Therefore, it is only natural that one of the early applications of deep learning was speech, and up to this day, a huge number of research papers have been published on the use of deep learning for speech-related applications specifically speech recognition. (Ali, Ismail, Imtinan, Mohammad, and Khaled, 2019).

Review of Relevant Approaches to Speech Recognition

There are different approaches or techniques to speech recognition. Some of these different approaches or techniques used will be review below. According to (Praphulla, Ratndeep, and Pukhraj, 2015) The speaker recognition system may be viewed as working in four stages i.e. Speech analysis, Feature extraction, Modeling, and Testing.

Machine Learning Approach

Machine learning is defined as the field of study that provides computers with the ability to learn from input data without being explicitly programmed to do so. The learning process is done iteratively from analyzed data and new input data. This iterative aspect allows computers to identify hidden insights and repeated patterns and use these findings to adapt when exposed to new data (Ali, Ismail, Imtinan, Mohammad, and Khaled, 2019). This

approach comprises three main techniques which are: Supervised learning, Unsupervised learning, and Semi-supervised.

Supervised learning includes training datasets at which data is presented in pairs, an input, and its corresponding correct output. The training of the model is based on a set of data called the training set after which it will be tested on a set of data called testing set. Some of the Supervised learning algorithms include Linear regression, Naïve Bayes, Support Vector Machine, etc.

Unlike Supervised learning, Unsupervised learning uses an input dataset without any labeled outputs to train the learning algorithm. There is no right or wrong output to each input object and no human intervention to correct or adjust as in supervised learning. Thus, Unsupervised learning is more subjective than Supervised. The main goal of unsupervised learning is to learn more about the data by identifying the fundamental structure or distribution patterns that are found in the data itself.Examples of Unsupervised learning algorithms are Kmeans which is suitable for Clustering problems and Apriori Algorithm which is also suitable for association problems. Lastly, is the Semi-supervised learning This method falls between the supervised and unsupervised learning methods where we have a large amount of input data, some of which are labeled and the rest are not. Many real-life learning problems fall under this area of machine learning. The reason for that is that semi-supervised requires less human intervention since it utilizes a very small amount of labeled data and a large amount of unlabeled data. Utilizing less labeled datasets is more appealing since such datasets are very hard to collect as well as expensive and may require access to domain experts. Unlabeled datasets on the other hand are cheaper and easier to get access to.

Both supervised and unsupervised learning techniques can be utilized to train the learning algorithm in semi-supervised learning. Unsupervised learning techniques canbe used to unfold hidden structures and patterns in the input dataset. Whereas supervised learning techniques can be utilized to make guess predictions on the unlabeled data, feed the data back to the learning algorithm as training data, and use gained knowledge to make predictions on new sets of data.

The Artificial Intelligence Approach (Knowledge-Based Approach)

It is the combination of the acoustic-phonetic approach and pattern recognition approach. The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies its brilliance in visualizing and analyzing. And then finally deciding on the measured acoustic features. The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies their intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features. A highly master system is used widely in this approach. In Knowledge-based approaches: Expert knowledge about variations in speech is hand-coded into a system. It is

beneficial for explicit modeling variations in speech, but unfortunately, such expert knowledge is difficult to obtain and use successfully. Thus this approach was judged to be impractical and an automatic learning procedure was sought instead (Praphulla, Ratndeep, and Pukhraj, 2015).

Connectionist Approaches (Artificial Neural Networks)

The artificial intelligence approach attempts to mechanize the recognition procedure according to the way a person applies their intelligence in visualizing, analyzing, and finally making a decision on the measured acoustic features. Among the techniques used within this class of methods are the uses of an expert system that integrates lexical, phonemic, semantic, syntactic, and pragmatic knowledge for segmentation and labeling. It also uses tools such as artificial NEURAL NETWORKS for learning the relationships among phonetic events. It mainly focuses on the representation of knowledge and the integration of knowledge sources. This method has not been widely used in commercial systems (Praphulla, Ratndeep, and Pukhraj, 2015).

Support Vector Machine (SVM)

One of the powerful tools for pattern recognition is SVM which uses a discriminative approach is a SVM. It uses linear and nonlinear separating hyperplanes for data classification. Despite this, SVMs can only classify fixed-length data vectors. This procedure may not be readily applied to tasks involving variable-length data classification. This data has to be transformed into fixed-length vectors before SVMs can be used. It is a generalized linear classifier with maximum margin fitting functions. This function gives regularization which helps the classifier generalize better. Traditional statistical and Neural Network methods control model complexity by using a small number of features. SVM curbs the model complexity by controlling the VC dimensions of its model. This method is independent of dimensionality and can employ spaces of very large dimensions' spaces, which grants the construction of a very large number of nonlinear features and then performing robust feature selection during training (Praphulla, Ratndeep, and Pukhraj, 2015).

Review of Relevant Literature to Speech Recognition

The main task of literary review is to get acquainted with other studies on the topic of the project. Many researchers have observed several times how they can prevent thisproblem, how they achieved their goals, and how they did the related work. A review was carried out on the concept overview, speech recognition, overview of related works, and summary of the review.

Abhishek, Balamurugan, Gokul, and Anitha (2019) proposed a system for a robot that can perform five basic commands go left, go right, go forward, go back, detecting humidity and temperature along with detecting objects in real time to interact with the user. The proposed system was an integration of four different modules namely: voice recognition integration, robot movement and sensor usage, object detection phase, and interactive virtual assistant. The voice recognition module focused on integrating the Google Assistant as a trigger to actions on the IFTTT (if This Then That) platform through the Adafruit cloud server. The

server consists of data from the device which is displayed to the user.

James et al (2018) presented the design and development of an Artificial Intelligence Voice Assistant for a University campus that can interact with the user and answer questions related to its domain of knowledge. The front end of the application is a simple, lightweight, and straightforward mobile application that is designed for ease of use for all age groups and requires minimal technical knowledge. The backend can be divided into three modules: the app module which is responsible for taking the user input, converting it from speech to text, and sending the query to Dialogflow, the Dialogflow module which transforms the input into a query and matches it with the most appropriate intent and the fulfillment module which stores the information about various entities related to the university that the user could ask the bot. The data is stored in the form of tables that can be quickly updated and managed by the admin. This module also contains the code for the webhook that is responsible for querying the database based on the user's question.

Adekola et al. (2019) proposed a system that uses voice recognition to allow people with disabilities to access door systems and also to provide security to lives and properties. Some of the methods and techniques used in this system are the use of voice recognition techniques this phase deals with the verification i.e. a sample of speech utterance is compared against a model already stored in the system to determine the best match. The use of fingerprint scans for biometric identification is based on individual registered fingerprints. The use facial recognition is also used for biometric identification but basically for analyzing patterns based on the person's facial textures or shape that has been registered. The use of electronic components like Arduino that is used to manufacture single-board microcontrollers and microcontroller kits for building digital devices and interactive platforms. Application of the acoustic-phonetic approach to investigate time domain features such as the mean squared amplitude of a waveform to abstract linguistic concepts such as phonemes, phrases, or utterances and application of support vector machine to analyze data used for classification and regression analysis.

Ratnadeep and Abdulmali (2018) reviewed the various speech recognition techniques and discussed the various speech recognition typologies like speaker-dependent, speaker-independent, isolated word recognizers, connected word systems, and spontaneous recognition systems. The various available speech corpora were presented in a detailed tabular format. They discussed the preprocessing stage which deals with removing noise from the data and the feature extraction stage in which techniques like Mel Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC), and Linear Predictive Cepstral Coefficient were discussed in depth. Next were theclassification techniques ranging from the Hidden Markov Method (HMM), Dynamic Time Warping (DTW), Artificial Neural Network, and Support Vector Machines.

Ashok and Vikas (2019) discussed speech recognition or Automatic Speech Recognition (ASR) in depth,describing the building blocks which are Signal Preprocessing, Feature

Extraction, Language Model, Decoder, and Speech Recognition. The signal preprocessing stage is simply turning the speech signals captured by the microphone which are analog into a digital form, the feature extraction stage is simply the parameterization of speech signals and some of the common feature extraction techniques like Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP) was discussed.

Yash, Rohan, Pushkar, Umar, and Ankita (2019) proposed a system for a voice-controlled personal assistant bot with smart storage. First of all, the system starts with respective wake words for the voice assistant. Following the wake word, the system proceeds by receiving voice commands from the users with the help of a microphone, further analyzing the command and processing the given, the device will create an appropriate response for the given command. The response is given to the user with the help of speakers. Users can also use the device for storing and streaming the data. Users can store excess data (Data taking up extra space on the local device) in the wireless storage system of the device with the help of an app on the host's device. This application interacts with the system. Uploading downloading and streaming of data from the system to the host device is now convenient. The application can also host various users simultaneously. The system implemented all the various Natural Language Processing (NLP) algorithms simultaneously with the server protocols.

Anil and Munikrishnaprasad (2020) proposed a voice-based message-scheduling Android application. The paper also discussed techniques for message scheduling and automatic speech-speech summarization based on speech unit extraction and concatenation, a two-stage summarization method consisting of important sentence extraction and word-based sentence compaction is investigated.

Prof. Frioz, Shrigowri, Swetha, Syeda, and Vinutha (2018)proposed a system for an artificial intelligence-based travel bot. It describes the implementation of a voice bot that a human can interact with by speaking to it and also receive responses from the voice bot using its speech synthesizer. It can take userinput in both speech and text, for this purpose many open source platforms were made available. Program-o which is an open source AIML engine written in PHP was used. It is an interpreter for AIML scripts for the voice bots, it uses MySQL database to store voice bot details. When a user sends a message to the voice bot, then according to the matched reply from the AIML, the answer is formulated and sent back to the user.

Chandu, Pooja, Nivedita, and Nikitha (2018) proposed a system for a chatbot to replace a helpdesk. They explained in clear detail the steps, techniques, and approaches required to create a chatbot. From using Artificial Intelligence Markup Language (AIML) which is an XML-complaint language, connecting with Google Dialogflow, and also usingthe Watson Assistant which combines several cognitive techniques to help build and train a bot-defining intents and entities and crafting dialog. George and Maya (2020) discussed the history, evolvement, and impact of voice assistants and smart speakersin everyday life. An in-depth

analysis of the techniques used from the inception of the idea to the more recent and efficient ones. It aimed to find if a man has made great strides in technology with the advent of speech recognition and how its presence or absence might have impacted society as a whole.

Sourabh, Gayatri, Drishti, and Jay (2020) proposed designs for a Human Voice Controlled Robot Embedded with Real-Time Obstacle Detection and Avoidance. Using different hardware parts like Arduino Uno (which is an open-source microcontroller), Bluetooth module HC-05 which is used to establish wireless connection, L298D Bridge which is a high voltage and high current dual driver that is designed to accept standard Transistor-Transistor logic, HC-SR04 Ultrasonic Sensor which is a distance sensor commonly used with microcontroller platform and software like the Arduino IDE which contains a text editor for writing code, a message area, a text console, a toolbar with buttons containing common functions and a series of menus. The programs are then uploaded to the Arduino hardware and the Android application that lets you connect to the HC-05 module via Bluetooth and allows you to send signals to the Arduino so that it can perform desired operations successfully.

Iannizzotto, Bello, Nucita, and Grasso (2018) proposed a system for a vision and speech-enabled virtual assistant for smart environments. The proposed architecture is fully modular. It is composed of a set of services, a graphical frontend, and a coordinator that leverages the services to offer the user a multimodal and involving interaction with the connected home automation and smart assistant systems. In the proposed architecture, the associated class currently contains three different modules, called Flite2 module, the MaryTTS module, and the GoogleTTS module. The MaryTTS module produces very natural utterances and supports multiple languages and voices. However, it has quite a large footprint (up to 1GB RAM), thus it is not suitable when other memory-intensive modules are involved. The Flite2 module produces natural, clear, and understandable utterances, it is lightweight in terms of memory footprint and computational needs, but currently supports just a few languages and offers few voices (mostly male and US English or Indic speaking). The GoogleTTS module relies on the well-known Google Cloud Speech API. Despite depending on cloud services and requiring a paid subscription, Google Cloud Speech API offers the benefits of supporting multiple languages and featuring both male and female, very natural, voices.

Ravikumar, Prateek, Sathvik, Rahul, and Mayura (2020) proposed a design for a virtual voice assistant. The device that uses the human voice for interacting with the device uses single commands as inputfor the device they usually consist of single phrases. When commands become ambiguous, the resulting actions can be misunderstood by the device. There is only one-way communication between the user and the device because the device cannot talk back for clarification. Voice-controlled Device uses Natural Language Processing to process the language spoken by humans understand the query process the query and respond to the human with the result. The understanding of the device means artificial intelligence needs to be integrated with the device so that the device can work smartly and can also control IoT applications and devices and can also respond to queries which will search the web for results

and process it. It is designed to minimize human efforts and control the device with just a human voice.

Akshata, Mahima, Laxmi, and Prof. Satish (2020) proposed a system for Interactive Speech Recognition using AI. For intelligent voice assistant application is done using a Server. This process involves the conversion of acoustic speech into a set of words and is performed by a software component. Accuracy of speech recognition systems differ in vocabulary size and confusability, modality of speech (isolated, discontinuous, or continuous speech, read or spontaneous speech), and task and language constraints. The system consists of five modules: feature extraction, phone model training, dictionary preparation, grammar estimation, and sentence decoding. They proposed a method based on DNN, which extracts informative features from the speech signal, and those features are then used by the Acoustic model for speech recognition.

Tae-Kook Kim (2020) gave a detailed design of a voice control system using embedded systems and open API AIs. A voice control system using Google Assistant, a representative service of open API artificial intelligence, and the conditional auto-run system, IF This, Then That (IFTTT) was designed. Voice recognition API which is an HTTP-based REST API interface that supports input and output based on the JavaScript Object Notation (JSON) format was used. To use the voice recognition API, audio files recorded at a sampling frequency of 16 kHz must be encoded into Base64 and delivered to the open API server via HTTP communication. Users use the Google Assistant service to send voice commands. Voice data sent to the Google server are converted to text after being recognized; the data are subsequently sent to the conditional auto-run mode IFTTT service. For IFTTT, one recipe performs one operation. The system allows users to implement their system through the advantageous voice recognition interface and various modules.

Debajyoti, Chonlameth, Suree, and Vijayakumar (2019) surveyed user experience with smart voice assistants. They attempted to investigate into the usability details, and the satisfaction obtained after using the VAs for people who speak English as their secondary language and compare them with native English speakers. In the usability study, they could not find any significant difference between the two groups of users. This indicates the positive outcomes of many previous researches on the voice recognition of the VA systems. The VAs are intelligent enough to understand and respond to the English instructions are being spoken in different accents. This is evident from the VAs allow them to be set up initially in multiple demographic variants from English (Australia), and English (UK) to a variety of other local flavors. Although none of the VAs (Siri and Alexa) had English as their built-in language, yet for the majority of our participants it was irrelevant. These are very positive results that imply the maturity of the VAs in understanding English having multiple accents.

Pradhyum, Shubhangi, and Manisha (2020) proposed a system for a Personal Digital Assistant. They considered different methodologies such as Speech Text, speech recognition, text-to-speech, etc. For example, weather.com is used to get the Weather forecast information by

using Python Module 'spyware'. Different modules with different functions were also created namely: Speech to Text module which news is fed to the system in the form of text which is then read out by the system, the Speech Recognition module which takes speech and is recognized as input and as output executes the said commands, Text to Speech module which takes text and input and converts it to speech either in male or female voices.

Polyakov et al (2018) presented an investigation and development of an Intelligent Voice Assistant for the Internet of Things Using Machine Learning. They discussed the steps and techniques involved in creating a voice assistant and also gave detailed information on already existing voice assistants like ALICE. To create their voice assistant PocketSphinx was used. PocketSphinx is a tool for automatic voice recognition, which works well on various low-power embedded systems, such as Raspberry Pi, and it is also cross-platform, which is the reason this framework was chosen. To generate the voice, the most used Festival engine was chosen, which runs on Linux operating systems and has quite good characteristics of voice generation. The next stage, on which the intellectual work of the voice assistant is carried out, is the creation of a system for recognizing the natural language of a person, i.e. recognition of intentions. For the system that can understand the user, they used training intellectual algorithms based on machine learning methods. But before you teach the system to understand the user, you need to perform several stages of data preparation.

Since in this case, the output characteristics can be a set of different classes (that is, n-best hypotheses of the user's intentions, the task will be a multi-class classification task where one label can contain labels from different classes). For example, a response can return "on; light; bathroom", where each part of the answer belongs to one of several classes. Collecting the necessary data is the most important process that will allow you to most accurately predict the results. In this work, they created a specialized system for managing a smart home. Considering this, the given examples of data sets will correspond to the training of the model for such systems. For any other tasks, the principle will remain unchanged, only the data sets will change.

Shiny, Rahul, Thushar, Ramya, and Varsha (2020) presented a proposal for a voice-controlled robot for guiding tourists or visitors in historical places and complex architectural buildings. The guide robot works with the help of Wi-Fi and guides the visitors with voice guidance. To do this design was divided into two parts: interfacing with the Wi-Fi module on the robot via hardware, and interacting with the Wi-Fi module on the robot via software. The software mainly used in this project is Google Assistant, IFTTT, and Blynx. IFTTT and Blynk are used to connect Google Assistant to Nodemcu. Nodemcu will be programmed in such a way that once the action occurs the respective action takes place in the microcontroller unit. Once it receives a signal from the blink via Wi-Fi which is inbuilt on the Nodemcu. To output the command once the action is completed by the Nodemcu we establish a connection between Arduino and Nodemcu. The Nodemcu triggers the respective part of the Arduino program and it is programmed in such a way that once the respective command comes to the Arduino the respective audio file is played as user guidance audio.

Subhash, Ullas, Prajwal, Siddesh, and Santhosh (2020) proposed a system for an Artificial Intelligence-based Voice Assistant. In this project, they installed the GTTS (Google Text to Speech) engine package to make the assistant speak like a normal human being.

Linda, Nilesh, and Rajat (2021) proposed a system for a Voice Control Human Assistant Robot to supply help to associates in nursing aged or disabled persons. This Robot was automated using machine learning algorithms to do things in particular areas like home. The Raspberry Pi module is the brain of the robot and the movement of the robot can be controlled by using the voice to do things like a human which can be used in many fields, especially for blind people. (Shalini, Jayasuruthi, and VinothKumar, 2018).

In summary, the reviews show that A lot of work has been put into place to ensure that the voice recognitionsystem can recognize the owner's voice and grant access to the owner. There is still a lot of work that needs to be done to ensure that the voice recognition system can perform well and optimally. Some key areas that should be considered when making improvements to the voice recognition system include Native language: Major Nigerian native languages such as Igbo, Yoruba, and Hausa currently do not have a speech-to-text engine. The ability of voice recognition to recognize the native language of the user can be worked and improved upon by training speech models in that language. Security: The exchange of data between the various components of the hardware and the software should be encrypted to ensure that the system cannot be hacked and data integrity is maintained. The recommendation is further that a custom encryption algorithm should be implemented. The summary of findings from the reviewed research works is shown in Table 2.1 below.

Firoz et al. (2018)	Artificial Intelligence-based Travel Bot	AIML engine written in PHP was used and also it uses MySQL database to store voice bot details.	The bot can only answer questions that it has the answers present in its database.
Akshata et al. (2020)	Interactive Speech Recognition Agent System using AI	Deep Neural Network	The system is too simple and cannot execute complex tasks.
Sourabhet al. (2020)	Human Voice Controlled Robot Embedded with Real-Time Obstacle Detection and Avoidance	Acoustic modeling, language modeling and Arduino UNO	For advancements in this research work, we can use a long-ranged module which will result in connectivity with the robot for long distances and as a result, overcome its limitation.
Pradhyum et al. (2020)	Personal Digital Assistant: A Case Study	Tokenized algorithm, Google voice API and SAPI5	Home Automation using microcontrollers can be included in the project.

Ravikumar et al. (2020)	Virtual voice Assistant	Natural Processing, Intelligence Language Artificia	The device uses single commands as input. When commands become ambiguous, the resulting actions can be misunderstood by the devices.
Linda et al.(2021)	Voice Control Human Assistance Robot	Machine Learning Algorithms, Raspberry Pi	The Robot can be improved by embedding this system into the wheelchair which will make an autonomous wheelchair

Table 2.1: Summary of findings from reviewed research works.

Author's Name and year	Title of work Reviewed	Methods	Remarks
Adekola et al. (2019)	Voice Recognition Door Access Control System	The use of electronic components like Arduino, Application of Acoustic phonetics, and Application of Support.	Improvements should be made to Native language and Security
Shalini et al. (2018)	Voice Recognition Robot Control Using Android Device	Arduino UNO, Bluetooth, L298N, and DC Motor for the hardware design. AndroidSDK, Eclipse IDE, and Android Virtual Devices (AVD) for the Software Design.	This system is limited to a set of commands i.e. it cannot recognize lots of words.
Ashok and Vikas (2019)	Speech Recognition: A Complete Perspective	Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficient(MFCC), Perceptual Linear Prediction (PLP).	Several parameters that need to be worked on are speech rate, environmental conditions, channel, and context of utterance.

Materials and Methods

This section of the study describes the method adopted to achieve the stated objectives of the proposed model. The proposed methodology provides complete measures for the speech recognition model (speech-to-text) which follows the Knowledge Discovery Database (KDD) process (see Figure 3.3).

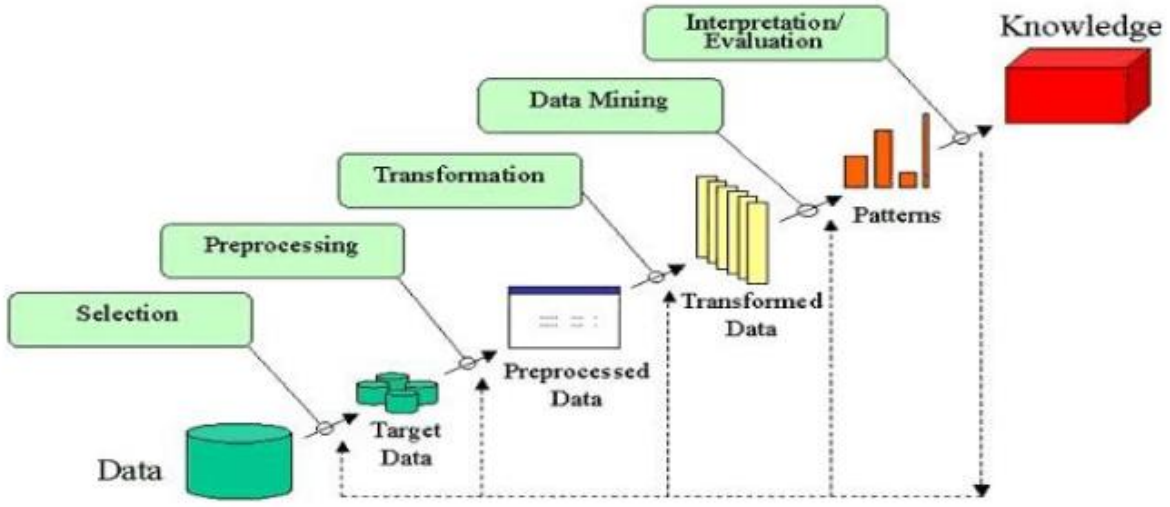


Figure 3.1: Knowledge Discovery in Database (KDD) Methodology

Data Selection: This aspect is concerned with the selection of data from available data sources. The data we make use of are in audio format and are gotten from Kaggle.

Preprocessing: This part has to do with the preparation of data in the form that will be fit for advanced evaluation and processing.

The first preprocessing step on the dataset is the conversion of audio samples to a .wav file. The second preprocessing step is the removal of empty spaces and noise reduction from audio samples, this was achieved using Audacity. This part has to do with the preparation of data in the form that will be fit for advanced evaluation and processing.

Feature Extraction: 20 Mel-frequency Frequency Cepstral Coefficient(MFCC) features are calculated for each audio file in the dataset. The MFCC is a method of categorizing the frequencies of a sound in a way that distinguishes phonemes.

Feature extraction is implemented using the Python library 'Librosa'. The following steps are used to calculate the MFCC cepstral features:

- Framing each signal into short frames of equal length with frame length as 2048 samples and hop length as 512 samples.
- Calculating the periodogram estimate of the power spectrum for each frame.
- Applying the Mel filter bank to the power spectra and summing the energy in each filter. The formula for converting from frequency to Mel scale is:

$$M(f) = 1125 \ln\left(1 + \left(\frac{f}{700}\right)\right)$$

Where m is the Mel filterbank, 1125 and 700 areconstant and f is the frequency iv.Taking the logarithm of all filterbank energies.

- Taking the Discrete Cosine Transform of the log filterbank energies.
- Selecting the required MFCC coefficients. In this case, 20 coefficients are selected.
- These coefficients are used because they approximate the human auditory system's response closely. The coefficients of each frame of an audio file are concatenated to form an array of MFCCs. The MFCC features extracted from an audio sample are outputted inthe form of a matrix with 20 coefficients for each frame of the sample, i.e.

$$MFCC = \begin{vmatrix} c0f0 & \dots & c0fm \\ \dots & \dots & \dots \\ c19f0 & \dots & c19fm \end{vmatrix}$$

where ci f0...ci fm are the values of coefficient i for frames 1...m the above matrix represents the MFCC coefficients for an audio sample with m frames. This 20 x m matrix needs to be transformed into a format that is recognized by the machine learning model. To contrast the three accents based on their feature sets, MFCC coefficients are sequentially concatenated. This retains enough information required to identify the accent from the feature set. The features are concatenated and flattened into a one-dimensional array of features for each frame. A sample vector of sequential MFCC features from a single audio sample is shown below.

$$M = [c_0 f_0 \dots c_0 f_m \dots \dots \dots c_{19} f_0 \dots c_{19} f_m]$$

Acoustic modeling and Language modeling:

The acoustic model takes audio as input and converts it to a probability over characters in the alphabet. The language model helps to turn these probabilities into words of coherent language. The language model (aka. the scorer), assigns probabilities to words and phrases based on statistics from training data. The language model knows that "I read a book" is much more probable than "I red a book", even though they may sound identical to the acoustic model.

Specification and Justification for the Selected Methodology

The extreme programming methodology was selected for the following reasons:

I. Provides a guiding strategy for the development process.

II. Improves the reusability of software modules. iii. Easy integration of machine

learning model on the system for deployment.

3.3 High-Level Model of the Proposed Model

The High-Level Model for Model (Figure 3.1) has 3 modules: Preprocessing/Feature Extraction Module, Parametric Features (Training/Testing Module), and Model generation Module. The user information module illustrated in Figure 3.2 depicts the functionality of the system.

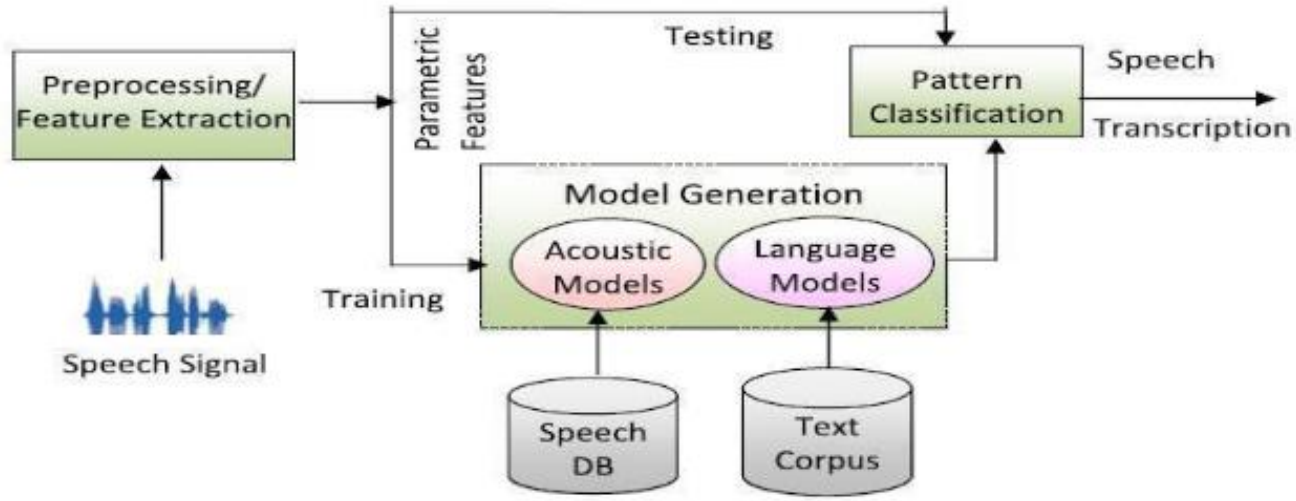


Figure 3.2: High-Level Model for theSpeech RecognitionModel(Ayushi et al., 2017)

Results and Discussion

Module

Specification The system developed in this research work, however,is divided into two modules: The training module and the Testing module.

Training Module:

This module received training accented audio samples of new users and stored them in the db.py database to enable recognition during the test phase.

Testing Module:

This module validates the test sample by comparing its features with trained features and identifies it as the text that is closely matched.

Experiment

During the development of the model, the available dataset was split into two sets.

These are the training set and the validation set. 80% portion of the dataset was used for the training set, while 20% of the dataset was used for the d validation set. During the development of this research work, Google Collab notebook was used to combine executable code. The training data was fit into the machine, and the validation set was used against the training set to obtain accuracy. Epoch, Val-loss, and loss were also used during computation. They provide understanding for the prediction and accuracy. Epoch indicates the number of passes of the entire training dataset the machine learning algorithm has completed. Val-loss is the value of the cost function for cross-validation data and loss is the value of the cost function for your training data. In the case of validation data, neurons using dropout do not drop random neurons.

Result and Analysis

The result shows that concatenating Mel-Frequency Cepstral Coefficients (MFCC) features sequentially and applying acoustic modeling and language modeling provide a solution to the problem of recognizing speech and also converting the speech to text. The experiment provides an accuracy of 83%. Table 4.1 shows the comparisonof the previous methodology with the proposed work.

Table 4.1: The comparison of proposed work with existing related work

Reference	Methods	Result
Ashok and Vikas (2019) Speech Recognition: A Complete Perspective	Linear Predictive Coding (LPC), Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Prediction (PLP).	68% accuracy was obtained from the experiment and it was able to convert spoken words in English to its equivalent English text format.
Adekola *et al.* (2019) Voice Recognition Door Access Control System	NLP, Arduino, Acoustic phonetic and Support vector machine.	A door access control system was built that can serve both abled-bodied and disabled individuals
Akshata *et al.* (2020) Interactive Speech Recognition Agent System using AI	Deep Neural Network	A speech recognition system was built which identifies and understands human speech and converts it to text. 70% of accuracy was achieved,
Proposed research work Speech Recognition Model in Yoruba Language	Mel-Frequency Cepstral Coefficients (MFCC), acoustic modeling and language modeling	The experiment provides an accuracy of 83% and the model was able to convert spoken words in English to its equivalent Yoruba text format.

Performance Evaluation

The performance evaluation of this system was done using different metrics and workloadsto examine how efficient the system is and also to check how well the system can interact using different data inputs. The performance of the system is an important aspect in the design and use of the system which ensures getting optimal performance of the system.

System Testing

This is a process of evaluating the software system opposing the user requirement elicitation and system specification. This ensures that the system is error-free, produces the correct result, secure, reliable, and efficient. Thesoftware system is tested at different levels or phases of testing; Unit testing, integration testing, system testing, and acceptance testing. The unit test was carried out in the course of the development process of the system to test each section of theprogram code. A system test was also carried out to test the requirement satisfaction of the system. Dynamic testing was also carried out to demonstrate the result of the running tests. Usability and accessibility testing were also carried out to show howthe user can easily interact with the interface.

Conclusion

This model was developed with the idea of allowing voice input in the English language and be converted into Yoruba text format. At the final stage of this project, we were able to develop a Speech Recognition Model that allows the computer to translate voice or text requests in English into Yoruba text by concatenating Mel-Frequency Cepstral Coefficients (MFCC) features sequentially and applying acoustic modeling and language modeling on the data set gotten. An accuracy level of 83% was obtained at the end of the project. The output of this project can be used in many aspects for the betterment of the people, there will be fewer communication and education barriers and ease of day-to-day life fordeaf, old people, and people with dysgraphia.

Suggested Area for Future Research

Taking the achievements and limitations of the research into consideration, we intend to keep making improvements to the accuracy of the system by expanding the dataset. Future work should also focus on applying the tools used in this research to some other datasets and comparing the accuracy.

REFERENCES

Abhishek, B. R., Balamurugan, S., Gokul, K. V., & Anitha, R. (2019). Smart virtual assistant bot using voice control and object detection. International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT), 5(1), 772-774.

Adekola, O. D., Akinsanya, A., Olufowobi, A., Babajide, O., Somefun, O. M., & Oduroye, A. (2019). Voice Recognition Door Access Control System. IOSR Journal of Computer Engineering (IOSR-JCE), 21(5), 1-12.

Akshata, T., Mahima, T., Laxmi, V., & Prof. Satish, K. (2020). Interactive Speech Recognition Agent System using AI. International Research Journal of Engineering and Technology (IRJET), 7(12), 1855-1860.

Ali, B. N., Ismail, S., Imtinan, A., Mohammad, A., & Khaled, S. (2019). Speech Recognition Using Deep Neural Networks: A Systematic Review. IEEE Access,7, 19143-19164. doi: 10.1109/ACCESS.2019.2896880.

Anil, K. K., & Munikrishnaprasad, C. (2020). Voice-Based Message Scheduling Android Application. Mukt Shabd Journal, 9(5), 1448-1454.

Ashok, K., & Vikas, M. (2019). Speech Recognition: A Complete Perspective. International Journal of Recent Technology and Engineering (IJRTE), 7(6c), 78-83.

Ayushi, Y. V., Krina, A. S., Yesha, A. K., & Nirali, P. (2017). Survey paper on Different Speech Recognition Algorithm: Challenges and Techniques. International Journal of Computer Applications, 175(1), 31-36.

Chandu,S. T., Pooja, H., Nivedita, S. S., & Nikitha, V. P. (2018). Chatbot to Replace Work of Helpdesk. International Journal of Engineering Research & Technology (IJERT), 6(15), 1-7.

Debajyoti, P., Chonlameth, A., Suree, F., & Vijayakumar, V. (2019). User Experience with Smart Voice Assistants: The Accent Perspective. 10th

International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 574-594). Kanpur: 10th International Conference on Computing, Communication, and Networking Technologies (ICCCNT) Press.

Frioz, K., Shrigowri, V. B., Swetha, K. H., Syeda, A., & Vinutha, K. (2018). Artificial Intelligence-based Travel Bot. International Journal of Engineering Research & Technology (IJERT), 6(15), 1-5.

George, T., & Maya, S. (2020). Voice Assistants and Smart Speakers in Everyday Life and in Education. Informatics in Education, 19(3), 473490. Jayashree, P., & Melvin, J. J. (2015). Machine Learning in Automatic Speech Recognition: A Survey. IETE Technical Review, 32(4), 240-251.

Li, D., & Xiao, L. (2013). Machine Learning Paradigms for Speech Recognition: An Overview. IEEE Transactions on Audio, Speech, and Language Processing, 21(5), 1060-1089.

Linda, M. J., Nilesh, V., & Rajat, S. (2021). Voice Control Human Assistance Robot. International Journal of Engineering Research & Technology (IJERT), 9(3), 394-399. Neha, J., & Somya, R. (2019). Speech Recognition Systems -A Comprehensive Study of Concept And Mechanism. Acta Informatica Malaysia (AIM), 3(11), 1-3.

Omprakash, M. P., Parish, K. A., & Evanjalin, P. R. (2017). Voice Recognition Robot for Surveillance. International Journal of Engineering Research & Technology (IJERT), 5(9), 1-11.

Polyakov, E. V., Mazhanov, M. S., Rolich, A. Y., Voskov, L. S., Kachalova, M. V., & Polyakov, S. V. (2018). Investigation and Development of the Intelligent Voice Assistant for the Internet of Things Using Machine Learning. Moscow Workshop on Electronic and Networking Technologies (MWENT), 1(18).

Pradhyum, B., Shubhangi, M., & Dr. Manisha, S. (2020). Personal Digital Assistant: A Case Study. International Journal of Engineering Research & Technology (IJERT), 8(5), 1-2.

Praphulla, A. S., Ratndeep, R. D., & Pukhraj, P. S. (2015). Speech Recognition Techniques: A Review. International Journal of Scientific & Engineering Research, 6(8), 1693-1698.

Ratnadeep, D. R., & Abdulmali, A. (2018). Automatic Speech Recognition

Techniques: A Review. Signal Processing and Computer Vision (pp. 464470). India: Department of Computer Science and Information Technology, Dr Bahasaheb Ambedkar Marathwada University, 464-470.

Ravikumar, N. R., Prateek, C., Sathvik, B., Rahul, K., & Mayura, T. D. (2020). Virtual Voice Assistant. International Research Journal of Engineering and Technology (IRJET), 07(04), 3399-3402.

Shalini, A., Jayasuruthi, L., & VinothKumar, V. (2018). Voice Recognition Robot Control Using Android Device. Journal of Computational and Theoretical Nanoscience, 15(6/7), 2197-2201.

Shiny, P. K., Rahul, R., Thushar, D., Ramya, K. H., & Varsha, M. A. (2020). Voice-Controlled Guiding Robot. International Journal of Engineering Research & Technology (IJERT), 8(11), 4-6.

Sourabh, M., Gayatri, D., Drishti, B., & Jay, K. (2020). Human Voice Controlled Robot Embedded with Real-Time Obstacle Detection and Avoidance. International Journal of Engineering Research & Technology (IJERT), 8(05), 1-5.

Subhash, S., Ullas, A., Prajwal, S. N., Siddesh, S., & Santhosh, B. (2020).

Artificial Intelligence-based Voice Assitant. Fourth World Conference on Smart Trends in Systems, Security and Sustainability(WorldS4), 593-596.

Yash, P., Rohan, P., Pushkar, P., Umar, P., & Ankita, G. (2019). Voice Controlled Personal Assistant Bot with Smart Storage. International Research Journal of Engineering and Technology (IRJET), 6(3), 75007502.