Speech Recognition Algorithm of Major Nigerian Languages (Yoruba, Hausa, Ibo) Using K-NN

Taiwo Mauyon KUPONU
LCU/PG/002632

Being a MSc Thesis Submitted to the Department of Computer Science, Faculty of Natural and Applied Sciences, Lead City University, Ibadan, Oyo State, Nigeria

In Partial Fulfilment of the Requirement for the Award of Master of Science (MSc) Degree in Information Science

2023

Certification

This is to certify that Taiwo Mauyon KUPONU with matriculation number LCU/PG/002632 carried out this research work titled "Speech Recognition Algorithm of Major Nigerian Languages (Yoruba, Hausa, Ibo) Using K-NN" in the Department of Computer Science, Faculty of Natural and Applied Sciences, Lead City University, Ibadan, Oyo State, for the award of Master of Science (MSc) in Computer Science and that this has not been previously submitted.

…….. .....
.....

Dr. Wilson Sakpere **Date**
Supervisor

…….. .....
.....

Dr. Wilson Sakpere **Date**
Head of Department

Dedication

This research work is dedicated to God, my parents, and my siblings.

Acknowledgement

Foremost, I would like to express my gratitude to the leadership of the Lead City University, Ibadan and also acknowledge the libraries used for creating a medium for us to acquire knowledge for self-reliance.

I acknowledge my supervisor Dr. Wilson Sakpere, for the continuous support for M.Sc study and research, for his motivation, enthusiasm, to guide me through the research. The immense knowledge and plentiful experience have encouraged me in all the time of my academic research and daily life. Besides my supervisor, my sincere thanks go to the Head of Department Dr. Wilson Sakpere, the Postgraduate Coordinator Dr. Azeez Waheed, all other lecturers and staff members in the department of computer science for their guidance encouragement, and insightful comments.

I thank my fellow coursemates, Mr. Oladejo Samuel Adetunji, Mr. Ayomide Feyi-Robinson, Mrs. Temilade Fashina, Mr. Kingsley Efekodedo, Mr. Wasiu Olayinka and others numerous to mention. I also thank Mr. Oluseye Akinmoluwa and Mr. Faruq Ayoade for their kind support on this project.

I thank my parents Rt. Rev'd Dr. S.G Kuponu and Mrs Victoria Kuponu for their guidance, parental advice and commitment instilled in me towards learning. May you both live long in good health to enjoy and the reap fruits of your labour. My sincere thanks to my friends and loved ones for the understanding, sacrifice, and prayers during my study. Also, to my siblings Mrs. Funmi Oni, Mrs. Elizabeth Akinyanmi,

Mrs. Esther Olowookere and Mrs. Kehinde Akinlabi for their continuous support and advice.

“Even though the above-mentioned institutions and persons have assisted in the process of this research work, I alone stand responsible for the errors, if any, found in the work”

Abstract

This study addresses the crucial need for enhanced voice recognition systems in the realm of human-machine interfaces, particularly with a focus on accent identification algorithms, and their application in the context of Nigerian English speakers. The research aims to improve the accuracy and efficiency of speech recognition for these major Nigerian languages using KNN to increase the efficiency and accuracy of accent identification by using a trained data with the three major Nigerian language. Voiced audio samples of speakers from these tribes speaking English and various platforms such as news media and radio recordings was scrapped and recorded and extracted. The audio data was then preprocessed and transformed from the time domain to the frequency domain using the Fourier transform. Matlab R2015A was employed for model training, encompassing input reading, window size and hop size definition, and noise reduction techniques such as high-pass filtering and spectral subtraction. For feature extraction, Mel Frequency Cepstral Coefficients (MFCC) were computed for each audio frame, subsequently aggregated to create fixed-length representations for each dialect sample for about sixty seconds in order to ensure uniformity in the inputs. The model underwent training with a classification algorithm KNN, followed by evaluation, which gave an accuracy rate of 84%. This result indicates that the model proficiently predicts the dialects within the context of English speech. The study's outcomes signify substantial progress in the development of an accent detection model tailored to the major Nigerian tribes: Yoruba, Hausa, and Igbo. The research is a significant stride toward more precise and effective voice recognition systems for Nigerian English speakers, contributing to the broader advancement of human-machine interfaces in an increasingly technology-driven world. It is recommended that future research explores alternative feature extraction techniques, particularly deep learning-based approaches capable of automatically learning relevant features from raw audio data.

Keywords: Accent. Accuracy, Algorithm, Dialect detection, Feature extraction, Fourier transform, Performance, Speech recognition, Voice recognition

Word Count: 298 Words

Table of Contents

Chapter One: Introduction

Chapter Two: Literature Review

Chapter Four: Results and Discussion

List of Tables

List of Figures

List of Acronyms

AI-	Artificial Intelligence
AM-	Acoustic Model
ASR-	Automatic Speech Recognition
CMS-	Cepstral Mean Subtraction
DNN-	Deep Neural Networks
DTW-	Dynamic Time Warping
GMM-	Gaussian Mixture Model
HMM-	Hidden Markov Model
KNN -	K-Nearest Neighbour
LPC-	Linear Predictive Coding

LRE-	Language Resources and Evaluation
MFCC-	Mel Frequency Cepstral Coefficient
ML-	Machine Learning
MLP-	Multi-Layer Perceptron
NLP-	Natural Language Processing
RASTA-PLP-	Relative Spectral Transform-Perceptual Linear Prediction
SI-ASR-	Speaker-Independent Automated Speech Recognition
SVM-	Support Vector Machine
VQ-	Vector Quantization

Chapter One

Introduction

1.1 Background to the Study

The process of automatic speech recognition is one that has been the focus of considerable investigation over the course of several decades. One of the most challenging challenges to solve and one of the topics that has been subjected to the most in-depth research is the facilitation of communication between humans and machines. Depending on the dataset and benchmark test that is used recently developed speech recognition algorithms are able to understand speech with an accuracy that is practically identical to that of humans[1]. This level of performance is only attainable when the system is used to recognise the speech of people from the target language's country of origin (i.e., the native speakers of the language represented by the dataset used to train the ASR system). Even the most technologically advanced speech recognition systems are unable to achieve human-like or even high levels of accuracy when used with individuals who are not native speakers of the language used by the ASR system[2,3]. The presence of patterns relating to the speaker's mother tongue, which can influence the speaker's pronunciation of the second language, is the primary cause of this reduction in proficiency[3]. Because of this, their language becomes skewed to some degree, which, in situations like these, causes the accuracy of the speech recognition system to decrease[1,3].

Because of the rapid pace of change taking place in today's global economy, education system, and mobility of the labour force, there is a pressing need to accurately recognise the speech of non-native speakers, who make up the overwhelming majority of internet users today[1]. Traditional methods for training speech recognition classifiers typically make use of supervised learning techniques[4,5].

While these methods are ideal for recognising speech in cases involving the world's most common languages, they are unable to produce classifiers of an adequate quality for individuals who are not native speakers[6].

An individual's accent may serve as a clue to his or her original language or mother tongue. The capacity to distinguish various accents can help enhance the quality of transcription in a text language by allowing for more precise preprocessing of recordings[7]. Communication is critical in our everyday lives and relationships with other humans and machines. Among the different modes of communication, which include writing, gesture, posture, and eye contact, the most well-known, convenient, and intelligible is speech. Verbal communication entails correct pronunciation, expressiveness, and fluency. Due to variances in speech articulation (sound), English, the most frequently spoken language, has been translated into a number of languages. Numerous reasons, including colonization, commerce, tourism, and migration, have aided in the expansion of English around the world, particularly in Africa, Asia, and South America[7].

English's growth has resulted in the emergence of several dialects of spoken English, including Nigerian English (NE), Singaporean English (SE), and Malaysian English (ME). As a result, it is pronounced in a variety of accents around the world.[2] One of the most difficult aspects of voice recognition is comprehending speech from non-native speakers. Nigeria was a British colony and was home to a diverse range of ethnic groups, each with its own distinct English accent. As a result, Nigerian accent speech occurs in phone conversations that are not familiar with the accent, making speech identification extremely challenging[1,8].

Accent detection and categorization can help enhance the recognition quality of speech. The automated speech recognition system can first determine the speaker's

ethnic origin and then utilize a trained automatic speech recognition system for that accent[9]. Apart from identifying the ethnic origin of a speaker, accent detection is critical in security-related applications such as criminal investigations. In real-world applications, the ability to discern accents from brief bits of audio collected from a distance becomes critical[9]. Accent is one of the constraints, along with gender, that affect the performance of Automatic Speech Recognition (ASR) systems[9]. As our nation, Nigeria, is made up of several ethnic groups, various accents of English pronunciation have developed in their speaking, inherited from their mother tongue's phonemes inventory[1,10]. This complexity is the most essential challenge for systems that do speaker-independent automated speech recognition (SI-ASR)[9]. As far as the market is concerned, there is no specific industrial ASR available for Nigerian English (NE) to address the divergence caused by this unidirectional diversity among the population.

The ASR is more of extracting the speech features and parameterising the features for a decision algorithm that will determine the output of the system. Some ASR algorithms either uses the Mel Frequency Cepstral Coefficient (MFCC) or relative spectral transform-perceptual linear prediction (RASTA-PLP) for their feature vectors and the Gaussian Mixture Model (GMM) or Hidden Markov Model (HMM) as the acoustic parametric model with different training criterion[11]. On the decision part different sequence discriminative training algorithms such as to improve the Minimum Classification Error (MCE) and Minimum Phone Error (MPE) are used to further improve the ASR accuracy[12,13].

The most recent developments in natural language processing have made it abundantly clear that the predominant mode of communication between people and computers in the not-too-distant future will most likely be in natural language. It is

interesting to note that some of the innate communication behaviours of humans are being demonstrated by computers. For instance, a substantial percentage of daily communications are currently taking place between humans and computers[14]. There is no longer a need to interact with other humans in order to pay-in or withdraw money because this process can now be done through an automatic teller machine (ATM). Interaction with smart phone assistants is now simple in order to set an alarm and organise a calendar, as well as find and connect with friends, or search for items on the internet. We may even be able to communicate with other humans who speak different languages by using a "speak and translate" system as a translator (for example, the free live voice and text translator on Apple iTune, which can hold written conversations in 100 different languages and speak 42 different languages). All of these are now realizable thanks to the language technology that are readily available today. It is conceivable that someday in the not-too-distant future, computers will achieve a level of performance in face-to-face communication that is either identical to that of humans or very similar to that of humans.

Language technology is a field that draws from a wide range of academic specializations, including linguistics, engineering, psychology, and computer science. The majority of it is composed of speech technology and natural language processing (NLP)[15]. While speech technology analyses data in its spoken form, natural language processing (NLP) automates the processing of data in its written or textual form using natural language. The process of putting thoughts into words and the process of extracting ideas from words are the two primary coordinated processes that are involved in communication. Context is used in each of these processes in order to decipher the various possible interpretations of ambiguous words and formulate the appropriate message. These are some of the most significant issues that computers

and communication face today. The study of how computers interact with natural language is the focus of the discipline known as natural language processing (NLP), which is a subfield of computer science, artificial intelligence (AI), and computational linguistics.

The processing of human natural language is not an easy undertaking because it includes determining the meanings of words, some of which may have more than one meaning. A term is said to have ambiguity when it can be interpreted in more than one way, and removing this ambiguity requires a comprehension of the word in relation to the context in which it is used[17]. Depending on the degree of ambiguity present in a language, the process of disambiguation can either be an easy or a difficult task. The natural language processing (NLP) tools that are developed to solve the task of word ambiguity for a language have the potential to be further utilised for the development of a machine translator, parser, chunker, word sense disambiguator, and other tools to assist in human-computer communication for that language[17]. As a result, the development of NLP resource tools for languages is an absolute requirement in this day and age. The study of natural language processing has shown a preference for a number of European languages. There are about 6000 different languages spoken across the globe, but only a select few have the natural language processing (NLP) resources necessary for the creation of NLP tools[18].

Language technology systems are overcoming linguistic obstacles; nevertheless, most African languages especially Nigerian accents have insufficient resources and have not been included in this area of research due to a lack of natural language processing (NLP) resources[19,20]. If nothing is done to preserve these languages, it is highly likely that they will die out, and its speakers will be unable to communicate with others around the world using their native tongues. Africa, which has a population of

approximately one billion people, is the second largest and the second most densely inhabited continent in the world[19]. Despite the fact that around 30 percent of the world's languages are spoken on the African continent and 13 percent of the world's population is made up of native speakers of those languages, Africa is considered a dark region[20]. On the Language Resources and Evaluation (LRE) Map, for instance, which is a large database on resources for natural language processing that is freely accessible, the number of English corpora and computational tools is 663. This indicates that English is the most studied language, followed by the languages of French and German, then Italian and Spanish. These are all European languages; on the other hand, there is hardly any evidence of African languages[20].

However, there are two issues to consider, as well as a potential solution, regarding the future of these less privileged languages and the position of the people who speak them as the global information society continues to expand. The problems are as follows: (1) a few large languages end up dominating the place, which leads to the gradual extinction of smaller languages; (2) a few large languages end up dominating the place, which leads to the marginalisation of speakers of the smaller languages even if those languages are preserved. The proposed answer is to make use of language and voice technology in order to ensure the participation of all Europeans in the European Union on an equal basis, regardless of the language that they speak. NLP research in African languages is an important aspect of research for a number of reasons, not the least of which is the fact that it is excluded.

African languages especially Nigerian have a linguistically rich variety of traits, and it is important for the advancement of natural language processing (NLP) research that these qualities be made known to a wider global audience. There are around 600

million mobile users on the African continent, which demonstrates the economic significance of learning an African language (more than Europe and America)[20].

It is worthwhile to conduct natural language processing (NLP) research in the field of low-resource languages because not only will it provide natural language processing (NLP) tools for the language, but it will also give insight on linguistic phenomena that are not found in already resource-rich languages. The findings of the NLP study could encourage more work to be done by NLP researchers, as well as participation in NLP research from native speakers of the language.

1.2 Statement of Problem

Accents may convey a great deal of information about a person's origins, such as their original language, country of origin, ethnic group, and accent categorization. As human-machine interfaces advance in the rapidly developing worldwide market of technologies, it is critical to enhance the voice recognition system. However, there is paucity in literature on accent identification algorithm (such as K-NN) for certain accent in Nigerian English speakers' especially (Yoruba, Hausa and Igbo). Thus, this work aims to use KNN to increase the efficiency and accuracy of accent identification by using a trained data with the three major Nigerian language. This research demonstrates the feasibility of accent-dependent automatic speech recognition by applying a supervised learning algorithm to the job of detecting and distinguishing three Nigerian ethnic groups (Yoruba, Igbo, and Hausa).

1.3 Research Aim and Objectives

The aim of this research is to develop a model to improve accurately the speech recognition algorithm of Nigeria major languages (Yoruba, Hausa, Ibo).

The specific objectives are to:

i. acquire a speech data using the three languages (Yoruba, Hausa and Igbo)

ii. create a comprehensive model for the accent recognition system and divide the speech data into training set and testing set using KNN algorithm.

iii. perform feature extraction and preprocessing of data using MFCC algorithm

iv. classify and detect Nigerian accented speech in (ii) by predicting the target class (Yoruba, Hausa and Igbo) and testing the developed model to identify it as the accent that closely matches.

v. Perform performance evaluation of the developed model

1.4 Scope of the Study

The project will cover the development of a model to improve accurately the speech recognition algorithm of Nigeria accent in the 3 major tribes (Yoruba, Hausa, Igbo) residing within Oyo State, Nigeria. These research procedures would facilitate the generation of data that would be used to improve accurately the speech recognition algorithm of Nigeria accent in the 3 major tribes.

1.5 Significance of the Study

Accent identification is a preprocessing step to speech recognition. This aids in more proficient speech recognition. Hence, problem of identifying accents of the three major Nigerian languages (Yoruba, Igbo, and Hausa) can be solved.

Government agencies as well as the private can employ the results of this study for criminal investigations. Thus, the results of this study will serve as a reference point for government and all other stake holder on issues on security and investigations

Academically, the study will add to knowledge on issues speech recognition or text to-speech. The outcome of this study will also serve as a reference material to students of computer science, lecturers and researchers; and it would also propel further research on the topic.

1.8 Operational Definition of Term

Accent: Can serve as a clue to someone's original language or mother tongue. The capacity to distinguish various accents can help enhance the quality of transcription in a text language by allowing for more precise preprocessing of recordings

Automated Speech Recognition System: Can first determine the speaker's ethnic origin and then utilize a trained automatic speech recognition system for that accent. The ASR is more of extracting the speech features and parameterizing the features for a decision algorithm that will determine the output of the system.

Language Technology: Is a multidisciplinary field comprising linguistics, psychology, engineering, and computer science. It is predominantly divided into speech technology and natural language processing (NLP)

Natural Language Processing: Is a field of computer science, artificial intelligence (AI), and computational linguistics which studies the interaction between computers and natural language

Verbal Communication: Entails correct pronunciation, expressiveness, and fluency

Endnotes

1. K Radzikowski, L Wang, O Yoshie, R Nowak. *Accent modification for speech recognition of non-native speakers using neural style transfer*. **EURASIP Journal on Audio, Speech, and Music Processing**. 2021 Dec;2021(1):1-0.

2. R Contreras, A Ayala, F Cruz. *Unmanned aerial vehicle control through domain-based automatic speech recognition. Computers*. 2020 Sep 19;9(3):75.

3. A Koenecke, A Nam, E Lake, J Nudell, M Quartey, Z Mengesha, C Toups, JR Rickford, D Jurafsky, S Goel. *Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences*. 2020 Apr 7;117(14):7684-9.

4. H Liu, B Lang, M Liu, H Yan. *CNN and RNN based payload classification methods for attack detection. Knowledge-Based Systems*. 2019 Jan 1;163:332-41.

5. S Dargan, M Kumar, MR Ayyagari, G Kumar. *A survey of deep learning and its applications: a new paradigm to machine learning.* Archives of Computational Methods in Engineering. 2020 Sep;27(4):1071-92.

6. K Radzikowski, R Nowak, L Wang, O Yoshie. *Dual supervised learning for non-native speech recognition*. **EURASIP Journal on Audio, Speech, and Music Processing**. 2019 Dec;2019(1):1-0.

7. FO Oladipo, RA Habeeb, AE Musa, C Umezuruike, OA Adeiza. *Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers*.2021

8. T McArthur, J Lam-McArthur, L Fontaine, editors. *Oxford companion to the English language*. Oxford University Press; 2018 May 14.

9. F Oladipo, RA Habeeb, AE Musa. *Accent Identification of Ethnically Diverse Nigerian English Speakers*. Available at SSRN 3666815. 2020 Jul 24.

10. F Idowu. "*Pronunciation intelligibility of Nigerian Speakers of English*." PhD diss., University of Roehampton, 2019.

11. AB Abdusalomov, F Safarov, M Rakhimov, B Turaev, TK Whangbo. *Improved Feature Parameter Extraction from Speech Signals Using Machine Learning Algorithm*. Sensors. 2022 Oct 24;22(21):8122.

12. SA Amuda, O Ibrahim. *Mathematical Profile Of Automatic Speech Recognition Algorithm*.

13. A Becerra, JI De La Rosa, E González. *Speech recognition in a dialog system: from conventional to deep processing*. Multimedia Tools and Applications. 2018 Jun;77(12):15875-911.

14. CT Carr. *Computer-mediated communication: A theoretical and practical introduction to online human communication*. Rowman & Littlefield; 2021 Apr 29.

15. D Mazzei, F Chiarello, G Fantoni. *Analyzing social robotics research with natural language processing techniques*. Cognitive Computation. 2021 Mar;13(2):308-21.

16. K Chowdhary. *Natural language processing. Fundamentals of artificial intelligence*. 2020:603-49.

17. B Cevoli, C Watkins, Y Gao, K Rastle. *Shades of meaning: Natural language models offer insights and challenges to psychological understanding of lexical ambiguity*.

18. B Bigi, OS Abiola, Caron B. *Resources and Tools for Automated Speech Segmentation of the African Language Naija (Nigerian Pidgin).* InLanguage and Technology Conference 2020 (pp. 164-173). Springer, Cham.

19. MD Correia. *Building safe food chains in developing countries: implications of a case study, Mozambique* (Doctoral dissertation, Universidade de Lisboa, Faculdade de Medicina Veterinária).

20. IE Onyenwe. *Developing methods and resources for automated processing of the african language igbo* (Doctoral dissertation, University of Sheffield).

Chapter Two

Literature Review

Conceptual Review

2.1.1 Speech

One of the earliest and most natural ways for humans to share information with one another is through the medium of speech. We acquire all of the necessary skills during our early childhood years, independent of any formal education, and we continue to depend on verbal communication throughout our entire lives. Because it is so second nature to us, we tend to underestimate the level of complexity that underlies the process of speech. The human vocal tract and articulators are biological organs with nonlinear properties, whose operation is not just under conscious control but also affected by factors ranging from gender to upbringing to emotional state[1]. The human vocal tract and articulators are also known as the larynx. As a direct consequence of this, the characteristics of a person's accent, pronunciation, articulation, roughness, nasality, pitch, volume, and speed of their vocalisations can vary widely[1].

In addition, our irregular speech patterns can become even more distorted during transmission as a result of environmental factors such as background noise and echoes, as well as electrical characteristics (if telephones or other electronic equipment are used)[1]. Even with these irregularities, speech can still be transmitted to the majority of the time so long as the language being spoken is one that is familiar to the listener. For centuries, people have endeavored to create machines that are capable of understanding and recognizing speech in the same way that humans do naturally[2]. It is

common knowledge that traditional computers are wired very differently from the human brain. In point of fact, it employs an entirely unique computational pattern in its operation. The human brain, on the other hand, uses a massively parallel collection of slow and simple processing elements (neurons), densely connected by weights (synapses) whose strengths are modified with experience, directly supporting the integration of multiple constraints, and providing a distributed form of associative memory[3]. Conventional computers use a very fast and complex central processor with specific programme instructions and locally addressable memory, whereas the human brain uses a massively parallel collection of slow and simple processing elements (neurons), densely

Communication between humans can be conceptualized as an all-encompassing model of the linguistic process of speech production[4]. In the perception of speech between the speaker and the listener; there are five distinct elements: formulation of speech; human vocal mechanism; acoustic air; perception of hearing; understanding and intervention[5]. The formation of the speaker's voice in their own mind is connected to the first component, which is referred to as the formulation of speech. In order to generate the actual waveform of speech, the human vocal mechanism (human vocal mechanism) employs this formulation. Waveform is communicated to the listener through the air (also known as acoustic air)[6].

During this transfer, the sound wave is susceptible to interference from outside sources, such as noise, which can result in a wave with a greater degree of complexity. Hearing system the precepts of listening from the mind of waveform (perception of the ear), and the listener begins to process this form of wave to understand its contents, so that the listener understands what the speaker is trying to tell you when the wave

reaches the listener's (ears) ears. One of the challenges associated with speech recognition is to "emulate" the way in which one would go about listening to the speech that was produced by the speaker[7].

During the process by which speech signals are processed, the listener's head and system are both actively engaged in a number of activities. The act of perceiving something can be thought of as the opposite of the process of producing speech[8]. Phonemes are the fundamental theoretical unit that describe how mental speech formation contributes to linguistic meaning[9]. Phonemes can be categorised into different sounds based on the properties of either of the two weather ways to wave or frequency characteristics, and this can be done by grouping phonemes according to the properties of either of the two.

However, speech is the variable signal, a communication process that is structured solidly, depends on known physical movements, is composed of acquaintances, various units (phonemes), and is unique for each speaker. Speech can be fast, slow, or variable speed; it can have high pitch, low pitch, or be whispered; it is subject to a wide variety of environmental noise; it may lack distinct boundaries between units (phonemes); and it has an unlimited number of words[10].

2.1.1.1 Properties of Human Voice

The frequency of a sound is the characteristic that stands out the most[11]. The ability to differentiate between sounds is facilitated by frequencies. When there is a high frequency involved, the sound is piercing and can be very annoying. When the frequency of a sound is decreased, the sound will become more profound. Waves that are produced as a result of the vibration of materials are referred to as sound waves. Around 10 kilohertz is the highest frequency that is capable of giving rise to a human

being. And the value that is the lowest is 70 Hz[12]. The highest and lowest possible values are listed here. This range of frequency varies greatly depending on the individual. In addition, the decibel scale is used to measure the intensity of a sound (dB). The frequency of a typical human language can range anywhere from 100 to 3200 hertz (Hz), and its magnitude can be anywhere from 30 to 90 decibels (DB). The range of frequencies that can be perceived by the human ear extends from 16 hertz to 20 thousand hertz. The sensitivity of the human ear can detect changes in frequency of up to 0.5%[13].

2.1.2 Speech Recognition

The process of converting an acoustic signal that was captured by a microphone or a telephone into a set of words is referred to as speech recognition[14]. Applications such as command and control, data entry, and document preparation can all use the recognized words as their final results. They are also capable of acting as the input to additional linguistic processing, which is necessary in order to achieve speech comprehension. The accuracy of the speech recognized by a speech recognition system is dependent on a number of different factors, including but not limited to the following [15]:

a. Vocabulary Size and Confusability: In general, it is not difficult to differentiate between a small set of words; however, the error rate will naturally increase as the size of the vocabulary grows. On the other hand, even a limited vocabulary can be difficult to understand if it contains words that can be confused with one another[16];

b. Speaker Dependence and Independence: A speaker dependent system, by definition, is designed for use by a single speaker, while a speaker independent system is designed for use by any speaker. However, a speaker dependent system

is designed for use by a single speaker. It is difficult to achieve speaker independence because the parameters of a system become tuned to the speaker(s) that it was trained on, and these parameters tend to be very speaker-specific[16];

c. Isolated Discontinuous, or Continuous Speech: The term "isolated speech" refers to individual words. When someone speaks in complete sentences with pauses inserted between the words, they are said to be using discontinuous speech. When we talk continuously, we use sentences that flow naturally. Recognition of isolated and discontinuous speech is relatively simple due to the fact that word boundaries can be identified and words are typically pronounced in a clear and concise manner. Continuous speech is more challenging because the boundaries between words are less clear and their pronunciations are more muddled as a result of the slurring of speech sounds [16];

d. Read vs. Spontaneous Speech: Either speech that is read from prepared scripts or speech that is uttered spontaneously can be used to evaluate the performance of various systems. It is much more difficult to understand spontaneous speech because it often contains disfluencies such as "uh" and "um," false starts, incomplete sentences, stuttering, coughing, and laughter; in addition, the vocabulary is essentially limitless, so the system must be able to deal intelligently with unknown words[16]. Spontaneous speech is significantly more difficult to understand than scripted speech.;

e. Change the Acoustics of the Room: The amount of noise is a crucial component of the ASR. In actuality, noisy conditions or acoustic costumes are when the limitations of today's engines become most obvious when using acoustic suppression ratios (ASR).

f. Temporal Variation: Various speakers speak at different speeds. Today the ASR engines only not able to adapt to.

g. Different Accents: Each individual speaks with his or her own distinct accent. There is a wide range of individual differences in how words are pronounced. This presents a significant challenge for ASR. Nevertheless, this is a problem that is not unique to ASR but rather affects people when they listen to audio.

h. Adverse Conditions: A system's performance can also be reduced by a range of adverse conditions. These include environmental noise (e.g., noise in a car or a factory); acoustical distortions; different microphones; limited frequency bandwidth (in telephone transmission); and altered speaking manner (shouting, whining, speaking quickly)[16].

A speech recognition (SR) system can, in its most basic form, either depend on the speaker or operate independently of the speaker. A speaker-dependent system is designed to be utilised by a single individual, and as a result, it is only capable of comprehending a single type of speech pattern during its training. A system that is independent of the speaker who is using it is designed to be usable by any speaker and is, of course, more difficult to achieve. These systems typically have error rates that are three to five times higher than those of speaker-dependent systems[16]. In general, the process of speech recognition can be broken down into a number of different approaches, as illustrated in Figure 2.1.

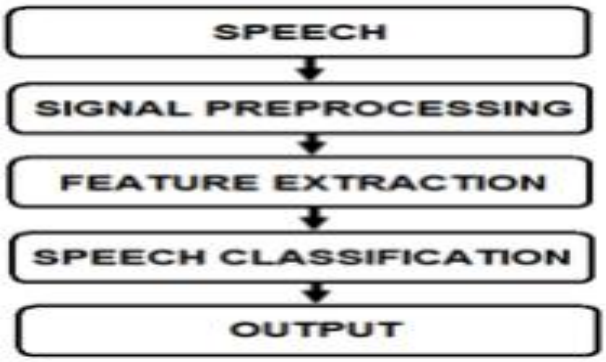


Figure 2.1. Speech Recognition Process[16].

The first block consists of the acoustic environment plus the transduction equipment (microphone, preamplifier and AD converter) that have a strong effect on the generated speech representations[16]. For instance, we can have additional impact generated from additive noise or room reverberation. The second block is intended to deal with these problems, as well as deriving acoustic representations that are both good at separating classes of speech sounds and effective at suppressing irrelevant sources of variation. The third block must be capable of extracting speech specific features of the pre-processed signal. This can be done with a variety of techniques like cepstrum analysis and the spectrogram. The fourth block tries to classify the extracted features and relates the input sound to the best fitting sound in a known 'vocabulary set' and represents this as output.

The commonly used approaches for speech classification include:

a. Template-Based Approaches: The process of comparing unidentified speech to a library of previously recorded words (templates) in order to identify the most appropriate match[17]. This has the benefit of using word models that are completely accurate; however, it also has the drawback that the prerecorded templates are fixed, which means that variations in speech can only be modelled

by making use of a large number of templates for each word, which eventually becomes impractical.

b. Knowledge-Based Approaches: A method in which "expert" knowledge about different varieties of speech is manually programmed into a system. It is difficult to acquire such expert knowledge and put it to good use, which is why this strategy was deemed to be impractical and replaced with the search for automatic learning procedures[18]. This strategy has the benefit of accurately modelling variations in speech; however, such knowledge is difficult to obtain and put to good use.

c. Statistically-Based Approaches: Hidden Markov Models (HMMs), which use automatic learning procedures, are used to model the statistical variations in speech. This method exemplifies the most recent and innovative approach available. The fact that statistical models are required to make a priori modelling assumptions, which are susceptible to inaccuracy and hinder the system's performance, is the primary drawback associated with these models[19].

d. Learning Based Approaches: It may be possible to use techniques from the field of machine learning, such as neural networks and genetic algorithm programming, in order to compensate for the shortcomings of HMMs. In these models of machine learning, explicit rules or other forms of domain-specific expert knowledge are not required; rather, they can be automatically learned through emulation or the evolutionary process[19].

e. The Artificial Intelligence Approach: is one that makes an effort to mechanise the recognition procedure in such a way that it corresponds to the manner in which a person employs their intelligence in the process of visually analysing the measured acoustic features in order to come to a conclusion[19,20]. Within this

methodology, the use of expert systems is prevalent. The artificial neural network (ANN) is a good example of the artificial intelligence approach. ANN provides a method for computing that is modelled after the functioning of biological nervous systems[20]. The idea of artificial neural networks is deeply rooted in the recognition that, despite the fact that the human brain performs functions approximately a million times slower than digital computers, the human brain is more effective when it comes to performing a complex set of tasks such as speech synthesis, visual information processing, handwriting analysis, and so on. This realisation is where the idea of ANNs originates. This can be partially explained by the fact that the human brain is primarily organised in the form of a parallel network of biological neurons. ANNs are physical cellular systems that are capable of learning from, storing, and applying the results of experiments. ANNs have been applied to an increasing number of complex real-world problems[19].

2.1.2.1 Accuracy of Speech Recognition

The accuracy of an SR system is commonly measured with WER[21].

$$\text{WER} = \frac{\text{number of substitutions + Insertions + Deletions}}{\text{Total number of words}}$$

The study of speech signals and the methods that can be used to prepare these signals is referred to as speech processing. Signals are typically organised in a portrayal, and computerised voice processing can be thought of as the connection between sophisticated signal processing and natural language processing[22]. Processing of natural language is a subfield of both artificial intelligence and linguistics[23]. It focuses on the challenges that come with understanding common human languages and automated technology. Normal dialect understanding frameworks change examples of human language into more formal introductions that are less difficult for computer

projects to control. Normal dialect era frameworks convert data from computer databases into the everyday human dialect of sound. Normal dialect era frameworks also convert examples of human language.

The conversion of spoken language into its accurate transcription is another aspect of the field known as automatic speech recognition (ASR). This process is performed by a computer. It is necessary, as a first step, to pre-process the recorded speech in order to use it as an input signal. This is essential due to the fact that "raw" speech-audio recordings typically vary in their type and form and are challenging for algorithms to identify. After the audio files have been transcoded into a standard format, such as acoustic vectors, those files, along with appropriate transcriptions, can be used to train automatic speech recognition (ASR) models. Already in the 1950s, people were making their first attempts at developing automatic speech recognition. The "Audrey" system, which was developed by Bell Laboratories, could already recognise the digits 1-9 with an accuracy that was greater than 90%. Deep learning and large amounts of data have contributed, in part, to the tremendous progress that has been made in ASR models in recent years. They are now an essential component of life in the modern era and can even be carried around with you in the form of voice assistants that are available on the majority of mobile devices. Applications range from basic assistance work and home control all the way up to translation from one language to another using only human speech.

2.1.2.2 Speech Coding

The compression of speech (in code) for transmission with codecs of discourse using methods of handling discourse and sound flag preparing is what is known as speech coding[23]. The techniques that are used include things like the compression of sound

information and human sound coding, both of which involve the application of knowledge gained from the study of psychoacoustics in order to transmit only the information that is relevant to hearing. For example, limit band voice coding only transmits data on the recurrence of 400 Hz to 3500 Hz band, but the reproduced flag is still satisfactory in terms of clarity. When it comes to the coding of the voice, the most important factor is ensuring that the discourse can still be understood while maintaining its "sensitivity." This is done while simultaneously reducing the amount of information that is transmitted[24]. It should be noted that the comprehensibility of discourse encompasses, in addition to the specific content, the speaker's character, feelings, inflection, timbre, and other aspects that are essential for perfect coherence. It is imperative that this fact be taken into consideration. It is possible that the corrupted discourse is completely understandable, but still subjectively irritating to the audience. This is why the idea that the sensitivity of the debased discourse is a property is more novel than the idea that understandability is a property of the debased discourse.

2.1.2.3 Speech Synthesis

The generation of human discourse that is produced by speech synthesis is a counterfeit. The transformation of ordinary content into discourse is accomplished by a framework known as "content to discourse" or "TTS." Various frameworks render typical semantic portrayals such as phonetic interpretations into discourse[23]. Connecting different segments of recorded discourse that have been saved in a database is another method for producing blended discourse. The frameworks have varying capacities in terms of the number of discourse units that can be stored; a framework that stores telephones or diaphones has the potential to give the greatest yield go but may require clarity. When applied to specific locations, the capability of

whole phrases or expressions enables the production of high-quality results. On the other hand, a synthesizer has the ability to

combine a model of the vocal tract with other human voice characteristics in order to produce an entirely "engineered" voice output. The quality of a discourse synthesizer can be evaluated according to how closely it resembles the sound of a human voice and how easily it can be understood by others. Listening to written works on a home computer is now possible thanks to a programme called coherent content to discourse. This software converts written content from a home computer into spoken language.

2.1.2.4 Speech Classifier

The issue of alleged example acknowledgment (ASR) is connected to a much larger and more comprehensive topic in logical and design-based alleged example recognition. The objects of interest are generally referred to as designs, and in the context of this discussion, they are sequences of acoustic vectors that are extracted from a conversation by making use of the tactics that were discussed in the prior section. The classes make references to the speakers who are present here. Because the ordering method for our circumstance is based on distinct components, it is also possible to refer to it as highlight coordinating. Dynamic Time Warping (DTW), Hidden Markov Modeling (HMM), and Vector Quantization (VQ) are the three highlight coordinating strategies that are considered to be the best in their respective categories and are used in the process of speaker recognition[24].

2.1.2.5 The Speech Recognition Process

The first step in the process of automatic speech recognition is known as speech analysis, which is also referred to as front end analysis and feature extraction[25]. The goal of this process is to extract the acoustic characteristics of the speech waveform. The output of the analysis front end is a set of parameters that represent the observed acoustic properties of input signals of speech. These parameters can then be put to further use in acoustic modelling in order to make it more compact and efficient. There are three primary categories of front-end processing techniques: linear predictive coding (LPC), mel-frequency cepstral coefficients (MFCC), and forecast linear perceptual (PLP), with the latter two being the ones that are most frequently utilised in automatic speech recognition (ASR) systems.

Linear Predictive Coding: LPC begins with the presumption that a speech signal is transmitted by a bell toward the end of a tube (voice sounds), with intermittent shrieks and pop sounds also included in the transmission. Despite the fact that it is obviously crude, this model is a reasonably close approximation to the reality of the process of discourse creation. The glottis, also known as the space between the vocal ropes, is responsible for producing the buzz, which can be characterized by its intensity (commotion) and its frequency (pitch). The tube is shaped by the vocal tract, which includes the mouth and the throat. The framing of the tube is described by the resonances that it produces. The movement of the tongue, lips, and throat in conjunction with sibilants and plosives results in the production of shrieks and snaps. The LPC analyses the estimation of the formant discourse flag, which involves removing the effects of the discourse flag in order to determine the strength and frequency of the rest of the buzz. The process of removing formants is referred to as switch sifting, and the signal that is left after the subtraction of the separated flag demonstrating is referred to as buildup. Formants and buildup flag numbers, which

depict the force and recurrence of the gossipy tidbits, can be stored away or transmitted to another location. LPC combines the flag of discourse by inverting the procedure, which is described as follows: use the buzz and various parameters to create a source flag; use the formants to create a channel (which speaks to tube); and run the source through the channel, which talks. This procedure is performed on the discourse flag short pieces, which are called outlines; generally speaking, thirty to fifty outlines for each second give coherent discourse with a high-pressure level. Because discourse signals change after some period of time, this procedure is performed on the discourse flag.

Approach Based on Probability: The probabilistic approach, which involves computing a score for matching spoken words with a speech signal, is the method that is most frequently utilised in automatic speech recognition systems today. A probability value is associated with each word or string of words in the vocabulary, and each of these associations corresponds to a speech signal. The score is derived from the phonemes in the acoustic model, and linguistic knowledge is used to determine which words can follow other words in the sentence. The recognition result will be determined by selecting the word sequence that received the highest score.

Pre-processing, feature extraction, decoding, and post-processing are the four stages that make up the SR process. These stages can be thought of as occurring in sequential order. The following is merely an illustration of one possible implementation of each step that can be found in various SR systems; however, these do not all look the same.

Pre-Processing: is the recording of speech with a sampling frequency of, for example, 16 kHz, and, a bandwidth limited signal can be reconstructed if the sampling frequency is more than double the maximum frequency, which means that frequencies

up to almost 8 kHz are constituted correctly[27]. Pre-Processing is the first step in the signal processing chain. It has been demonstrated that data sent over a telephone network at frequencies ranging from 5 Hz to 3.7 kHz is adequate for recognition; consequently, 8 kHz is more than sufficient. Because they are considered to be noise, we can get rid of any frequencies that are lower than 100 Hz. The removal of the segments that come before and after the user's speech as well as those that come between the start of the recording and when the user begins speaking is an essential component of the pre-processing step.

This is done to combat the fact that an SR system will assign a probability, even if it is very low, to any sound-phoneme combination, allowing background noise to insert phonemes into the recognition process. This is done to counteract the fact that an SR system will do this. When examined over a short period of time (5-100 ms), speech signals are slowly timed varying signals, and their characteristics are relatively stable when viewed in this context. As a result, acoustic observations are extracted in frames that are typically 25 milliseconds long during the step known as feature extraction. Calculating a multi-dimensional vector for the acoustic samples in that frame, and then performing a fast Fourier transformation on that vector, allows one to transform a function of time, such as a signal in this instance, into the frequencies of the signal's components[29].

In Cepstral Mean Subtraction, abbreviated as CMS: The step of cepstral mean subtraction (CMS), which is used to normalise differences between channels, microphones, and speakers, is one of the most common steps in the feature extraction process[30]. During the decoding process, calculations are performed to determine which string of words provides the most likely match to the feature vectors. The results of these calculations are then displayed. In order for this step to be successful,

there must be three things present: an acoustic model that includes a hidden Markov model (HMM) for each unit (phoneme or word), a dictionary that includes possible words and the sequences of phonemes that make up those words, and a language model that includes the likelihoods of words or word sequences.

In decoding, the SR systems tries to find the word or sequence of words w* best matching the observation X, giving the equation 2.1 with p(w) being from the language model and p(X|w) from the phoneme sequence in the vocabulary calculated by equation 2.231.

$$w^* = \text{argmax}_w \, (p(X|w)\, p(w)) \qquad 2.1$$

$$p(X|w) = \text{argmaxs}(\prod(p(x|sj)p(sj)) \qquad 2.2$$

However, as the number of possible state sequences increases, it will no longer be possible to calculate all of the probabilities. An optimal recursive solution that can be used to solve this problem is the Viterbi search algorithm, which estimates the state sequence that is most likely to occur next[31].

Mel frequency Cepstrum Coefficients: These are derived from a particular type of representation of the cepstral of the audio clip, which is known as a "spectrum-of-a-spectrum." The frequency bands in the Mel-frequency cepstrum are placed logarithmically (on the mel scale), which approximates the response of the human auditory system more closely than the frequency bands that are spaced linearly and are obtained directly from the FFT and DCT[33]. This is the primary distinction between the cepstrum and the Mel-frequency cepstrum. This makes it possible to process data more effectively, for instance in the process of audio compression. On the other hand, in contrast to ultrasound, CSBMS do not have a land of ear mode; as a result, they might not accurately represent the perceived loudness. The following is a common method for deriving CSBMS: Consider a signal, and perform the Fourier transform on

it (an extract from the window). A windowed and superimposed triangular map showing the amplitudes of registration for the retrieved over spectrum on the Mel scale Consider the Mel list register of amplitudes and apply the discrete cosine transform to it as though it were a signal. The amplitudes of the spectrum that was produced are denoted by the CSBMS[33].

The Connection Temporal Classification, or CTC: Given that different speakers' pronunciations of individual phonemes, letters, and even words can vary significantly from one another, GMM and HMM models require a large number of distinct hidden states in order to function properly[33]. This is of the utmost importance in the temporal context, as it is impossible to determine with absolute certainty which segment of the input audio sequence corresponds to which segment of the output.

A solution for this problem was provided with the connection temporal classification (CTC) algorithm[24]. This algorithm tries to assign an input sequence X = [x1, x2, ..., xT], e.g. audio files, to an output sequence Y = [y1, y2, ..., yT], e.g. characters, while ensuring temporal flexibility. In other words, the sequences do not necessarily have to be of the same length and no direct assignment has to take place. Thus, an assignment of X2 to Y2 would not necessarily preclude X3 being assigned to Y2 as well. As in the HMM model, the best matching assignment of sequence Y to sequence X is searched for:

$$\text{argmaxs } P(y \mid x) \qquad 2.3$$

However, since multiple assignments of Y are possible, a problem arises. For example, if the word" Hello" is predicted, the possible transcription could result in: [h, h, e, e, l, l, o, o][24]. To counteract this, repeating letters have to be removed, which would result in the word Helo. To address this, a new Blank ε is introduced to separate repeating letters. With this blank token, a new prediction would form the following output: [h, h,

e, e, l, l, ε, l, o, o]. This can be expressed by choosing the most likely tokens a for each time-step t, leading to the following equation[24]:

arg maxA 2.4

This would lead to a correct transcription of Y =" Hello" when all repeating letters and Blank Tokens are removed in A

2.1.3 Speech Recognition Models

2.1.3.1 Hidden Markov Models

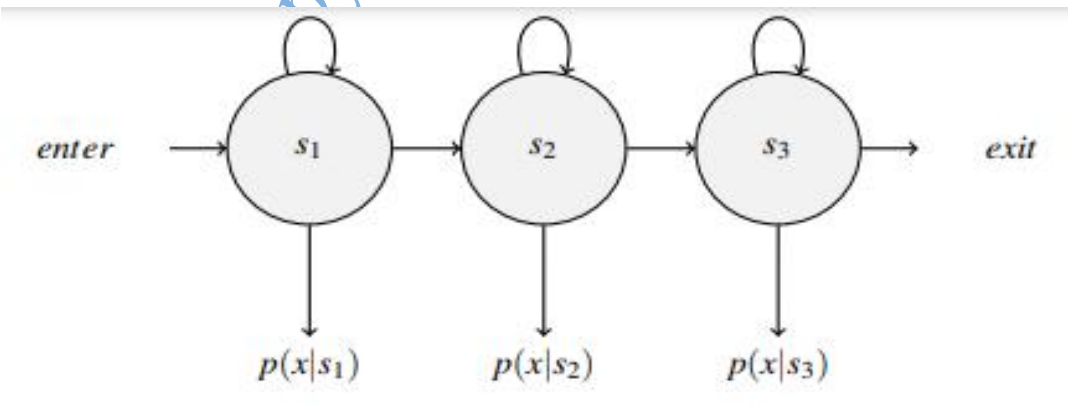


Figure 2.2: **A Three-State Hidden Markov Model[24].**

Figure 2.2 shows an illustration of a three-state HMM where each state si has a probability density p(x|si) that states the probability density for the acoustic

observation x for the state si. The word S can be formed using the three states s1, s2, and s3. HMMs are trained on speech data, and if the data comes from a sufficient number of speakers, the model can be considered to be independent of the speaker from whom the data originated. In the field of acoustic modelling, HMMs can be utilised for either spectral or temporal analysis[24,32].

2.1.3.2 Gaussian Mixture Models

With enough components, these models can model probability distributions to any level of accuracy[24,34]. The Gaussian mixture model, or GMM, is commonly used for determining how well each HMM state fits a frame of the acoustic input, also known as the probability. After it has been trained, a GMM-HMM system's accuracy can be improved even further through the process of fine-tuning[24]. GMMs are utilised in the process of pattern recognition and model the probability density function in a manner that is analogous to that of neural networks. One example of this would be sound signals. In a GMM/HMM model, there is an HMM associated with each class (for example, phonemes) that contains a unique set of hidden states. Each class has several hidden states to account for the various ways in which it can express itself linguistically. If this model is provided with an acoustic input x at this point, it will work to estimate the most likely sequence of hidden states s, which in this case would be phonemes. Now, with an acoustic observation x in hand, a GMM/HMM model will make an attempt to compute the state s with the maximum likelihood[24,35].:

$$\arg\max_s P(s \mid x) \qquad 2.5$$

The probability can be re-written using Bayes rule

$$\text{argmax}_s \frac{P(x \mid s)\, P(s)}{P(x)} \qquad 2.6$$

In this case, the probability P(x), also known as the probability that this observation will take place, is statistically independent of the state s, which means that it can be disregarded in the process of maximisation. A language model has the capability of determining the probability P(s), also known as the probability of the phoneme[24]. The last remaining probability P(x | s) can now be trained by a multidimensional GMM.

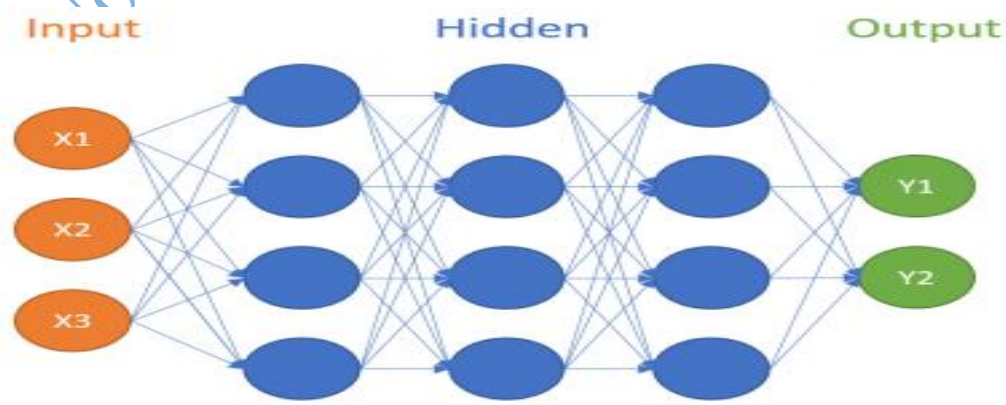
These GMM/HMM models have been the standard in ASR models for decades and have been constantly improved by new algorithms to enhance the finding of model parameters, such as maximum a posteriori or maximum likelihood linear regression. Unfortunately, the performance of the GMM/HMM models is highly dependent on the size of the training datasets. The reason for this is that the GMM model requires a representation that contains all relevant information, since the distribution of input features is performed directly. In order for a GMM/HMM system to achieve an accuracy of over 95%, 100,000 hours of training data would be required[37].

2.1.3.3 Deep Neural Network Model

For a number of decades, the majority of ASR systems relied on GMM/HMM models as their foundation. This changed as computing power continued to steadily increase, particularly in the graphical domain, which made it possible to carry out a large number of computations in parallel. When computing power began to be used in the field of machine learning for deep neural networks, it started to become relevant for the area of automatic speech recognition (ASR) (DNN). A network that is made up of multiple neurons is referred to as a deep neural network.

These neurons are typically arranged in layers, with each preceding layer being completely connected to the layer that comes after it[38]. The first layer of these networks is also referred to as the input layer because it is the location where the input sequence is applied. Due to the fact that the classifications are returned at this stage, the final layer is also referred to as the output layer. If characters are to be predicted, the size of the output layer needs to be proportional to the size of the number of possible classifications, such as the size of the alphabet. The layers that are in between are sometimes referred to as the hidden layers, and they are the ones in which the actual teaching and learning takes place. Figure 2.2 provides an illustration of a straightforward depiction of such a DNN.

Fig 2.3 Representation of a Deep Neural Network[24].



the incorrect hypothesis Y′ with a correct prediction Y, resulting in an error δ:

$$\delta = |Y - Y'| \quad 2.7$$

This error can be utilised to fine-tune the weights that are stored within the network's neurons. Back-propagation is the term used to describe this process. Given that the input sequences X contain a predictable pattern, which is mapped by Y, the prediction of the network becomes continuously better throughout the training process by adjusting the weights depending on the error. This is provided that the input sequences contain ions of audio features, which were then forwarded to conventional GMM/HMM networks[38].

2.1.3.4 Language Models

Because of algorithms like CTC, the vast majority of today's ASR models do not any longer rely on explicit language models to perform the task of prediction[40]. As a result, it is feasible to train these models from beginning to end, that is, beginning with audio input and ending with text output. This provides a significant benefit in that acoustic ASR models can be trained on speech without first needing to comprehend the context or the grammar of the language. On the other hand, this independence is both a benefit and a drawback because it makes it impossible to validate any predicted character by examining it in its natural environment.

For instance, if a blank token is not correctly predicted, the word "Helo" in this case is a valid prediction that cannot be corrected by the acoustic model[40]. This is because a blank token does not contain any information that can be used to predict it. The Language models come into play at this point, calculating the probabilities for both individual words and word combinations. In this way, these models are able to correct straightforward typos, such as letters that have been switched around or that are missing, but they can also detect and correct errors in the context of a word combination. It is even possible to make a prediction about the next words that are

most likely to be used. In the following sections, the forty most prevalent language models are dissected and discussed.

Gram Language Model: N-Gram models are the ones that are used the most frequently in linguistic research. In an N-Gram model sequences are decomposed into individual fragments. One has complete discretion over the size of the fragments. These may stand in for individual letters, phonemes, or even entire words[41]. The one-gram model, also known as the unigram model, is the simplest representation of an N-gram[40,41]. The probabilities of each individual fragment are the only information that is stored in this model. This straightforward model already demonstrates a high level of efficiency because it recognises and rectifies incorrectly spelled words. For instance, depending on the size of the vocabulary, the word "Helo" either has no probability at all or only a very small probability of being used. As a direct consequence of this, it will be changed to the well-known word "Hello."

However, this model does have a significant drawback in that it eliminates words from the vocabulary that were not previously known to the user. Additionally, context-dependent errors such as "they went swimming" are not yet recognised in this context. This is due to the fact that the word "the," along with the word "they," is a component of the unigram model[40]. The solution to this problem can be found in higher N-gram models, such as the 2-gram model, which is also known as the bigram model, or the 3-gram model, which is also known as the trigram model. These models are capable of storing not only the probabilities for individual fragments, but also the probabilities for the fragments that came before them. In a trigram word language model, the two preceding words would also be considered[40]:

$$\mathbf{P(wn \mid wn^{-1}, wn^{-2})} \qquad 2.8$$

Because this combination has a higher probability than the first one, the language model would probably correct the sentence "the went swimming" to the sentence "they went swimming" in the example that was shown, because this combination has a higher probability. The computational resources become a limiting factor due to the fact that the number of possible combinations increases exponentially with each degree of the N-Gram model, so this despite the fact that the degree of the N-Gram model could in theory be extended indefinitely. However, in practise, technical limits are typically reached in this case[40].

Transformer Language Model: The use of Transformer models is highly recommended for Natural Language Processing endeavours. These models have the capability of self-supervised learning, which enables them to learn the linguistic patterns of words and even the context in which they are used. Because of this, Transformer models can also be utilised effectively when developing a language model[42]. Similar to the MLM learning approach, these methods can identify misspelt words or even words that have been incorrectly classified[40]. On the other hand, if these transformer models are trained on complete words, then unknown words also present a challenge here.

The model is dependent on learned vector representations as a result of the initial embedding layer that is present in both the encoder and the decoder[40]. As a consequence of this, unseen words might be given character-based vector representations, and the model won't be able to accurately capture them. This has a negative impact, as well, on the practise of zero-shot learning. This issue can, however, be sidestepped by training sequence-to-sequence models on other word representations; doing so will allow for the problem to be resolved. Not only do Transformer models use the n most recent words as the basis for prediction, but they

also use the entire input sequence. This is a significant improvement over the traditional n-gram language models that have been used. As a result, the Transformer model has access to the entirety of the input context whenever it attempts to predict the following word.

When trained on full words, however, language models may rely on predetermined sizes of their vocabularies. This presents a significant obstacle for words that are not commonly used (OOV). The use of subwords is one solution to this problem[40,43]. Subwords are used to segment words into their component parts, as their name suggests. However, before anything else, the size of the subword vocabulary needs to be defined[43]. The alphabet is the smallest conceivable vocabulary size that can be used. For instance, the sequence of characters [t,o,k,e,n,i,z,a,t,i,o,n] is obtained when the word "tokenization" is segmented using this subword alphabet vocabulary. In this particular scenario, many different relationships between frequently occurring characters and even whole words can already be learned using this method if an n-gram level that is sufficiently high is selected.

The vocabulary of one's own training dataset should be considered the optimal size for the largest possible vocabulary size. If the word "tokenization" is included in the dataset used for training, the subword model will select "tokenization" as the segmentation for the whole string. If a high n-gram is to be trained on, however, these models can quickly become computationally intensive due to the extremely large vocabulary sizes involved. Because of this, the sizes of the subword vocabularies that are chosen in practise typically fall between these two boundaries. In situations where the vocabulary size is smaller than the vocabulary size of the training dataset, known words are disassembled into more manageable components[44]. For instance, if the word "token" appears multiple times in the dataset, there is a good chance that the

word "tokenization" will be broken down into its component parts [token and iztation] if the vocabulary size is not large enough to store the entire word "tokenization"[40]. This will occur in the event that the vocabulary size is insufficient to store the entire word.

This has the distinct advantage that the previously unknown word "modernization" can now be broken up into components that are already familiar, such as "mordern" and "ization." The subword byte pair encoding is one of the methods that can be utilised to carry out this subword segmentation[40]. Characters that are frequently found together are easily identifiable thanks to the subword byte pair encoding's use of substitution. When applied to neural machine translation, the use of subword byte pair encoding has already demonstrated a significant improvement in the recognition of uncommon words. The use of the unigram language model is yet another possibility for carrying out subword segmentation. A unigram model is utilised in this model, just like the N-Gram Language Model is utilised in this model. In contrast to the N-Gram Language Model, the probabilities of the subword combinations, as opposed to the probabilities of the whole words P(w), are the ones that are relevant in this particular instance[45].

$$\mathbf{P(X)}=\prod_{i=1}^{N} P(S_i) \qquad 2.9$$

With the vocabulary v in hand, a maximisation algorithm can be used to determine the P(s) that represent the most frequent subwords. Iteratively determining the subwords is necessary due to the fact that the vocabulary set itself is a mystery in the real world. Here, first, a vocabulary set that is sufficiently large is heuristically created from the training corpus. This vocabulary set is comprised of the combinations of all characters and the substrings that occur most frequently. It is possible to calculate the probability

of the subwords P(Si) by making use of the EM algorithm if the initial vocabulary is defined[40].

Taking this approach, the loss I of the subword Si can also be calculated after the subword Si has been eliminated from the vocabulary. The lossi in this situation is equivalent to a reduction in the marginal likelihood of each and every subword. If the subwords are arranged in descending order according to the lossi, then the top n percent of the subwords can be chosen. This process can be repeated as many times as necessary until the vocabulary size that you desire is achieved. In order to prevent the use of words that are not in one's vocabulary, it is essential to ensure that the subwords only consist of a single character. SentencePiece is a programming model that includes both an implementation of the subword byte pair encoding and the subword unigram language model. SentencePiece places an emphasis on both the speed of computation and the ease of implementation[46].

2.1.4 Nigerian Major Languages

2.1.4.1 Hausa Language

There are more people in sub-Saharan Africa who speak Hausa as their first language than speak any other language. Hausa is a major language. Additionally, it is classified as a Chadic language within the Afro-Asiatic language family and is spoken by approximately 50 million people across the countries of Nigeria, Niger, Cameroon, Togo, and Ghana[47]. The northern states of Nigeria and the southern regions of the neighbouring Republic of Niger are where the vast majority of the language's speakers can be found[47]. It is a language that is spoken in the northern states of Nigeria, and it is one of the major languages spoken in Nigeria, along with Yoruba and Igbo. Hausa is spoken in Nigeria. There are a significant number of Fulani people who speak Hausa as their mother tongue. There are also communities in the Diaspora

that speak Hausa. However, it is the most important widespread West African language, rivalled only by Swahili as an African lingua Franca, and it has expanded rapidly as a first or second language, especially in Northern Nigeria[47]. This is because Northern Nigeria is home to a large number of people who speak the language. Since more than 150 years ago, serious research has been conducted on the Hausa language, making it one of the sub-Saharan African languages with the best documentation and the most research done on it overall[47].

In addition to the 22 letters of the English alphabet (A/a, B/b, C/c, D/d, E/e, F/f, G/g, H/h, I/i, J/j, K/k, L/l, M/m, N/n, O/o, R/r, S/s, T/t, U/u, W/w, Y/z), the Hausa alphabet also includes four additional characters that are known as "hook letters" The following digraphs are considered to be fundamental in Hausa: dy, fy, gw, gy, kw, ky, y, w, sh, and ts. Hausa has five vowel alphabets: a, e, I o, u. The numbers 0 through 9 are written in Hausa in the following order: Sifiri, aya, Biyu, Uku, Hudu, Shida, Bakwai, Takwas, and Tara respectively[47]. There are three fundamental tones in the Hausa language, and they are low, high, and a falling tone in the middle. It is possible for each of the five vowels /a/, /e/, /i/, and /o/ and /u/ to have a low tone, a high tone, a mid-tone, or a falling tone. In addition to this, it differentiates between short vowels and long vowels, both of which can have an impact on the meaning of a word. In written standard Hausa, neither the vowel lengths nor the tones are indicated in any way[47].

2.1.4.2 Yorùbá Language

Yorùbá, like the majority of African languages, is a tonal language. It is also one of the twelve languages that belong to the Edekiri sub-branch of the great family of the West Benue-Congo language branch of the Niger-Congo phylum of African languages[48]. Yorùbá was one of the first languages to be written down. Native

speakers can be found in the south-western region of Nigeria (the second largest ethnic group in number). There are approximately 30 million native speakers in Nigeria[48]. It is also spoken in the Republic of Benin, Ghana, Sierra Leone, and Côte d'Ivoire. Togo is another country where it is spoken. In addition to being spoken in Africa, Yoruba is also spoken in countries such as Brazil and Cuba, as well as Trinidad and Tobago, where a significant number of native speakers of the language can be found[48]. Unlike non-tonal languages like French and Malay, in which word meaning can be inferred from spelling, Standard Yoruba (SY) is a tonal language like Cantonese and Thai. As a result, the tone of pronunciation that is associated with each syllable of a SY word determines the meaning of that word[48]. The fact that SY is homographic contributes to the overall complexity of the language. One single word in a homographic language can have multiple meanings, depending on how it is pronounced in different tones[49].

Standard Yorùbá is the most common variety of the Yorùbá language, despite the fact that the language has many subtypes (SY). The SY alphabet includes seven vowels (a, e,, I o, u), eighteen consonants (b, d, f, g, gb, h, j, k, l, m, n, p, r, s,, t, w, y), five nasalized vowels (an, n, in, un), and two syllabic nasals The letter combination SY uses for the diagraph gb, which is a consonant, consists of two letters. SY has three different tone levels, which are denoted by the acute accent symbol (ˊ), the macron (), and the grave accent symbol (ˋ) respectively[50]. These tone levels are high tone, mid tone, and low tone. Additionally included in SY are the contrasting tones of rising (R) and falling (L), respectively. The realisation of tones occurs on vowels and occasionally on nasal consonants. It is possible to create a SY syllable by combining vowels (V), consonants (C), and/or nasal vowels (n), which results in the following possible combinations of syllables: CV, CVn, V, N, and Vn[50].

2.1.4.3 The Igbo Language

Igbo, like a lot of other languages, has multiple dialects; there are about thirty of them, and each one has a different contrastive pitch. It also has tones and vowel harmony characteristics, and it has those things in common[51]. The majority of the differences between dialects are found in their lexical, phonological, and syntactic structures. The Owerri and Umuahia dialects, which are spoken in the capital cities of the two eastern states, Imo and Abia, serve as the basis for the standard dialect. There are 28 consonants, including b, gb, ch, d, f, g, gh, gw, h, j, k, kw, kp, l, m, n, ny, n, p, r, s, sh, t, v, w, and z, and 8 vowels, which are divided into two harmony groups based on Advanced Tongue Root (In the process of word formation, the consonants sh and v are utilised infrequently[51]. In order to form Igbo words, the vowels from the two harmony groups are combined according to the rules governing vowel harmony. In some languages, such as Igbo, a phenomenon known as vowel harmony occurs when all of the vowels found in a word belong to the same group[51].

Letter	A	B	Ch	D	E	F	G	Gb
Pronounciation	/a/	/b/	/ʧ/	/d/	/e/	/f/	/g/	/g͡b/
Letter	Gh	Gw	H	I	Ị	J	K	Kp
Pronounciation	/ɣ/	/gʷ/	/ɦ/	/i/	/ị/	/dʒ/	/k/	/k͡p/
Letter	Kw	L	M	N	Nw	Ny	Ṅ	O
Pronounciation	/kʷ/	/l/	/m/	/n/	/ŋʷ/	/ɲ/	/ŋ/	/o/
Letter	Ọ	P	R	S	Sh	T	U	Ụ
Pronounciation	/ɔ/	/p/	/ɹ/	/s/	/ʃ/	/t/	/u/	/ʊ/
Letter	V	W	Y	Z				
Pronounciation	/v/	/w/	/j/	/z/				

Fig 2.4: **The Standard Orthographical Graphemes for Igbo**[51]

2.2 Methodological Review

In the following paragraphs, the components of a conventional ASR system will be discussed. Feature Extraction, Acoustic Modeling, Language Modeling, and Lexical Modeling are the four main components that make up a general automatic speech recognition (ASR) system.

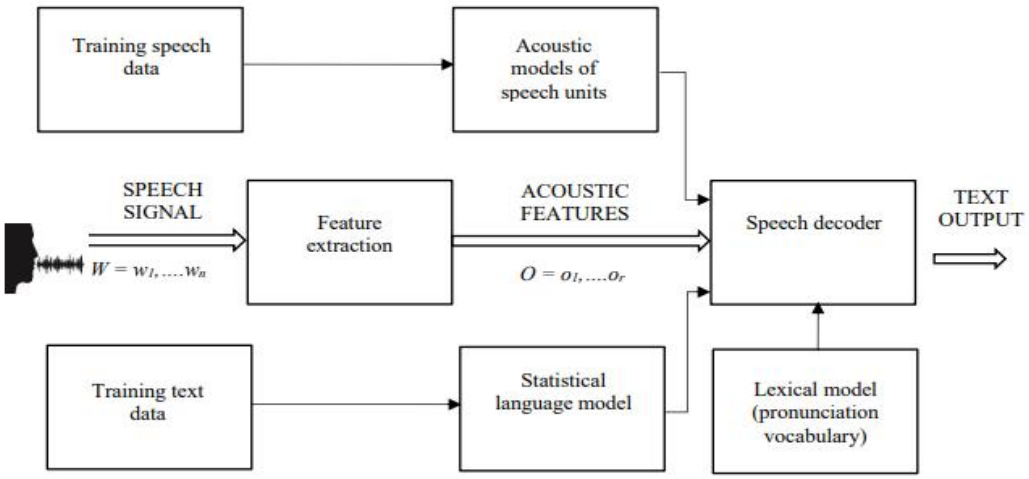


Fig.2.5: Components in an ASR System[52].

2.2.1 Feature Extraction

The feature extraction phase is considered to be the front-end of any ASR system. This phase receives the audio signal as its input and produces the digital representation of the audio signal as its output[52]. Raw audio can be directly given as input to the ASR system, or it can be converted to the frequency domain and either passed as spectrograms or a feature extraction technique can be applied on the frequency domain representation of the audio signal[52]. Another option is for raw audio to be converted to the time domain and then either passed as spectrograms or converted back to the time domain. Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), and Perceptual Linear Training speech data are some of the various feature extraction techniques. Acoustic models of speech units, Training Speech Data Feature extraction, Speech decoder Training, text data Acoustic models of speech units, Training Speech Data Feature extraction The statistical language model, the lexical model (pronunciation vocabulary), the prediction coefficients (PLP), and the bottleneck layer features (Multi-Layer Perceptron (MLP) outputs, specifically using autoencoders) etc[52,53].

Acoustic Modeling: In automatic speech recognition (ASR), the relationship between an audio signal and the corresponding phonemes, characters, or words is represented by means of something called an Acoustic Model (AM)[52,54]. AMs construct statistical representations of meaningful speech units based on the information that they are given. Hidden Markov Models were traditionally used in the construction of ASR models (HMM). The simplest form was when each phoneme was modelled using its own individual HMM, and the probabilities of the HMM were modelled using Gaussian Mixture Models (GMM)[52,54]. This was the most common form. In more

recent approaches, the feature extraction system, the phoneme decoding system, and the acoustic model are all modelled as being part of the same end-to-end deep network. Due to the fact that the phonemes in each language are distinct from one another, the acoustic model is unique to the language that is being modelled. However, by employing speaker normalisation techniques, the acoustic model can be generalised so that it can be applied to varying pronunciations of the same language.

Language Modeling: Following the output of the sequence of phonemes, characters, and words corresponding to the input utterance by the acoustic model, the language model then makes corrections based on prior probabilistic statistics[55]. The Language Model (LM) is a probability distribution over a set of words that describes the probability of sounds, characters, and words occurring together in sequence. This distribution is applied to the entire set of words[52]. Similar to the acoustic model, the language model is typically collected on a large written corpus. This corpus is distinct from the acoustical corpus and is dependent on the language that is being modelled. The probabilities of a sequence occurring are very different from one language to the next[52].

Lexical Modeling: Because it functions as a link between the audio-based acoustic model and the text-based language model, the Lexical Model (Dictionary) is an essential component of any automatic speech recognition (ASR) system[52,56]. An ASR system requires the lexicon to fulfil not one but two distinct functions: 1) It contains a list of all the words that the ASR model is capable of possibly recognising, and 2) It assists traditional HMM models in the process of building decoder models for each phoneme[57]. The dictionary is divided into two sections: the first section contains the words that the ASR system is able to recognise, and the second section contains the phoneme composition that is used to produce those words. In order to get a good

performance out of the ASR system, it is often essential to include all of the possible words and phonemes in the language in the lexicon[57].

Mel Frequency Cepstral Coefficients (MFCC): The Mel Frequency Cepstral Coefficient (MFCC) is a popular feature representation method used for speech signals58. It is based on the concept of Cepstrum, which represents the power of a signal. The steps involved in calculating the MFCCs of a speech signal are as follows:

i. Pre-Emphasis: the first step in the process of MFCC feature extraction is to boost the amount of energy contained in the signal at high frequencies[59]. This is done because the spectrum of voiced segments has less energy at higher frequencies than it does at lower frequencies. The spectrum of a waveform is the summation of sinusoids, each of which has a particular amplitude and phase. This phenomenon is referred to as a Spectral tilt[52,60]. Sounds such as [r], [g], and [j], as well as [b] are voiced, whereas sounds such as [s], [p], and [k] and [t] are unvoiced. The vocal folds vibrate when producing voiced sounds but do not do so when producing unvoiced sounds. This is the primary distinction between the two types of sounds. They are produced from the beginning of the vocal tract the majority of the time[60]. Increasing the amount of energy at high frequencies provides the acoustic model with additional data. This contributes to an improvement in the performance of the phone recognition.

ii. Windowing: Speech is a non-periodic signal, and the properties it possesses shift over the course of time[61]. As a result, information is extracted from a signal region that is small enough so that the speech signal appears relatively stationary. This results in an improvement in the spectral information for phone recognition. Nevertheless, when calculating the Discrete Fourier Transform (DFT), it assumes that the small signal region is one period of a continuous periodic signal[62]. This is

done so in order to speed up the process. Because of their aperiodic nature, speech signals are susceptible to experiencing discontinuities. These discontinuities have the potential to have an effect on the spectrum by manifesting themselves as high frequency components that were absent from the initial signal. Windowing is a strategy that can be utilised to reduce the severity of these effects. The amplitude of the discontinuities that occur at the boundaries of each signal region can be reduced through the use of windowing. The Hamming window is typically used for this purpose because its performance in the calculation of MFCCs is superior to that of the rectangular window. When compared to the main lobe, the results produced by the Hamming window are significantly cleaner and more amenable to frequency-selective analysis because the side lobes are significantly suppressed by the Hamming window[52].

iii. Discrete Fourier Transform: Spectral information is always extracted from the windowed signals utilising Discrete Fourier Transform in order to obtain the energy of the signal at various frequency bands[63].

iv. Mel Filter Bank and Log: The human ear is not sensitive to each frequency band in the same way throughout its range. Its sensitivity to lower frequencies is greater than its sensitivity to higher frequencies. The mel-scale modelling tool can be used to simulate this property[64]. One unit of pitch is called a mel. Because human perception of frequency is more non-linear than linear, the mel-scale has a closer relationship to human hearing than a time-frequency domain representation like the spectrogram does. The mel-scale applies a pre-emphasis that is not applied when the spectrogram is being constructed[64]. This pre-emphasis is applied to the higher frequencies

v. Cepstrum: Speech is produced when the output of the glottal source is passed through the vocal tract, which has a filtering characteristic due to its ability to form a variety of shapes[65]. This results in the production of speech. In order to differentiate the glottal source from the filter, the cepstrum is utilised. Finding the inverse of the discrete Fourier transform (DFT) of the Mel filter bank output66 is the first step in calculating the cepstrum. A typical cepstrum will have a large peak at the fundamental frequency F0, which will represent the glottal pulse. Additionally, a typical cepstrum will have higher harmonic components at lower amplitudes, which will represent the filters in the vocal tract[67]. We are able to distinguish between the source and the filter by utilising the cepstrum values that fall between the second and thirteenth peaks and ignoring the peak that corresponds to the fundamental frequency. The information about the energy is not conveyed by the cepstral coefficients. As a result, an energy component has been added to it. The sum of the squares of the samples taken at various points in time constitutes the energy of the samples taken in any particular frame. Consider a signal x that is windowed from time t1 to time t2 in the following manner: The driving force behind this signal can be found in the form of:

$$E=\sum_{i=t1}^{t2} x\,(i)^2 \qquad 2.10$$

The slope of the formants changes from the stop burst to the release, which results in speech signals that are not constant. Therefore, we plan to include these different iterations of the features. A delta value, also known as a velocity value, is added to each of the 13 cepstral values. A change in the corresponding cepstral or energy feature from one frame to the next corresponds to a change in the delta values. It is possible to calculate it as:

$$dt = \frac{c(t+1)-c(t-1)}{2} \quad 2.11$$

Where d(t) represents the delta value, and c(t) is the cepstral value at time t. Similarly, double delta features are added, which correspond to the change between frames in the corresponding delta values.

2.3 Related Work

The review of the literature reveals that several authors have contributed significantly to the development of speech recognition.

In a study titled "accent classification of the three major Nigerian indigenous languages using 1D CNN LSTM network model" the researchers looked at the three major indigenous languages of Nigeria[68]. The purpose of this paper is to investigate the words in the three major indigenous languages of Nigeria that are most sensitive to accent, and it also makes use of machine learning (ML) to find a solution to the problem of accent classification (AC) in all three languages. Python was utilised in the development and implementation of a speech-based algorithm. Speech data were collected from 300 speakers, and an algorithm called mel-frequency cepstral coefficient (MFCC) was used to extract distinct features. These features are used to differentiate speakers of the three native languages. A combination of a one-dimensional convolutional neural network (also known as a 1D CNN) and a long short-term memory (also known as an LSTM network model) was trained with the acquired speech data (1D CNN LSTM). The findings of the experiment indicate that the accuracy of classification is 94.9%.

Using machine learning and deep learning models' researchers conducted a study on how to classify accents[69]. In order to develop an efficient instrument for accent classification, the authors of this paper developed a model by making use of machine

learning and deep learning algorithms. Validation of the proposed system was performed on a dataset taken from the Speech Accent Archive, which was hosted on Kaggle by George Mason University. According to the findings, when compared to other machine learning models, such as CNN and LSTM, Decision Trees performed the best, achieving an accuracy of 97.36%. This was the case even though CNN and LSTM are considered to be deep learning models.

In the research paper titled "Using Brain-Like Audio Features to Improve Speech Recognition" the authors proposed a new method for the extraction of features. They referred to it as "SHH (spike-H)," and it was designed to be analogous to the human brain. It was able to achieve higher speech recognition rates than earlier methods[70]. After that, the features that were extracted with the help of the proposed model are fed into the classification model. New parallel CRNN model with an attention mechanism that takes into account both temporal and spatial features is proposed here. According to the findings of the experiments, the proposed CRNN is able to achieve an accuracy of 94.8 percent on the Aurora dataset. Experiments on audio similarity have also demonstrated that SHH is better able to differentiate between different audio features. In addition, the experiments involving ablation demonstrate that SHH can be applied to digital speech recognition.

In a separate but related study on the use of machine learning techniques for accent recognition[71]. This study aims to distinguish the American accent from the British accent, as well as the French accent, the German accent, the Italian accent, and the Spanish accent. The task of accent recognition makes use of two distinct algorithms that are based on machine learning: the Support Vector Machine (SVM) and the k-Nearest Neighbor algorithm (K-NN). Both approaches involve optimising the user-defined hyper parameters in order to achieve high levels of precision in the results. In

addition to this, the method of k-fold cross validation is utilised in order to guarantee the accuracy of the findings. According to the findings of the experiments, the accuracy achieved by the SVM equipped with the Radial Basis Function (RBF) kernel is the highest. As a consequence of this, applications that require speaker accent recognition are good candidates for the use of SVM with an RBF kernel.

According to the findings of a study that used audio surveillance to investigate the detection of potential collision hazards for the safety of construction workers[72]. The purpose of this study is to propose the development of a less expensive technology for the prevention of collisions that makes use of audio signals to detect the presence of mobile equipment. The problem is addressed in the study by developing a novel sound detection model that employs artificial intelligence (AI) to detect the sound of collision hazards buried in a great deal of ambient noises. This improves the auditory situational awareness of construction workers who are exposed to loud noises and helps them avoid injuries. This study consisted of three stages: (1) the collection of audio data from construction equipment; (2) the development of an innovative audio-based machine learning model for the automated detection of collision hazards; and (3) the conducting of field experiments to investigate the effectiveness and latency of the system. The findings demonstrated that the model detects equipment accurately and is able to provide timely notification to the workers of potentially dangerous situations.

The goal of this project is to create a hybrid deep CNN-based multi-accent recognition system for the English language.

The database of non-native Indian English speakers' accents is going to be given the name IndicAccentDB in this paper, and it will be innovative and well-structured. The unbalanced dataset (gender bias) and speaker mismatch problems that were seen in the past are addressed by IndicAccentDB through the inclusion of speech samples

from six different states. The work that is being proposed also discusses the requirements for creating the IndicAccentDB database as well as the pre-processing tasks that are going to be performed on the dataset. In addition, in order to construct the reliable Multi-Accent Recognition System, we conducted research and development on various accent classification models, including 1D-CNN, Support Vector Machines, Random Forest, Decision tree, ResNet18, ResNet50, and xResNet18. These models were trained on MFCC and Mel-Spectrogram features (MARS). In the end, we assessed how well the proposed models performed on the novel database and compared the findings by using evaluation metrics such as precision, accuracy, F1-score, and recall. According to the findings we gathered, xResNet18 displayed a high level of accuracy when identifying the various accent classes.

According to the findings of a study titled "Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers[74]." This study provides support for accent-dependent automatic speech recognition by applying a supervised learning algorithm to the task of recognising three Nigerian ethnic groups (Yoruba, Igbo, and Hausa) and distinguishing between them based on their accents. This is accomplished by constructing sequential Mel-Frequency Cepstral Coefficients (MFCC) features from the frames of the audio sample. This research was carried out in Nigeria. According to the findings of our research, an effective method for recognising and categorising accents is to concatenate the MFCC features in a sequential fashion and then apply a supervised learning strategy. This approach achieves both high efficiency and high accuracy in its results.

According to research published in a paper with the title "Language Accent Detection with CNN Using Sparse Data from a Crowd-Sourced Speech Archive"[76].

In this study, the authors made a contribution to the task of accent recognition for a group of up to nine European accents in English. Additionally, they tried to provide some evidence in favour of specific hyper parameter choices for neural network models while also searching for the best input speech signal parameters to improve the baseline accuracy of accent recognition. To be more specific, they utilised a CNN-based model that was trained on the audio features extracted from the Speech Accent Archive dataset. This dataset is a collection of accented speech recordings that was compiled by using contributions from the general public. The author demonstrated that incorporating time–frequency and energy features (such as spectrograms, chromograms, spectral centroids, spectral rolloffs, and fundamental frequencies) into the Mel-frequency cepstral coefficients (MFCC) has the potential to increase the accuracy of the accent classification when compared to the conventional feature sets of MFCC and/or raw spectrograms.

According to the results of the experiments, amplitude mel-spectrograms plotted on a linear scale are the input data type that has the greatest influence on the model[76]. In order to produce state-of-the-art classification results, amplitude mel-spectrograms on a linear scale, which are the correlates of the audio signal energy, are used. This brings the recognition accuracy for English with Germanic, Romance, and Slavic accents ranged from 0.964 to 0.987, thus outperforming existing models of classifying accents that use the Speech Accent Archive. In addition to that, they investigated how the rhythm of speech influences the accuracy of recognition. According to the results of our preliminary research, the audio recordings were utilised for additional accent classification research in their unaltered, unedited form, meaning that all of the pauses were maintained.

A study was done using CNN Model trained on amplitude Mel-Spectrograms to classify speakers of English with accents[77]. The purpose of this paper is to contribute to the advancement of a classification method for accented speech by making use of a CNN-based model that was trained and tested on English with Germanic, Romance, and Slavic accents. The machine learning model's input feature set was investigated in order to locate the optimum combination of time-frequency and energy characteristics of the speech that was being fed into the model. In addition, the authors adjusted the model's hyperparameters as well as the dimensionality of the features that were input. They argued that when compared to conventional feature sets based on MFCCs and raw spectrograms, mel-scale amplitude spectrograms on a liner scale appear to be more effective in accent classification tasks. Even though our models only made use of a small amount of data from the Speech Accent Archive, they were still able to produce cutting-edge classification results for English spoken with Germanic, Romance, and Slavic accents. In comparison to other models that classify accents using the same dataset, the accuracy of our models that were trained on linear scale amplitude mel-spectrograms ranged from 0.964 to 0.987, making them superior to those other models.

According to the findings of a study titled "Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers"[78]. automatic speech recognition was able to accurately identify the accents. This study provides support for accent-dependent automatic speech recognition by applying a supervised learning algorithm to the task of recognising three Nigerian ethnic groups (Yoruba, Igbo, and Hausa) and distinguishing between them based on their accents. This is accomplished by constructing sequential Mel-Frequency Cepstral Coefficients (MFCC) features from the frames of the audio sample. This research was carried out

in Nigeria. According to the findings of our research, an effective method for recognising and categorising accents is to concatenate the MFCC features in a sequential fashion and then apply a supervised learning strategy. This approach achieves both high efficiency and high accuracy in its results.

To investigate an acoustic comparison of the english accents of Malaysian and Nigerian Speakers[79]. This study compares the spectral and cepstral acoustics of speakers of English with a Malaysian accent and a Nigerian accent in order to determine the effect that accents have on the spectral and cepstral features of speech. Researchers at the ARS have paid a lot of attention to accent because it is a major source of ASR performance degradation, which is why they have paid so much attention to it. The vast majority of ASR applications were developed using speech samples from native English speakers, despite the fact that the majority of the people who could use these applications speak English as a second language.

In a study titled "Comparative Phonological Analysis of Varieties of English Spoken by Native Speakers of Nigerian Languages (Hausa, Igbo, Kanuri, and Yoruba) for the Determination of Speakers' Origins[80]." researchers compared the English dialects spoken by native speakers of these Nigerian languages: Hausa, Igbo, Kanuri, and Yoruba. The investigation is broken up into two parts: I the provision of segmental descriptions of four different Nigerian English accents (Yoruba, Hausa, Igbo, and Kanuri); (ii) accent classification experiments to evaluate the relative effectiveness of four different methods in classifying four different Nigerian English accents. A corpus of the four different Nigerian English accents was compiled using the responses of sixty individuals, with each accent being represented by fifteen individuals. Impressionistic analysis, supplemented with a little acoustic corroboration, was performed on the corpus. The task of accent classification included a total of 118

participants, drawn from three human groups: Nigerian non-linguists (each L1 group represented by 20 respondents), 25 Nigerian linguists (6 Hausa, 9 Igbo, 5 Kanuri, and 5 Yoruba), 13 phoneticians based in the United Kingdom, and an automatic accent recognition system called Y-ACCDIST. According to the findings, each of the four approaches has some degree of promise in terms of accent recognition. In spite of this, the overall findings suggest that native speakers, regardless of their linguistic backgrounds, were significantly more accurate in identifying speakers of their own accent groups. According to the findings, the phoneticians based in the United Kingdom as well as Y-ACCDIST were the most accurate in distinguishing Yoruba-English. It is possible to speculate, in light of the fact that Yoruba-English speakers make use of linguistic stereotypes such as [h]-elision and [h]-epenthesis in their speech, that language analysis carried out by linguists who are not native speakers of the language variety in question can be more reliable if the language variety in question contains some stereotypes.

In a study very similar, researchers attempted to recognise English speech using deep learning and multiple features[81]. This study proposes a deep learning speech recognition algorithm that combines speech features and speech attributes. The research object for this study is English speech. First, the deep neural network supervised learning method is used to extract the high-level features of the speech, then the output of the fixed hidden layer is selected as the new speech feature for the newly generated network, and the GMM–HMM acoustic model is trained with the new speech features; second, the speech attribute extractor based on deep neural network is trained for multiple speech attributes, and the extracted speech attributes are classified into phoneme by deep neural network; finally, the deep neural network is used to classify the phone The results of the experiments show that the English

speech recognition algorithm that was proposed is based on a deep neural network with multiple features, and it is able to directly and effectively combine the two methods by combining the speech features and the speech attributes of the speaker in the input layer of the deep neural network. Additionally, it is able to significantly improve the performance of the English speech recognition system.

A survey of CNN-Based network intrusion detection was published[82]. The deep neural network feature fusion method is used in this paper to perform speech recognition in an effective manner. This method is used to effectively fuse the extracted monomodal features. Second, when establishing the architecture of the business English translation system, you should use the edge computing method. In conclusion, the simulation test analysis demonstrates that the business English translation framework developed in this paper is effective in its operation. Our suggestion resulted in an improvement in accuracy that was at least 10% higher than that of the other available methods, and it also resulted in a significant reduction in the amount of time spent building the model. This research aims to discuss how to deal with the many differences that exist between the source language and the target language, how to improve the readability of the translation, and how to meet the cultural cognition and needs of the reader. The purpose of this research is to discuss how to deal with the many differences that exist between the source language and the target language.

In a framework for human-drone interaction that is based on the recognition of multiple languages[83]. This study offers an alternative to using innovative methods known as natural user interfaces, which permit users to interact with drones in an intuitive manner using speech. These methods were developed as a result of this research. However, if there are multiple languages spoken in the same area, it is

possible that the number of users will decrease if there is only one language available for communication. In addition, the noise from the environment and the propellers makes speech recognition a challenging endeavour. The controlling of the movement of drones is intended to be accomplished through the utilisation of a multilingual speech recognition system that is comprised of English, Arabic, and Amazigh. These languages were chosen because of their widespread use across many regions, particularly in the Middle East and North Africa (MENA) region of the world. A strategy consisting of two stages is suggested as the best way to accomplish this objective. A model for multilingual speech recognition that is based on deep learning is developed during the first stage of the process. After that, a quadrotor unmanned aerial vehicle (UAV) is used to test the developed model in real-world conditions. In order to make the network more resilient, it was trained with 38,850 records that contained a mixture of noise and commands as well as unknown words. It has been determined that the overall accuracy of the class is greater than 93%. After that, we put the newly designed system through its paces by having 16 different people participate in a series of experiments in which they gave voice commands to the system. The accuracy that was achieved was approximately 93.76% for the English recognition and 88.55% and 82.31%, respectively, for the Arabic and Amazigh languages. In the end, a quadrotor unmanned aerial vehicle (UAV) was used to create a hardware implementation of the designed system. Tests conducted in real time have demonstrated that the approach possesses a great deal of potential as a substitute mode of human–drone interaction, in addition to providing the benefit of simplified control.

In addition, in a speech recognition implementation where MFCC and DTW algorithms were used for home automation[84]. This research has built a speech

recognition system that can be used as a device to control the devices in a smart home by identifying the commands spoken by users, particularly when the user is in a state of clean speech. The command that is carried out is a string of words that has been chosen in advance. In order to extract voice commands, the MFCC algorithm is used to match spoken words with templates generated by the Dynamic Time Warping (DTW) algorithm. This is done in order to facilitate the extraction of voice commands. The DTW algorithm is able to determine the difference between two time series that spans of time that are distinct from one another. The results of the accuracy of this system were successfully carried out by these algorithms at a rate of 86.67%, with a required time of 5.28 seconds on average to identify the commands.

Using support vector machine to research speaker recognition for digital forensic audio analysis[85]. For the purpose of this investigation, the data collected came in the form of evidence in the form of a recording of a conversation that took place over the phone as well as a recording of a comparison of some unexpected findings. In order to recognise the speaker, the part that has already been completed is the classification of speaker recognition using the Support Vector Machine (SVM) classification method. Excellent accuracy was achieved in the classification of the speaker's introduction when the SVM method was utilised. From the test results, the SVM method's use resulted in an accuracy rate of 86.67% for the test with the same sentence and up to 67% for different sentences to recognize the speaker with the values of C 0.01 and γ (Gamma) 0.0001.

In a separate study titled "Human Emotion Detection with Speech Recognition Using Mel-frequency Cepstral Coefficient and Support Vector Machine[86]." researchers looked at the correlation between human emotions and speech. The Mel-Frequency Cepstral Coefficient was the feature extraction method that was utilised in this

investigation. It was used in conjunction with a human emotion detection system that utilised sound signals (MFCC). This technique was selected because, in comparison to other systems, MFCC comes the closest to simulating the response of the human auditory system. The most recent approach to the classification of data is called Support Vector Machine (SVM), and it was developed in the 1990s by Chervonenkis and Vapnik. SVM is an example of supervised machine learning, which is frequently utilised in numerous studies for the classification of human speech recognition. The RBF kernel from SVM Multi-Class was the one that was utilised the majority of the time in a number of earlier studies. This is due to the fact that SVM makes use of the Radial Basis Function (RBF) kernel, which has improved accuracy. This research found that a frame size of 0.001 seconds, 80 filter banks, [0.3 - 0.7] gamma, and 1.0 C values produced the highest accuracy ratio possible, which was 72.5%.

In a study titled "Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets[87]." the researchers looked at a number of different data sets. In this study, spectrograms and mel-spectrograms are utilised to investigate which method of feature extraction best represents emotions and how significant the differences in effectiveness are in this particular setting. According to the findings of the studies that were carried out, mel-spectrograms are the data type that is best suited for training CNN-based speech emotion recognition systems (SER). The research experiments made use of the following five well-known datasets: The Interactive Emotional Dyadic Motion Capture, the Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), the Surrey Audio-Visual Expressed Emotion (SAVEE), the Toronto Emotional Speech Set (TESS), and the Toronto Emotional Speech Set (TESS) are some of the datasets that were used in this study (IEMOCAP). There were six

distinct categories of feelings that were used: joy, anger, sadness, fear, and disgust, in addition to neutral. On the other hand, due to the characteristics of the IEMOCAP dataset, some of the experiments were designed to only recognise four of the available emotions. A comparison of the effectiveness of classification across the various datasets was also carried out, along with an attempt to develop a universal model that could be trained using any and all datasets. When attempting to identify all four feelings, this strategy resulted in an accuracy of 55.89%. On a total of four datasets, the most accurate model for recognising six emotions was trained, and it achieved 57.42% accuracy after being evaluated (CREMA-D, RAVDESS, SAVEE, TESS). Furthermore, another study was developed that demonstrated that improper data division for training and test sets has a significant influence on the test accuracy of CNNs. This was demonstrated by the fact that the study was able to show the relationship between the two. As a result, a significant amount of work was put into resolving the issue of improper data division between the training and test sets, which was found to have an impact on the results of studies that were previously published. The experiments that were carried out used the well-known ResNet18 architecture in order to demonstrate the dependability of the research results and to show that these issues are not specific to the custom CNN architecture that was proposed in the experiments. After that, the correctness of the labels in the CREMA-D dataset was investigated with the help of a pre-made questionnaire.

Research into approximation techniques for the fast Fourier transform (FFT), with an eye toward their use in speech recognition[88]. Several different approximative designs for computing the FFT are presented in this article. Adjusting the word length at each stage of computation allows for a balance to be struck between the two competing priorities of accuracy and performance. Two different algorithms for modifying word

length while maintaining a predetermined error margin are proposed. The first algorithm aims for performance, so it can achieve a higher operating frequency. The second algorithm aims for performance, so it can achieve an approximate FFT for an area-limited design compared to the conventional fixed design. Both of the proposed algorithms demonstrate that it is possible to strike a productive balance at stage-level between the hardware utilisation and performance requirements. The proposed designs for an approximate FFT are implemented on an FPGA, and the results of the experiments show that the hardware utilisation can be reduced by at least nearly 40% when using the first approximate algorithm. The performance of the designs is improved by more than 20% thanks to the second algorithm. When compared to a coarse design, a fine granularity architecture can cut the amount of FPGA resources needed for a 256-point FFT computation by almost ten percent. This architecture is being researched as well. Finally, the proposed approximate designs are applied to a feature extraction module in an isolated word recognition system. This results in a reduction in the number of LUTs and FFs for the Mel frequency cepstrum coefficients (MFCC) extraction module by up to 47.2% and 39.0%, respectively, as well as a reduction in power of up to 27.0%, all while maintaining an accuracy of less than 2%.

To investigate a speech recognition system that makes use of an improved mel frequency cepstral coefficient along with windowing and framing method[89]. In recent years, a great number of speech recognition systems have been developed in order to solve a variety of problems that have arisen in applications that are used in the real world. A novel speech recognition system that combines enhanced mel frequency cepstral coefficient with windowing and framing method is one of the systems that we have proposed. In order to get rid of the Gaussian white noise that is present in the

input speech signal, a method called windowing and framing is utilised. The nonnegative matrix factorization algorithm is utilised in an efficient manner by the de-noising block for the purpose of factorising the Mel-magnitude spectra of the noisy input audio signal. In addition, the mel-frequency cepstral coefficients, also known as MFCC, are utilised in order to locate the more significant characteristics that are present in the speech signal. Last but not least, the Laplace smoothing technique serves as the language model for the purpose of recognising the audio signals. The proposed Mel frequency cepstral coefficient with Windowing and Framing-based speech recognition system is demonstrated with the help of the MATLAB software. The wavelet-based feature extraction method and the artificial neural network-based method are two other methods for speech recognition that we have evaluated and contrasted with the speech recognition system that we have proposed. The results of the experiments demonstrated that the proposed Mel frequency cepstral coefficient with windowing and framing based speech recognition system performed very well.

Using a neural network and a pre-processing technique for speech recognition in the Mongolian language [90]. In this study, the authors developed a neural network model that is capable of recognising a small selection of Mongolian words. This research was published in the journal "Neural Networks." In the Mongolian language, we have selected these four words. These words were selected in order to facilitate the further development and production of a unique device that features an audio interface. In this experiment, we utilised audio recordings that were captured in a computer by means of a microphone in a typical audience with a minimum amount of ambient noise. The audio recording database that was used to train the neural network is comprised of the speeches of 11 different individuals (7 men and 4 women). One of

them is between the ages of 20 and 30, three others are between 60 and 70, and the remaining two are between 30 and 40. The work is carried out on a standard desktop computer equipped with an Intel Core i5 processor from the third generation and 8 gigabytes of DDR IV memory.

In a similar vein, on emotional speech recognition based on weighted distance optimization system[91]. The researchers presented a number of models to recognise human emotion based on the speaker's words. The Gaussian mixture model is one of the models that has gained a lot of attention (GMM). When there are a large number of features and some of those features are correlated, the GMM may have one or more of its components as ill-conditioned or singular covariance matrices. This can happen when the number of features is high. For the purpose of recognising emotional speech, a brand-new system that is predicated on weighted distance optimization (WDO) has been developed as part of this investigation. The primary goal of the WDO system, also known as WDOS, is to improve recognition accuracy while simultaneously addressing the limitations of the GMM. We discovered that WDOS has had a great deal of success by conducting a comparative analysis of all of the different emotional states as well as the characteristics of each individual emotional state. The accuracy of WDOS's performance for the Japanese language is exceptional, coming in at 86.03%. When compared with GMM and k-mean, the accuracy of Japanese emotion recognition is improved by 18.43% using this method.

In order to better understand pattern mining as an approach to improving speech emotion recognition[92]. The authors of this paper present a novel method that can be used to extract a new set of high-level features that can be used for the classification of emotions. For this purpose, the author first reduce the dimension of discrete-time speech signals, then we perform a quantization operation on the new signals and

assign a distinct symbol to each quantization level, then we use the symbol sequences representing the signals to extract discriminative patterns that are capable of distinguishing different emotions from one another, and finally, we generate a separate set of features for each emotion based on the patterns that were extracted from the signals. The results of the experiments show that pattern features perform better than Energy features, Voicing features, MFCC features, Spectral features, and RASTA features. They also show that combining pattern-based features with acoustic features results in an improvement in classification performance that is even greater than the previous improvement.

In a study on application of machine learning to speech analysis with the purpose of identifying speaker accents within the context of audio forensics deep learning and machine learning frameworks are being used to help with forensic investigations[93]. The work that was done to recognise the speaker's accent utilised support vector machine, k nearest neighbours, XG boost, linear discriminant analysis, quadratic discriminant analysis, and decision tree algorithms. The experiment that was carried out on the dataset for accent recognition demonstrated that Mel Frequency Cepstral Coefficients and the kNN classifier are capable of identifying and differentiating six accents that belong to different speakers with greater accuracy.

According to the findings of a study titled "Arabic Speech Recognition by Stationary Bionic Wavelet Transform and MFCC Using a Multi-layer Perceptron for Voice Control[94]." The researchers were able to successfully recognise Arabic speech. In this chapter, we will go into detail about our approach to the recognition of Arabic speech using a mono-locutor and a reduced vocabulary, both of which are introduced in the literature. This strategy involves using our proper speech database, which is comprised of Arabic speech words and was recorded by a single interlocutor, as the

first step in the process of producing a voice command. The next thing that needs to be done is to pull characteristics from the words that were recorded. The third step is to perform a classification of those features that were extracted. In order to accomplish this extraction, the first thing that is done is to apply the stationary bionic wavelet transform, abbreviated as SBWT, to each recorded word. After that, the Mel Frequency Cepstral Coefficients, also known as MFCCs, are computed using the vector that is obtained by concatenating the stationary bionic wavelet coefficients that were previously obtained. After that, the MFCCs that were obtained are concatenated for the purpose of constructing one input of a multi-layer perceptron (MLP), which is utilised for the feature classification. We have used ten different Arabic words throughout the phases of learning and testing the MLP that was used, and each of those words has been repeated 25 times by the same voice. The effectiveness of the proposed method was evaluated using a simulation programme, which revealed that the proposed approach had a classification rate of 98%.

In addition to this, in research done on a robust automatic speech recognition system based on deep learning for the Hindi language[95]. In this article, a seven-layer, one-dimensional convolutional neural network called HindiSpeech-Net is proposed with the goal of recognising various speech samples of the Hindi language in their appropriate categories. A large dataset consisting of 2400 speech samples in the Hindi language were collected in ten distinct classes under real-world conditions. This was followed by signal filtering and augmentation in order to improve the dataset for the purpose of producing a robust model and to prevent overfitting. The gathered dataset was then evaluated based on a variety of performance parameters, and the results were separated into a training set, a validation set, and a test set. On the test set, the trained HindiSpeech-Net model achieved an accuracy of 92.92%. The proposed framework is

more efficient from a computational cost standpoint, operates in real time, and is amenable to incorporation into embedded systems.

In a separate but related study, researchers developed an intelligent voice recognition system utilising fuzzy logic and the bag-of-words technique[96]. In this paper, a method for recognising voice commands based on a fuzzy logic system that is capable of perceiving fuzzy commands, also known as commands containing fuzzy terms such as 'close,' 'closer,' 'close to,' 'closer than,' and 'very far,' is described. The method is based on a fuzzy logic system. The developed method has the capability of being customised for the needs of a particular user. In order to increase the expressiveness of the language used for the control of a moving robot, a system of fuzzy logic has been developed. This system is used to recognise commands that are linguistically incorrect.

Specifically, in the article titled "Approximate designs for fast Fourier transform (FFT) with application to speech recognition[97]." Several different approximative designs for computing the FFT are presented in this article. Adjusting the word length at each stage of computation allows for a balance to be struck between the two competing priorities of accuracy and performance. Two different algorithms for modifying word length while maintaining a predetermined error margin are proposed. The first algorithm aims for performance, so it can achieve a higher operating frequency. The second algorithm aims for performance, so it can achieve an approximate FFT for an area-limited design compared to the conventional fixed design. Both of the proposed algorithms demonstrate that it is possible to strike a productive balance at stage-level between the hardware utilisation and performance requirements. The proposed designs for an approximate FFT are implemented on an FPGA, and the results of the experiments show that the hardware utilisation can be reduced by at least nearly 40%

when using the first approximate algorithm. The performance of the designs is improved by more than 20% thanks to the second algorithm. When compared to a coarse design, a fine granularity architecture can cut the amount of FPGA resources needed for a 256-point FFT computation by almost ten percent. This architecture is being researched as well. Finally, the proposed approximate designs are applied to a feature extraction module in an isolated word recognition system. This results in a reduction in the number of LUTs and FFs for the Mel frequency cepstrum coefficients (MFCC) extraction module by up to 47.2% and 39.0%, respectively, as well as a reduction in power of up to 27.0%, all while maintaining an accuracy of less than 2%. According to the findings of a study on low-power speech keyword recognition using precision adaptive MFCC based on R2SDF-FFT and approximate computing[98]. The purpose of this paper is to present system-architecture-circuits co-designs for the purpose of extracting MFCC features for speech keyword recognition. By combining the 8-stage radix-2 single-path delay feedback FFT (R2SDF-FFT) and the precision self-adaptive architecture with approximate computing, it is possible to strike a balance between the level of accuracy achieved and the amount of power consumed, regardless of the type of background noise present. The R2SDF-FFT structure, when combined with fine-grained bit-width quantization, has the potential to reduce memory size by 35.7%. In order to further improve the FFT computing energy efficiency, it has been suggested that approximate multiplication and addition with Dual-Vdd be used. Finally, the author presentented the precision self-adaptive MFCC architecture with the proposed FFT. This architecture can be dynamically configured to use two calculation modes with different hardware settings depending on the input speech background noise. This allows the architecture to be more flexible. The power consumption of the proposed design is able to be reduced by up to 76.3% when it is

implemented and evaluated using 22 nm technology, while simultaneously increasing by 0.8% in terms of accuracy.

In a paper on analysis of Yorùbá Automatic Speech Recognition. A brief review of research progress on Yorùbá Automatic Speech Recognition (ASR) is presented. The focus of this review is on variability as a factor contributing to the performance gap between HSR and ASR. The purpose of this paper is to x-ray the advances recorded, major obstacles, and chart a way forward for the development of ASR for Yorùbá that is comparable to those of other tone languages and of developed nations. This is accomplished by conducting exhaustive literature reviews on ASR, with a primary focus on Yorùbá. Although significant headway has been made in the development of ASR in the developed world, this is not the case for the majority of developing nations, particularly those in Africa. Yorùbá, along with the vast majority of languages spoken in Africa, does not have the human or material resources necessary for the development of a functional ASR system, let alone for reaping the potential benefits of using such a system. According to the findings, achieving the ultimate goal of having ASR performance comparable to that of humans requires an in-depth understanding of the factors that contribute to variability[99].

In a research paper on the development of an automatic speech recognition system for tonal languages[100]. This study presents the results of a comprehensive survey on Automatic Speech Recognition (ASR) for tonal languages that are spoken in different parts of the world. There is a discussion on the tonal languages of the Asian, Indo-European, and African continents; however, the tonal languages of the American and Australasian continents are not discussed. The presentation of the work done in previous years on the ASR of tonal languages from different continents, including Asian continent tonal languages such as Chinese, Thai, Vietnamese, Mandarin, Mizo,

and Bodo; Indo-European continent tonal languages such as Punjabi; Indo-European continent tonal languages such as Swedish; and African continent tonal languages such as Yoruba and Hausa; and Indo-European continent tonal languages such as Swedish; and African continent tonal languages such In conclusion, the findings are used to guide an examination of the synthesis analysis. Discussion centres on a wide range of concerns and difficulties connected with tonal languages. It has been observed that a significant amount of work has been done for the tonal languages of the Asian continent, such as Chinese, Thai, Vietnamese, and Mandarin. On the other hand, very little work has been reported for the tonal languages of the Mizo and Bodo languages, Indo-European languages such as Punjabi, Latvian, and Lithuanian, and the tonal languages of the African continent, such as Hausa and Yoruba.

According to the results of a survey on automatic speech recognition technology for tonal languages[101]. This study presents the results of a comprehensive survey on Automatic Speech Recognition (ASR) for tonal languages that are spoken in different parts of the world. There is a discussion on the tonal languages of the Asian, Indo-European, and African continents; however, the tonal languages of the American and Australasian continents are not discussed. The presentation of the work done in previous years on the ASR of tonal languages from different continents, including Asian continent tonal languages such as Chinese, Thai, Vietnamese, Mandarin, Mizo, and Bodo; Indo-European continent tonal languages such as Punjabi; Indo-European continent tonal languages such as Swedish; and African continent tonal languages such as Yoruba and Hausa; and Indo-European continent tonal languages such as Swedish; and African continent tonal languages such In conclusion, the findings are used to guide an examination of the synthesis analysis. Discussion centres on a wide range of concerns and difficulties connected with tonal languages. It has been

observed that a significant amount of work has been done for the tonal languages of the Asian continent, such as Chinese, Thai, Vietnamese, and Mandarin. On the other hand, very little work has been reported for the tonal languages of the Mizo and Bodo languages, Indo-European languages such as Punjabi, Latvian, and Lithuanian, and the tonal languages of the African continent, such as Hausa and Yoruba.

In a study on creating a free and accessible digital archive of Yoruba speech[102]. This article presents an open-source speech dataset for the Yoruba language, which is one of the most widely spoken low-resource languages in West Africa and is understood by at least 22 million people. In addition to being one of the official languages of Nigeria, Benin, and Togo, Yoruba is also spoken in a number of other countries in Africa and even further afield. The corpus is comprised of over four hours' worth of recordings made at 48 kHz by 36 different male and female volunteers, along with the transcriptions of those recordings that include disfluency annotation. Full diacritization has been applied to the transcriptions, which is an essential component for accurate pronunciation and lexical disambiguation. The annotated speech dataset that is described in this paper is primarily intended for use in text-to-speech systems, as well as serve as adaptation data in automatic speech recognition and speech-to-speech translation, and provide insights in West African corpus linguistics.

In a study titled "Development of Acoustic Models and Language Models[103]" it was found that the first automatic Fongbe continuous speech recognition system was developed. In this paper, we report on our efforts to develop an ASR system for a new language that receives inadequate resources (Fongbe). The development of acoustic models and language models for continuous speech decoding in Fongbe is the purpose of this body of work. The problem that arises when attempting to use an ASR system to translate from the African language of Fongbe, which is spoken primarily in the

countries of Benin, Togo, and Nigeria, is that Fongbe does not have any language resources available. As a first step in this project, we have compiled Fongbe text and speech corpora, which are explained in more detail in the following sections. Modeling of acoustics has reached the graphemic level, and language modelling has resulted in the creation of two language models that can be compared in terms of their levels of performance. In order to investigate the effect that tone diacritics have on the language models, we also simplified vowels by removing them. This was done by removing the tones.

Tone transcription has been shown to improve ASR performance in languages with extremely limited resources[104]. The Chibchan language Bribri, which is spoken in Costa Rica and is considered to be a language with very few resources available, is used as an example in this paper to demonstrate a systematic comparison of various transcription styles. The models that are the most successful split the tone from the vowel in order to allow the ASR algorithms to learn tone patterns independently. The character error rate (CER) showed improvements ranging from 4% to 25% with these models, while the word error rate showed improvements ranging from 3% to 23%. (WER). This is true for the more traditional GMM/HMM algorithms as well as the end-to-end CTC ones. Additionally, the first attempt at training ASR models for Bribri is presented in this paper. The models with the best overall performance had a CER of 33% and a WER of 50%. These models were trained on only 68 minutes of data, and as a result, they demonstrate the potential of ASR to generate additional training materials, as well as to assist in the documentation and revitalization of the language. This is despite the fact that using hand-engineered representations is a disadvantage.

In the case of the Hindi language, a robust automatic speech recognition system that is based on deep learning was developed[105]. In this article, a seven-layer, one-dimensional convolutional neural network called HindiSpeech-Net is proposed with the goal of recognising various speech samples of the Hindi language in their appropriate categories. A large dataset consisting of 2400 speech samples in the Hindi language were collected in ten distinct classes under real-world conditions. This was followed by signal filtering and augmentation in order to improve the dataset for the purpose of producing a robust model and to prevent overfitting. The gathered dataset was then evaluated based on a variety of performance parameters, and the results were separated into a training set, a validation set, and a test set. On the test set, the trained HindiSpeech-Net model achieved an accuracy of 92.92%. The proposed framework is more efficient from a computational cost standpoint, operates in real time, and is amenable to incorporation into embedded systems.

In a study on the recognition of emotions in spoken Chinese based on a combination of deep neural networks and acoustic features[106]. The primary objective of this research is to develop a touch-sensitive interface that can be converted into a voice-operated interface for use with AI system service robots or smart home voice assistants as they become increasingly popular. A Deep Neural Network (DNN) model that was specifically designed for the purpose of developing a Chinese speech emotion recognition system was proposed in this research. Within the scope of this investigation, the proposed model's training attributes are comprised of 29 acoustic characteristics derived from acoustic theory. A number of different audio adjustment methods, such as waveform adjustment, pitch adjustment, and pre-emphasize, are proposed as a result of this research in order to amplify datasets and improve the accuracy of training. This research was successful in achieving an accuracy of 88.9%,

on average, in the recognition of emotions in the CASIA Chinese sentiment corpus. The investigation's findings indicate that the deep learning model and audio adjustment method proposed in this study are capable of accurately identifying the sentiments conveyed in short Chinese sentences, and that they have the potential to be implemented in Chinese voice assistants or combined with other dialogue applications.

In the use of Tonal Non-Tonal Classifier for the Recognition of Children's Punjabi Speech to improve the word recognition rate for tonal languages, the research presented here classifies speech into tonal and non-tonal components[107]. This is done in order to retrieve the prosodic features. A feature that is related to pitch is known as a prosodic feature, and it is capable of effectively understanding the tone of the word and has a high level of accuracy when recognising words. Prosodic features that have been extracted are fed to the ASR system one at a time, and only later are they combined. The results of MFCC-based automatic speech recognition (ASR) are compared with the results of combining prosodic features with Mel Frequency Cepstral Coefficients (MFCC). When the performance of the system is analysed, it is discovered that the WER can be decreased by incorporating prosodic features, resulting in a 14% increase in the system's Relative Improvement (RI). The results are compared to the result at the beginning of the study, and an analysis of the system's performance is carried out.

In the study titled "Voice Conversion Based Augmentation and a Hybrid CNN-LSTM Model for Improving Speaker-Independent Keyword Recognition on Limited Datasets," researchers looked at ways to improve speaker-independent keyword recognition using limited data sets. In this paper, we proposed a voice conversion (VC) - based augmentation to increase the limited training dataset as well as a fusion of a

convolutional neural network (CNN) and a long-short term memory (LSTM) model for robust speaker-independent isolated keyword recognition. Both of these augmentations are intended to increase the size of the training dataset. A cumbersome and time-consuming task, speaker-independent speech recognition requires a sufficient amount of voice data to be collected and prepared in advance. An auxiliary classifier conditional variational autoencoder (ACVAE) method was utilised by the authors in order to circumvent this challenge. This resulted in the generation of new raw voices derived from the original voices. The accuracy of the linguistic content was maintained through the use of parallel VC. They used robust deep neural network algorithms to evaluate the performance of the suggested voice conversion enhancement techniques. The baseline for this analysis was determined to be the original training data, excluding any voice that was generated using other data augmentation and regularisation techniques. Incorporating voice conversion augmentation into the baseline augmentation techniques and applying the CNN-LSTM model were shown to improve the accuracy of isolated keyword recognition in the study's findings[108].

In the publication entitled "Computational sociophonetics utilising automatic speech recognition[109]." A preliminary attempt to apply Universal Dependencies to Bribri, an Indigenous language from Costa Rica that is a member of the Chibchan family, is presented in this paper. Due to the lack of previous work on Bribri NLP, a suggestion was made for a dependency parser, and a listing of structures that were difficult to parse was also provided (eg flexible word order, verbal sequences, arguments of intransitive verbs and mismatches between the tense systems of Bribri and UD). Issues with tokenization, data normalisation, and the training of tools like POS taggers, which are necessary for the parsing, are some of the challenges that the authors listed

as some of the difficulties associated with performing NLP with an extremely low-resource Indigenous language. Other difficulties include: We gathered a total of 150 sentences (760 words) from resources that are readily available to the public, such as grammar books and corpora. After that, a context-free grammar was utilised for the initial parse, and the headfloating algorithm described in Xia and Palmer (2001) was used in order to automatically generate dependency parses.

During the process of building an automatic speech recognition system for the documentation of Cook Islands Mori[110]. The process of data processing and training an automatic speech recognition (ASR) system for Cook Islands Mori (CIM) is discussed in this paper. Cook Islands Mori is an Indigenous language that is spoken by approximately 22,000 people in the South Pacific. As part of the research, the authors prepared two experiments and transcribed four hours of speech from native adults and seniors who spoke the language. First, they trained three different automatic speech recognition (ASR) systems; one of them was statistical and was called Kaldi, and the other two were based on Deep Learning and were called DeepSpeech and XLSR-Wav2Vec2. The character error rate was the same for both Wav2Vec2 and Kaldi (CER=61), but the word error rate for Wav2Vec2 was significantly higher (WER=232) than the word error rate for Kaldi (WER=182). This provides evidence that Deep Learning ASR systems are reaching the performance of statistical methods on small datasets, as well as that they are capable of working effectively with extremely low-resource Indigenous languages such as CIM. In the second experiment, the authors trained models using Wav2Vec2 with speakers that were held out of the way. Although the performance was worse than before (CER=157, WER=4616), the system still demonstrated a significant amount of learning.

The development of an acoustic model for automatic speech recognition in the Hausa language is the primary focus of this research. An acoustic model for the Hausa language is going to be designed and developed as part of this project's goals. The Hausa speech corpus database is utilised to compile a word-level phonemes dataset. This is then used to accomplish the goal. The next step is to put an algorithm for acoustic modelling that uses deep learning into action. The accuracy of the model was determined to be 83% thanks to the use of a Convolutional Neural Network in its construction[110].

During the process of developing acoustic models and language models [111]. In this paper, we report on our efforts to develop an ASR system for a new language that receives inadequate resources (Fongbe). The development of acoustic models and language models for continuous speech decoding in Fongbe is the purpose of this body of work. The problem that arises when attempting to use an ASR system to translate from the African language of Fongbe, which is spoken primarily in the countries of Benin, Togo, and Nigeria, is that Fongbe does not have any language resources available. As a first step in this project, we have compiled Fongbe text and speech corpora, which are explained in more detail in the following sections. Modeling of acoustics has reached the graphemic level, and language modelling has resulted in the creation of two language models that can be compared in terms of their levels of performance. In order to investigate the effect that tone diacritics have on the language models, we also simplified vowels by removing them. This was done by removing the tones.

In a study that was very similar to this one, the authors used harmonic pitch to classify the accents of native and non-native children[112]. Using harmonic pitch estimation and Mel Frequency Cepstral Coefficients (MFCCs) to train the k-Nearest Neighbor (k-NN)

classifier, the purpose of this paper is to extract reliable acoustic and prosodic speech cues of accent in order to classify native and non-native preschool children. This will be accomplished by classifying the children based on their native language. The findings of the experiments show that the proposed robust model performs better than a number of different feature extractors in the classification of native and non-native children's accents in terms of accuracy and F-measure, and it is also better at discriminating against environments with a lot of background noise.

In addition, there was research done on audio enhancement for non-native children's speech recognition using discriminative learning[113]. The primary purpose of this investigation is to create a non-native children's speech recognition system that is built on top of feature-space discriminative models, such as feature-space maximum mutual information (fMMI) and boosted feature-space maximum mutual information (fbMMI). An effective performance can be achieved by utilising the collaborative power of speed perturbation-based data augmentation on the original children's speech corpora. In order to investigate the effect that children who are not native speakers of a language have on the accuracy of speech recognition systems, the corpus focuses on the various ways in which children express themselves when they speak, including both read speech and spontaneous speech. According to the results of the experiments, feature-space MMI models with steadily increasing speed perturbation factors perform significantly better than traditional ASR baseline models.

Within the paper entitled "Acoustic Nudging-Based Model for Vocabulary Reformulation in Continuous Yorùbá Speech Recognition"[114]. This research paper presents a model for reformulating the persistence of automatic speech recognition errors that involve the user's acoustic irrational behaviour and distortion of speech

recognition accuracy. The model is based on acoustic nudging, which is described in the previous sentence. Gaussian mixture model (GMM) was helpful in achieving better accuracy and system performance in Yorùbá language translation projects by addressing the low-resource nature of the language. According to the results that were put into action, it was found that the acoustic nudging-based model that had been proposed led to an improvement in accuracy as well as system performance when measured by Word Error Rate (WER), validation, testing, and training accuracy. When compared to previously established models, the evaluation findings indicated that the mean WER was 4.723%. When compared to earlier models developed using GMM (1.1% error rate), GMM-HMM (0.5%), CNN (0.8%), and DNN (1.4% error rate), this approach achieves a lower error rate. As a result, the work that was done was successful in locating a basis for advancing the current understanding of under-resourced languages and developing a model that is accurate and precise with regard to speech recognition.

The Development of a Reliable Speech-to-Text Algorithm for Nigerian Speakers of English[115]. This article presents a STT algorithm that is able to withstand the heavy accent used by speakers of English in Nigeria. During the stage of the project known as "data acquisition," approximately 27,000 isolated speech samples were gathered from five different ethnic groups. Thirty percent of the samples came from Yoruba speakers, twenty-nine percent came from Hausa speakers, twenty percent came from Igbo speakers, and the remaining twenty-one percent came from Fulani and Ijaw speakers. After the data had been preprocessed, the features were extracted with the help of the Mel-Frequency Cepstral Coefficients (MFCCs), which had 13 coefficients. The Hidden Markov Model (HMM) with a variable number of states was selected as the method of recognition that was utilised. The findings of this research indicate that

an average accuracy of 86% was accomplished for the 10-word vocabulary isolated speech that was taken into consideration with the number of states equal to 5. In addition, an average accuracy of 86% was accomplished by using a vocabulary of 10 words and a number of HMM states equal to 7. An accuracy rate of 90% was achieved on average for the 10-word vocabulary with the number of states set to 9.

In the course of an in-depth analysis of the published research on Hausa Natural Language Processing[116]. In this research paper, using a keyword index and article title search, we present a systematic analysis of the current literature that is applicable to HNLP that can be found in the Google Scholar database from the year 2015 to the year June 2020. Only a few research papers on HNLP research have been published recently. These papers focus on areas such as part-of-speech tagging (POS), name entity recognition (NER), words embedding, speech recognition, and machine translation. This is because Natural Language Processing (NLP) relies on a substantial quantity of data that has been annotated by humans in order to train intelligent models. As a result of the substantial amount of research that has been conducted on NLP in English and other languages, researchers are now becoming interested in HNLP. The primary goals of this paper are to encourage research, to identify possible subject areas for additional research in the HNLP, and to offer assistance to researchers in the process of developing additional examinations for studies that are pertinent to the field.

In a separate but related study on the identification of accents among Nigerians of varying ethnic backgrounds[117.] This study provides support for accent-dependent automatic speech recognition by applying a supervised learning algorithm to the task of recognising three Nigerian ethnic groups (Yoruba, Igbo, and Hausa) and distinguishing between them based on their accents. This is accomplished by

constructing sequential Mel-Frequency Cepstral Coefficients (MFCC) features from the frames of the audio sample. This research was carried out in Nigeria. According to the findings of our research, an effective method for recognising and categorising accents is to concatenate the MFCC features in a sequential fashion and then apply a supervised learning strategy. This approach achieves both high efficiency and high accuracy in its results.

Assessing Hausa large vocabulary continuous speech recognition[118]. The authors of this study investigated and developed a Large Vocabulary Continuous Speech Recognition (LVCSR) system that is compatible with the Hausa language as part of this research project. In this article, we provide an overview of the Hausa language and speech database that was recently compiled as a part of our GlobalPhone corpus. They were able to make significant advancements by automatically replacing pronunciation dictionary entries that were inconsistent or flawed, including information on tone and vowel length, utilising state-of-the-art techniques for acoustic modelling, and crawling large quantities of text material from the internet for language modelling. When applied to read newspaper speech, a system that combines the best grapheme- and phoneme-based 2-pass systems achieves a word error rate of 13.16% on the development set and 16.26% on the test set.

Within the scope of this study, the effect of pitch enhancement in Punjabi children's speech recognition system was investigated under a variety of different acoustic conditions[119]. A Punjabi children's speech recognition system is developed in this work using a variety of acoustic matching and mismatching conditions. One major problem in children's speech recognition is the differences in the acoustic attributes of the children and adult speech signals, which leads to the poor recognition rate for the children's speech. This paper demonstrates how pitch enhanced features extracted

from the front-end feature extraction process play an important role under mismatched acoustic conditions. The recognition rate of the children's speech recognition system using different age group datasets increases after enhancing the pitch with the Cepstral analysis in the feature extraction process. This is in comparison to the normal acoustics features extracted using the Mel Frequency Cepstral Coefficient (MFCC) feature extraction process. Kaldi toolkit is used for building the children's speech recognition models at different phoneme levels. Results show the improvement of 0.03% to 16.47% WER under different acoustic conditions using pitch enhanced features.

In a work on Automatic speaker verification systems and spoof detection techniques: review and analysis[120]. Reviewing and evaluating the significant advances that have been suggested by a variety of researchers and scientists is the purpose of this particular piece of writing. The frontend of these systems is designed using a variety of feature extraction strategies, including some that are more traditional, such as autoregressive and cepstral, as well as some that are more modern and based on deep learning. The backend of an ASV system is where the extracted features are learned and classified. This can be done using traditional machine learning models or deep learning models, both of which are also the primary focus of the review that is being presented. Since the advent of practical work in this field, experimental studies have adopted the practice of continuously modifying datasets and evaluation measures in order to develop reliable systems. The majority of the contributing spoofed speech datasets and evaluation protocols are dissected and analyzed here. Spoofing attacks on ASV systems can come in a variety of forms, including speech synthesis (SS), voice conversion (VC), replay, mimicry, and twins. This work enables the defense mechanism of ASV by providing knowledge of techniques for the generation of

attacks similar to those being targeted. The results of this survey herald the beginning of a new era in the development of ASV systems and highlight the beginning of a new generation (G4) in the methods of SS attack development. The paper makes its best efforts to give newcomers to this area the complete idea of ASV and also sheds some light on some of the spoofing attacks that can be targeted during implementation of the future ASV systems. With the increase in advancement of deep learning techniques, the paper also makes its best efforts to give newcomers to this area the complete idea of ASV.

An empirical comparison of deep learning and traditional classifiers for speaker verification in emotional talking environments is presented in information[121]. In this paper, an empirical comparison of the performances of traditional classifiers, such as the Gaussian Mixture Model (GMM), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Artificial neural networks (ANN), and deep learning classifiers, such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU), in addition to the ivector approach, is conducted for a text-independent speaker The deep models undergo hyperparameter tuning using the Grid Search optimization algorithm. The models are trained and validated using a private Arabic Emirati Speech Database, a public Crowd-Sourced Emotional Multimodal Actors (CREMA) database, and a Ryerson Audio–Visual Database of Emotional Speech and Song dataset (RAVDESS) database. The results of experiments show that deep learning architectures do not automatically perform better than traditional classifiers. In point of fact, evaluation was carried out utilising Area Under the Curve (AUC) scores in addition to Equal Error Rate (EER) scores. According to the findings, the GMM model produces the lowest EER values and the best AUC scores across all datasets among the classical classifiers. This is true

regardless of the dataset used. In addition, the ivector model surpasses all the fine-tuned deep models (CNN, LSTM, and GRU) based on both evaluation metrics in the neutral, as well as the emotional speech. In addition, the GMM outperforms the ivector using the Emirati and RAVDESS databases.

In a further study on the detection of multitasking synthetic speech on Indian Languages, the authors found that. This study's objective is to evaluate the efficacy of various methods for detecting synthetic spoofing using a multilingual, loosely constrained Indian language set as its test subject. This paper aims to achieve a multitasking spoofing detection by identifying real/spoof utterance identification as well as the regional language spoofing attack vector. Specifically, this will be done by using a combination of real and spoof utterances. In order to accomplish this, the features and the classifiers that are best candidates for the detection of synthetic spoofing and the identification of language are selected in the appropriate manner. Our methodology compares the performances of three main different classifiers GMM, SVM, DNN on the vector formulated from the accumulation of MFCC features. Hindi, Malayalam, Tamil, Telugu are the four languages which are taken into account for the comparison. It has been discovered that, out of all of these classifiers, SVM and DNN produce the best results, with EER rates of 1.98% and 1.19% respectively[122].

In a study of the emotional information acoustic characteristics of synthetic speech phoneme/ei[123]. This paper, the author compared the differences of acoustic characteristics between synthetic speech and emotional speech under the same text from the perspective of the lack of emotional expression that is present in synthetic speech by using Praat software for a single phoneme /ei/. When the results were analysed, it was determined that the differences in the emotional information were primarily in the small dispersion of synthetic speech fundamental frequency, the

dispersion of synthetic speech intensity was much smaller than that of real speech with large emotional fluctuations, the harmonic waves in narrowband spectrograms were nearly straight without bending and jittering, and the formant centre frequencies interlacing degree is small. These were the conclusions that were reached as a result of the analysis. The absence of harmonic waves at frequencies above 3000 Hz and the obvious difference in the direction of the tail end of the second formant are two examples of the common differences between synthetic speech and neutral and emotional speech. The absence of harmonic waves at these frequencies is one of the hallmarks of synthetic speech.

According to the findings of a study on spectral warping and data augmentation for low resource language ASR systems operating under mismatched conditions[124]. The proposed work in this paper tries to address the both challenges i.e. acoustic and linguistic variations challenge, and data scarcity problem, thereby improves performance of a children speech ASR system for Punjabi language. The proposed work makes use of formant modification as a technique for spectral warping in order to reduce the amount of acoustic and linguistic variation that exists between children's speech and adult speech. This is done in order to address the first problem. Further, to address the second issue of data scarcity, this paper discusses training of ASR models on augmented children speech data. Also, the work proposes an MFCC-FDLP hybrid approach for front end feature extraction by combining the Mel-Frequency Cepstral Coefficients (MFCC) features extraction technique with the Frequency Domain Linear Prediction (FDLP) technique. Both of these techniques are well established. In order to carry out the implementation of the data augmentation, the end-to-end Text to Speech (TTS) generative model known as Tacotron 2 was utilised. In order to implement continuous Punjabi language ASR systems, the work that is being

proposed uses MFCC, FDLP, and a hybrid of MFCC and FDLP techniques for front end feature extraction. Additionally, it uses Time Delay Neural Network (TDNN) for backend acoustic modelling and a trigram language model. To increase robustness of the proposed ASR system, we have included a batch of lexically diverse words in our pronunciation model which achieved a relative improvement of 29.44%.

According to the findings of a study done on the role of computational intelligence in the processing of speech acoustics[125]. In this paper, the author examined some of the most significant difficulties associated with speech recognition for various languages. According to an analysis of the relevant published material, the inaccessibility of standard databases containing information in minority languages is a significant barrier to research recognition all over the world. The research on speech recognition of Indian languages (with the exception of Hindi), in comparison to the research on speech recognition of non-Indian languages, has not yet achieved the expected milestone. When it comes to developing ASR for minority languages, the system that is most commonly used is a combination of MFCC and DNN–HMM classifier. On the other hand, when it comes to developing ASR for majority languages, researchers are using much more advanced algorithms of DNN. It has also been observed that there is not a lot of research done in this area, and there is a need for even more research to be carried out, particularly in the case of minority languages. This is something that needs to be done.

In the development of a reliable speech-to-text algorithm for Nigerian speakers of English[115]. This article presents a STT algorithm that is able to withstand the heavy accent used by speakers of English in Nigeria. During the stage of the project known as "data acquisition," approximately 27,000 isolated speech samples were gathered from five different ethnic groups. Thirty percent of the samples came from Yoruba

speakers, twenty-nine percent came from Hausa speakers, twenty percent came from Igbo speakers, and the remaining twenty-one percent came from Fulani and Ijaw speakers. After the data had been preprocessed, the features were extracted with the help of the Mel-Frequency Cepstral Coefficients (MFCCs), which had 13 coefficients. The Hidden Markov Model (HMM) with a variable number of states was selected as the method of recognition that was utilised. The findings of this research indicate that an average accuracy of 86% was accomplished for the 10-word vocabulary isolated speech that was taken into consideration with the number of states equal to 5. In addition, an average accuracy of 86% was accomplished by using a vocabulary of 10 words and a number of HMM states equal to 7. An accuracy rate of 90% was achieved on average for the 10-word vocabulary with the number of states set to 9.

In the course of an in-depth analysis of the published research on Hausa Natural Language Processing[116]. In this research paper, using a keyword index and article title search, we present a systematic analysis of the current literature that is applicable to HNLP that can be found in the Google Scholar database from the year 2015 to the year June 2020. Only a few research papers on HNLP research have been published recently. These papers focus on areas such as part-of-speech tagging (POS), name entity recognition (NER), words embedding, speech recognition, and machine translation. This is because Natural Language Processing (NLP) relies on a substantial quantity of data that has been annotated by humans in order to train intelligent models. As a result of the substantial amount of research that has been conducted on NLP in English and other languages, researchers are now becoming interested in HNLP. The primary goals of this paper are to encourage research, to identify possible subject areas for additional research in the HNLP, and to offer assistance to researchers in the

process of developing additional examinations for studies that are pertinent to the field.

In a related study on the identification of accents among Nigerians of varying ethnic backgrounds[117]. This study provides support for accent-dependent automatic speech recognition by applying a supervised learning algorithm to the task of recognising three Nigerian ethnic groups (Yoruba, Igbo, and Hausa) and distinguishing between them based on their accents. This is accomplished by constructing sequential Mel-Frequency Cepstral Coefficients (MFCC) features from the frames of the audio sample. This research was carried out in Nigeria. According to the findings of our research, an effective method for recognising and categorising accents is to concatenate the MFCC features in a sequential fashion and then apply a supervised learning strategy. This approach achieves both high efficiency and high accuracy in its results.

Assessing Hausa large vocabulary continuous speech recognition[118]. The authors of this study investigated and developed a Large Vocabulary Continuous Speech Recognition (LVCSR) system that is compatible with the Hausa language as part of this research project. In this article, the authors provide an overview of the Hausa language and speech database that was recently compiled as a part of our GlobalPhone corpus. They were able to make significant advancements by automatically replacing pronunciation dictionary entries that were inconsistent or flawed, including information on tone and vowel length, utilising state-of-the-art techniques for acoustic modelling, and crawling large quantities of text material from the internet for language modelling. When applied to read newspaper speech, a system that combines the best grapheme- and phoneme-based 2-pass systems achieves a word error rate of 13.16% on the development set and 16.26% on the test set.

Within the scope of this study, the effect of pitch enhancement in Punjabi children's speech recognition system was investigated under a variety of different acoustic conditions[119]. A Punjabi children's speech recognition system is developed in this work using a variety of acoustic matching and mismatching conditions. One major problem in children's speech recognition is the differences in the acoustic attributes of the children and adult speech signals, which leads to the poor recognition rate for the children's speech. This paper demonstrates how pitch enhanced features extracted from the front-end feature extraction process play an important role under mismatched acoustic conditions. The recognition rate of the children's speech recognition system using different age group datasets increases after enhancing the pitch with the Cepstral analysis in the feature extraction process. This is in comparison to the normal acoustics features extracted using the Mel Frequency Cepstral Coefficient (MFCC) feature extraction process. Kaldi toolkit is used for building the children's speech recognition models at different phoneme levels. Results show the improvement of 0.03% to 16.47% WER under different acoustic conditions using pitch enhanced features.

In a work on automatic speaker verification systems and spoof detection techniques: review and analysis[120]. Reviewing and evaluating the significant advances that have been suggested by a variety of researchers and scientists is the purpose of this particular piece of writing. The frontend of these systems is designed using a variety of feature extraction strategies, including some that are more traditional, such as autoregressive and cepstral, as well as some that are more modern and based on deep learning. The backend of an ASV system is where the extracted features are learned and classified. This can be done using traditional machine learning models or deep learning models, both of which are also the primary focus of the review that is being

presented. Since the advent of practical work in this field, experimental studies have adopted the practice of continuously modifying datasets and evaluation measures in order to develop reliable systems. The majority of the contributing spoofed speech datasets and evaluation protocols are dissected and analysed here. Spoofing attacks on ASV systems can come in a variety of forms, including speech synthesis (SS), voice conversion (VC), replay, mimicry, and twins. This work enables the defence mechanism of ASV by providing knowledge of techniques for the generation of attacks similar to those being targeted. The results of this survey herald the beginning of a new era in the development of ASV systems and highlight the beginning of a new generation (G4) in the methods of SS attack development. The paper makes its best efforts to give newcomers to this area the complete idea of ASV and also sheds some light on some of the spoofing attacks that can be targeted during implementation of the future ASV systems. With the increase in advancement of deep learning techniques, the paper also makes its best efforts to give newcomers to this area the complete idea of ASV.

An empirical comparison of deep learning and traditional classifiers for speaker verification in emotional talking environments is presented[121]. In this paper, an empirical comparison of the performances of traditional classifiers, such as the Gaussian Mixture Model (GMM), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Artificial neural networks (ANN), and deep learning classifiers, such as Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU), in addition to the ivector approach, is conducted for a text-independent speaker The deep models undergo hyperparameter tuning using the Grid Search optimization algorithm. The models are trained and validated using a private Arabic Emirati Speech Database, a public Crowd-Sourced

Emotional Multimodal Actors (CREMA) database, and a Ryerson Audio–Visual Database of Emotional Speech and Song dataset (RAVDESS) database. The results of experiments show that deep learning architectures do not automatically perform better than traditional classifiers. In point of fact, evaluation was carried out utilising Area Under the Curve (AUC) scores in addition to Equal Error Rate (EER) scores. According to the findings, the GMM model produces the lowest EER values and the best AUC scores across all datasets among the classical classifiers. This is true regardless of the dataset used. In addition, the ivector model surpasses all the fine-tuned deep models (CNN, LSTM, and GRU) based on both evaluation metrics in the neutral, as well as the emotional speech. In addition, the GMM outperforms the ivector using the Emirati and RAVDESS databases.

In a further study on the detection of multitasking synthetic speech on Indian Languages study's objective is to evaluate the efficacy of various methods for detecting synthetic spoofing using a multilingual, loosely constrained Indian language set as its test subject[112]. This paper aims to achieve a multitasking spoofing detection by identifying real/spoof utterance identification as well as the regional language spoofing attack vector. Specifically, this will be done by using a combination of real and spoof utterances. In order to accomplish this, the features and the classifiers that are best candidates for the detection of synthetic spoofing and the identification of language are selected in the appropriate manner. Our methodology compares the performances of three main different classifiers GMM, SVM, DNN on the vector formulated from the accumulation of MFCC features. Hindi, Malayalam, Tamil, Telugu are the four languages which are taken into account for the comparison. It has been discovered that, out of all of these classifiers, SVM and DNN produce the best results, with EER rates of 1.98% and 1.19% respectively.

In a study of the emotional information acoustic characteristics of synthetic speech phoneme/eiIn[123]. This paper, the author compared the differences of acoustic characteristics between synthetic speech and emotional speech under the same text from the perspective of the lack of emotional expression that is present in synthetic speech by using Praat software for a single phoneme /ei/. When the results were analysed, it was determined that the differences in the emotional information were primarily in the small dispersion of synthetic speech fundamental frequency, the dispersion of synthetic speech intensity was much smaller than that of real speech with large emotional fluctuations, the harmonic waves in narrowband spectrograms were nearly straight without bending and jittering, and the formant centre frequencies interlacing degree is small. These were the conclusions that were reached as a result of the analysis. The absence of harmonic waves at frequencies above 3000 Hz and the obvious difference in the direction of the tail end of the second formant are two examples of the common differences between synthetic speech and neutral and emotional speech. The absence of harmonic waves at these frequencies is one of the hallmarks of synthetic speech.

According to the findings of a study on spectral warping and data augmentation for low resource language ASR systems operating under mismatched conditions[124]. The proposed work in this paper tries to address the both challenges i.e., acoustic and linguistic variations challenge, and data scarcity problem, thereby improves performance of a children speech ASR system for Punjabi language. The proposed work makes use of formant modification as a technique for spectral warping in order to reduce the amount of acoustic and linguistic variation that exists between children's speech and adult speech. This is done in order to address the first problem. Further, to address the second issue of data scarcity, this paper discusses training of ASR models

on augmented children speech data. Also, the work proposes an MFCC-FDLP hybrid approach for front end feature extraction by combining the Mel-Frequency Cepstral Coefficients (MFCC) features extraction technique with the Frequency Domain Linear Prediction (FDLP) technique. Both of these techniques are well established. In order to carry out the implementation of the data augmentation, the end-to-end Text to Speech (TTS) generative model known as Tacotron 2 was utilised. In order to implement continuous Punjabi language ASR systems, the work that is being proposed uses MFCC, FDLP, and a hybrid of MFCC and FDLP techniques for front end feature extraction. Additionally, it uses Time Delay Neural Network (TDNN) for backend acoustic modelling and a trigram language model. To increase robustness of the proposed ASR system, we have included a batch of lexically diverse words in our pronunciation model which achieved a relative improvement of 29.44%.

In a study done on the role of computational intelligence in the processing of speech acoustics[125]. In the paper, we examined some of the most significant difficulties associated with speech recognition for various languages. According to an analysis of the relevant published material, the inaccessibility of standard databases containing information in minority languages is a significant barrier to research recognition all over the world. The research on speech recognition of Indian languages (with the exception of Hindi), in comparison to the research on speech recognition of non-Indian languages, has not yet achieved the expected milestone. When it comes to developing ASR for minority languages, the system that is most commonly used is a combination of MFCC and DNN–HMM classifier. On the other hand, when it comes to developing ASR for majority languages, researchers are using much more advanced algorithms of DNN. It has also been observed that there is not a lot of

research done in this area, and there is a need for even more research to be carried out, particularly in the case of minority languages. This is something that needs to be done.

2.4 Summary of Gaps in Literature

This section presented a review or literature related and relevant to the research topic. The section started with the conceptual review where relevant concepts (speech, Speech Recognition, speech recognition models, Nigerian major languages) were defined and explained. This was followed by the theoretical framework which analyzes the components of a conventional ASR system (Feature Extraction, Acoustic Modeling, Language Modeling, and Lexical Modeling). Empirical findings on speech recognition using machine learning and other algorithms, accent identification, language identification, speech algorithms in some dialects of the country and other languages and dialects outside Nigeria were discussed, revealed and presented.

It is evident from the preceding sections that the accent and speech recognition developed have a limited level of intelligence. Also, the works are limited to single Nigerian language (Only either Yoruba, Hausa or Igbo). Several academics concur on the necessity to enhance the artificial intelligence skills and data sets of accent recognition of the major Nigerian language accent[95,97,100]. In light of this highlighted gap, the primary topic of this study will be: what are the current methods to improve accurately the speech recognition algorithm of major Nigerian Languages (Yoruba, Hausa, Ibo)? This question, which has not previously been examined in speech

recognition, will serve as the basis for further study. According to the literature analysis, the following languages were anlyzed using speech recognition. They include Hausa, Yoruba and Igbo (individually), Punji, Mongolian, Hindi, tonal, Bribri, fongbe amongst other. This paper proposes improving the speech recognition algorithm of major Nigerian Languages (Yoruba, Hausa, Ibo).

Endnotes

1. S.S.S BHABAD. *Speech Recognition & Rectification For Articulatory Handicapped People* (Doctoral Dissertation, Savitribai Phule Pune University). 2019

2. C Zhang, Y Lu. *Study on artificial intelligence: The state of the art and future prospects*. **Journal of Industrial Information Integration**. 2021 Sep 1;23:100224.

3. AP Stampfl, Z Liu , J Hu, K Sawada, H Takano, Y Kohmura, T Ishikawa, JH Lim, JH Je, CM Low, A Teo. SYNAPSE: *An international roadmap to large brain imaging*. Physics Reports. 2023 Feb 9;999:1-60.

4. L Bloomfield, E Lane, M Mangalam, DG Kelty-Stephen. *Perceiving and remembering speech depend on multifractal nonlinearity in movements producing and exploring speech*. **Journal of the Royal Society Interface**. 2021 Aug 4;18(181):20210272.

5. TJ Hixon, G Weismer, JD Hoit. *Preclinical speech science: Anatomy, physiology, acoustics, and perception*. Plural Publishing; 2018 Aug 31.

6. YÜ Sönmez, A Varol. *In-Depth analysis of speech production, auditory system, emotion theories and emotion recognition.* In 2020 8th International Symposium on Digital Forensics and Security (ISDFS) 2020 Jun 1 (pp. 1-8). IEEE.

7. SC Hedger, IS Johnsrude. *Speech perception under adverse listening conditions*. In Speech Perception 2022 (pp. 141-171). Springer, Cham.

8. JG Bernstein, OA Stakhovskaya, KK Jensen, MJ Goupell. *Acoustic hearing can interfere with single-sided deafness cochlear-implant speech perception*. Ear and hearing. 2020 Jul;41(4):747.

9. O Räsänen, S Seshadri, M Lavechin, A Cristia, M Casillas. ALICE: *An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings.* Behavior Research Methods. 2021 Apr;53(2):818-35.

10. S Saxena. *Speech And Pattern Recognition For Emotion Classification Using Machine Learning (Doctoral dissertation).* 2019

11. MZ Anwar, Z Kaleem, A Jamalipour. *Machine learning inspired sound-based amateur drone detection for public safety applications.* IEEE Transactions on Vehicular Technology. 2019 Jan 17;68(3):2526-34.

12. J Araújo Alves, F Neto Paiva, L Torres Silva, P Remoaldo. *Low-frequency noise and its main effects on human health—A review of the literature between 2016 and 2019.* Applied Sciences. 2020 Jul 28;10(15):5205.

13. B Zonooz, E Arani, KP Körding, PA Aalbers, T Celikel, AJ Van Opstal. *Spectral weighting underlies perceived sound elevation.* Scientific reports. 2019 Feb 7;9(1):1-2.

14. N Prasangini, H Nagahamulla. *Sinhala Speech to Sinhala Unicode Text Conversion for Disaster Relief Facilitation in Sri Lanka.* In 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS) 2018 Dec 21 (pp. 1-6). IEEE.

15. R Haeb-Umbach, S Watanabe, T Nakatani, M Bacchiani, B Hoffmeister, ML Seltzer, H Zen, M Souden. *Speech processing for digital home assistants: Combining signal processing with deep-learning techniques.* IEEE Signal processing magazine. 2019 Oct 30;36(6):111-24.

16. SM Abdou, AM Moussa. *Arabic speech recognition: Challenges and state of the art. Computational linguistics, speech and image processing for arabic language.* 2019:1-27.

17. EE Adam. *Deep learning based NLP techniques in text to speech synthesis for communication recognition.* **Journal of Soft Computing Paradigm (JSCP).** 2020 Dec 18;2(04):209-15.

18. P Kapil, A Ekbal. *A deep neural network based multi-task learning approach to hate speech detection.* Knowledge-Based Systems. 2020 Dec 27;210:106458.

19. H Pawa, N Gaikwad and S Kulkarni. 2020. *A Study of Techniques and Processes Involved in Speech Recognition System.*

20. SA IBRAHIM. *Speech Recognition Based On Convolutional Neural Networks (Doctoral dissertation, University of Gezira).*

21. M Stenman, 2015. *Automatic speech recognition An evaluation of Google Speech.*

22. N Das, S Chakraborty, J Chaki, N Padhy, N Dey. *Fundamentals, present and future perspectives of speech enhancement.* **International Journal of Speech Technology.** 2021 Dec;24(4):883-901.

23. SJ Priscilla, M Vanithalakshmi. *Aggression Monitoring In Speech Using Semantics and Pitch.* **Global Journal of Pure and Applied Mathematics.** 2017;13(9):5437-45.

24. M Stenman, 2015. *Automatic speech recognition an evaluation of Google* Speech.

25. V Bhardwaj, V Kukreja. *Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions.* Applied Acoustics. 2021 Jun 1;177:107918.

26. T Kumar, M Mahrishi, G Meena. *A comprehensive review of recent automatic speech summarization and keyword identification techniques*. Artificial Intelligence in Industrial Applications. 2022:111-26.

27. H Wang, D Wang. *Towards robust speech super-resolution.* IEEE/ACM transactions on audio, speech, and language processing. 2021 Jan 25;29:2058-66.

28. YA Ibrahim, SA Faki, TI Abidemi. *Automatic Speech Recognition Using Mfcc In Feature Extraction Based Hmm For Human Computer Interaction In Hausa.* Annals. Computer Science Series. 2019 Dec 1;17(2).

29. S Debnath, P Roy. *Automatic speech recognition based on clustering technique.* In Emerging Technology in Modelling and Graphics 2020 (pp. 679-688). Springer, Singapore.

30. D Deshwal, P Sangwan, D Kumar. *Feature extraction methods in language identification: a survey.* Wireless Personal Communications. 2019 Aug;107(4):2071-103.

31. M Stenman. *Automatic speech recognition An evaluation of Google Speech.*2015

32. BF Zaidi, SA Selouani, M Boudraa, M Sidi Yakoub. *Deep neural network architectures for dysarthric speech analysis and recognition.* Neural Computing and Applications. 2021 Aug;33(15):9089-108.

33. V Rajadnya & K Joshi. (2021, December). *Raga Classification Based on MFCC and Variants.* In 2021 IEEE 2nd International Conference on Technology, Engineering, Management for Societal impact using Marketing, Entrepreneurship and Talent (TEMSMET) (pp. 1-6). IEEE.

34. D Wang, X Wang, S Lv. *An overview of end-to-end automatic speech recognition.* Symmetry. 2019 Aug 7;11(8):1018..

35. F Zhang, S Han, H Gao, T Wang. *A Gaussian mixture based hidden Markov model for motion recognition with 3D vision device*. Computers & Electrical Engineering. 2020 May 1;83:106603.

36. V Bhardwaj, MT Ben Othman, V Kukreja, Y Belkhier, M Bajaj, BS Goud, AU Rehman, M Shafiq, H Hamam. *Automatic Speech Recognition (ASR) Systems for Children: A Systematic Literature Review.* Applied Sciences. 2022 Apr 27;12(9):4419.

37. Y Liu, Y Qian, N Chen, T Fu, Y Zhang, K Yu. *Deep feature for text-dependent speaker verification.* Speech Communication. 2015 Oct 1;73:1-3.

38. A Krizhevsky, I Sutskever, GE Hinton. *Image net classification with deep convolutional neural networks.* Communications of the ACM. 2017 May 24;60(6):84-90.

39. S Ambrogio, P Narayanan, H Tsai, RM Shelby, I Boybat, C Di Nolfo, S Sidler, M Giordano, M Bodini, NC Farinha, B Killeen. *Equivalent-accuracy accelerated neural-network training using analogue memory.* Nature. 2018 Jun;558(7708):60-7.

40. C Bensch. (2021). *Continuous Learning In Automatic Speech Recognition (Doctoral dissertation, Maastricht University).*

41. SL Aouragh, A Yousfi, S Laaroussi, H Gueddah, M Nejja. *A new estimate of the n-gram language model.* Procedia Computer Science. 2021 Jan 1;189:211-5.

42. AT Liu, SW Li, HY Lee. Tera: *Self-supervised learning of transformer encoder representation for speech. IEEE/ACM Transactions on Audio, Speech, and Language Processing.* 2021 Jul 8;29:2351-66.

43. H Babii, A Janes, R Robbes. *Modeling vocabulary for big code machine learning.* arXiv preprint arXiv:1904.01873. 2019 Apr 3.

44. Z Alyafeai, MS Al-shaibani, M Ghaleb, I Ahmad. *Evaluating various tokenizers for Arabic text classification.* Neural Processing Letters. 2022 Aug 18:1-23.

45. A Martinez, K Sudoh, Y Matsumoto. *Sub-Subword N-Gram Features for Subword-Level Neural Machine Translation.* **Journal of Natural Language Processing.** 2021;28(1):82-103.

46. K Bostrom, G Durrett. *Byte pair encoding is suboptimal for language model pretraining.* arXiv preprint arXiv:2004.03720. 2020 Apr 7.

47. YA Ibrahim, SA Faki, TI Abidemi. *Automatic Speech Recognition Using Mfcc In Feature Extraction Based Hmm For Human Computer Interaction In Hausa. Annals.* Computer Science Series. 2019 Dec 1;17(2).

48. OA Adetunmbi, OO Obe, JN Iyanda. *Development of Standard Yorùbá speech-to-text system using HTK*. **International Journal of Speech Technology**. 2016 Dec;19(4):929-44.

49. A Atanda, S Yusof, M Hariharan. *Yorùbá automatic speech recognition: A review*. In Rural ICT Development (RICTD) International Conference 2013 (Vol. 1, No. 1, pp. 116-121).

50. AE Akinwonmi. *Development of a Prosodic Read Speech Syllabic Corpus of the Yoruba Language.* Development. 2021 Jun;7(36).

51. IE Onyenwe. *Developing methods and resources for automated processing of the african language igbo* (Doctoral dissertation, University of Sheffield).2017

52. Kruthika Prasanna Simha. *Improving Automatic Speech Recognition on Endanger omatic Speech Recognition on Endangered Languages.* Thesis. Rochester Institute of Technology, 2019.

53. S Bhatt, A Jain, A Dev. *Feature extraction techniques with analysis of confusing words for speech recognition in the Hindi language.* Wireless Personal Communications. 2021 Jun;118(4):3303-33.

54. R Yang, G Cheng, H Miao, T Li, P Zhang, Y Yan. *Keyword Search Using Attention-Based End-to-End ASR and Frame-Synchronous Phoneme Alignments.* IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021 Oct 15;29:3202-15.

55. KP Simha. *Improving Automatic Speech Recognition on Endangered Languages*. Rochester Institute of Technology; 2019.

56. G Chrupała. *Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques.* **Journal of Artificial Intelligence Research.** 2022 Feb 18;73:673-707.

57. KP Simha. *Improving Automatic Speech Recognition on Endangered Languages*. Rochester Institute of Technology; 2019.

58. FS Al-Anzi, D AbuZeina. *The capacity of Mel Frequency Cepstral Coefficients for speech recognition.* **International Journal of Computer and Information Engineering.** 2017 Sep 1;11(10):1149-53.

59. ZW Ding, XF Li, X Huang, MB Wang, QB Tang, JD Jia. *Feature extraction, recognition, and classification of acoustic emission waveform signal of coal rock sample under uniaxial compression.* **International Journal of Rock Mechanics and Mining Sciences**. 2022 Dec 1;160:105262.

60. AA Abdulsatar, VV Davydov, VV Yushkova, AP Glinushkin, VY Rud. *Age and gender recognition from speech signals*. **In Journal of Physics: Conference Series 2019**
Dec 1 (Vol. 1410, No. 1, p. 012073). IOP Publishing.

61. KR Borisagar, RM Thanki, BS Sedani. *Generation of Speech Signal and Its Characteristics.* In Speech Enhancement Techniques for Digital Hearing Aids 2019 (pp. 13-27). Springer, Cham.

62. E Bezzam, S Kashani, P Hurley, M Simeoni. *pyFFS: A Python Library for Fast Fourier Series Computation.* arXiv preprint arXiv:2110.00262. 2021 Oct 1.

63. G Wang, X Wang, C Zhao. *An Iterative Hybrid Harmonics Detection Method Based on Discrete Wavelet Transform and Bartlett–Hann Window.* Applied Sciences. 2020 Jan;10(11):3922.

64. P Jain, NR Kasture, T Kumar. *Comparative Study of Speaker Recognition Techniques in IoT Devices for Text Independent Negative Recognition.* Scalable Computing: Practice and Experience. 2020 Aug 1;21(3):359-68.

65. NP Narendra, B Schuller, P Alku. *The detection of Parkinson's disease from speech using voice source information.* IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021 May 7;29:1925-36.

66. K Jagadeeshwar, T Sreenivasarao, P Pulicherla, KN Satyanarayana, K Mohana Lakshmi, PM Kumar. *ASERNet: Automatic speech emotion recognition system using MFCC-based LPC approach with deep learning CNN.* **International Journal of Modeling, Simulation, and Scientific Computing.** 2022 Nov 30:2341029.

67. I Mikuöová. *Estimating Vocal Tract Resonances of Synthesized High-Pitched Vowels Using CNN* (Doctoral dissertation, Technische Universität Wien).

68. AO Salau, TD Olowoyo, SO Akinola. *Accent classification of the three major nigerian indigenous languages using 1d cnn lstm network model.* In Advances in Computational Intelligence Techniques 2020 (pp. 1-16). Springer, Singapore

69. A Purwar, H Sharma, Y Sharma, H Gupta, A Kaur. *Accent classification using Machine learning and Deep Learning Models.* In2022 1st International Conference on Informatics (ICI) 2022 Apr 14 (pp. 13-18). IEEE.

70. J. Wang., B. Li and J Zhang. *Research Article Use Brain-Like Audio Features to Improve Speech Recognition Performance.*2022

71. M Muttaqi, A Degirmenci, O Karal. *US Accent Recognition Using Machine Learning Methods.* In 2022 Innovations in Intelligent Systems and a Conference (ASYU) 2022 Sep 7 (pp. 1-6). IEEE.

72. K Elelu, T Le, C Le. *Collision Hazard Detection for Construction Worker Safety Using Audio Surveillance.* **Journal of Construction Engineering and Management.** 2023 Jan 1;149(1):04022159

73. S Darshana, H Theivaprakasham, GJ Lal, B Premjith, V Sowmya, K Soman. MARS: *A Hy.brid Deep CNN-based Multi-Accent Recognition System for English Language*. In 2022 First International Conference on Artificial Intelligence Trends and Pattern Recognition (ICAITPR) 2022 Mar 10 (pp. 1-6). IEEE.)

74. FO Oladipo, RA Habeeb, AE Musa, C Umezuruike, and OA Adeiza, 2021. *Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers.*

75. FO Oladipo, RA Habeeb, AE Musa. *Accent Identification of Ethnically Diverse Nigerian English Speakers*. Available at SSRN 3666815. 2020 Jul 24.

76. V Mikhailava, M Lesnichaia, N Bogach, I Lezhenin, J Blake, E Pyshkin. *Language Accent Detection with CNN Using Sparse Data from a Crowd-Sourced Speech Archive. Mathematics. 2022 Aug 13;10(16):2913.*

77. M Lesnichaia, V Mikhailava, N Bogach, I Lezhenin, J Blake, E Pyshkin. *Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms.* Proc. Interspeech 2022. 2022:3669-73.

78. FO Oladipo, RA Habeeb, AE Musa, C Umezuruike, OA Adeiza. *Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers*

79. AF Abdulwahab, SA Mohd Yusof, H Husni. *Acoustic Comparison of Malaysian and Nigerian English Accents.* **Journal of Telecommunication, Electronic and Computer Engineering (JTEC)** 9 (3-5), 141-146, 2017

80. UG Muhammad. *A Comparative Phonological Analysis of Varieties of English Spoken by Native Speakers of Nigerian Languages (Hausa, Igbo, Kanuri and Yoruba) for the Determination of Speakers' Origins (Doctoral dissertation, University of York).*

81. Z Song. *English speech recognition based on deep learning with multiple features.* Computing. 2020 Mar;102(3):663-82.

82. L Mohammadpour, TC Ling, CS Liew, A Aryanfar. *A Survey of CNN-Based Network Intrusion Detection.* Applied Sciences. 2022 Aug 15;12(16):8162.

83. K Choutri, M Lagha, S Meshoul, M Batouche, Y Kacel, N Mebarkia. *A Multi-Lingual Speech Recognition-Based Framework to Human-Drone Interaction*. Electronics. 2022 Jun 9;11(12):1829.

84. AS Haq, M Nasrun, C Setianingsih, MA Murti. *Speech recognition implementation using MFCC and DTW algorithm for home automation.* Proceeding of the Electrical Engineering Computer Science and Informatics. 2020 Oct;7(2):78-85.

85. R Mardhotillah, B Dirgantoro, C Setianingsih. *Speaker Recognition for Digital Forensic Audio Analysis using Support Vector Machine.* In 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) 2020 Dec 10 (pp. 514-519). IEEE.

86. M Nasrun, C Setianingsih. *Human Emotion Detection with Speech Recognition Using Mel-frequency Cepstral Coefficient and Support Vector Machine.* In 2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS) 2021 Apr 28 (pp. 1-6). IEEE.

87. M Zielonka, A Piastowski, A Czyżewski, P Nadachowski, M Operlejn, K Kaczor. *Recognition of Emotions in Speech Using Convolutional Neural Networks on Different Datasets.* Electronics. 2022 Jan;11(22):3831.

88. W Liu, Q Liao, F Qiao, W Xia, C Wang, F Lombardi. *Approximate designs for fast Fourier transform (FFT) with application to speech recognition.* IEEE Transactions on Circuits and Systems I: Regular Papers. 2019 Aug 23;66(12):4727-39.

89. S Lokesh, MR Devi. *Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method.* Cluster Computing. 2019 Sep;22(5):11669-79.).

90. Z Batzorig, O Bukhtsooj, AG Chensky, T Galbaatar. *Speech recognition in Mongolian language using a neural network with pre-processing technique*. In2020 International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE) 2020 Mar 12 (pp. 1-5). IEEE

91. MN ElBedwehy, GM Behery, R Elbarougy. *Emotional Speech Recognition Based on Weighted Distance Optimization System*. **International Journal of Pattern Recognition and Artificial Intelligence.** 2020 Oct 19;34(11):2050027.

92. U Avci. *A Pattern Mining Approach for Improving Speech Emotion Recognition.* **International Journal of Pattern Recognition and Artificial Intelligence.** 2022 Nov 24:2250045.

93. VG Mahesh, AN Raj, R Nersisson. *Implementation of Machine Learning-Aided Speech Analysis for Speaker Accent Identification Applied to Audio Forensics.* InAiding Forensic Investigation Through Deep Learning and Machine Learning Frameworks 2022 (pp. 174-194). IGI Global.

94. T Mourad. *Arabic Speech Recognition by Stationary Bionic Wavelet Transform and MFCC Using a Multi-layer Perceptron for Voice Control*. In The Stationary Bionic Wavelet Transform and its Applications for ECG and Speech Processing 2022 (pp. 69-81). Springer, Cham.)

95. U Sharma, H Om, AN Mishra. *HindiSpeech-Net: a deep learning based robust automatic speech recognition system for Hindi language.* Multimedia Tools and Applications. 2022 Oct 24:1-21.

96. OS Stefanenko, LV Lipinskiy, AS PolyakoJ, JA Khudonogova, ES Semenkin. *An intelligent voice recognition system based on fuzzy logic and the bag-of-words technique.* InIOP Conference Series: Materials Science and Engineering 2022 Mar 1 (Vol. 1230, No. 1, p. 012020). IOP Publishing.

97. W Liu, Q Liao, F Qiao, W Xia, C Wang, F Lombardi. *Approximate designs for fast Fourier transform (FFT) with application to speech recognition.* IEEE Transactions on Circuits and Systems I: Regular Papers. 2019 Aug 23;66(12):4727-39.

98. B Liu, X Ding, H Cai, W Zhu, Z Wang, W Liu, J Yang. *Precision adaptive MFCC based on R2SDF-FFT and approximate computing for low-power speech keywords recognition*. IEEE Circuits and Systems Magazine. 2021 Nov 15;21(4):24-39.

99. SA Yusof, AF Atanda, M Hariharan. *A review of Yorùbá Automatic Speech Recognition.* In2013 IEEE 3rd International Conference on System Engineering and Technology 2013 Aug 19 (pp. 242-247). IEEE.).

100. J Kaur, A Singh, V Kadyan. *Automatic speech recognition system for tonal languages: State-of-the-art survey.* Archives of Computational Methods in Engineering. 2021 May;28(3):1039-68.

101. J Kaur, A Singh, V Kadyan. *Automatic speech recognition system for tonal languages: State-of-the-art survey*. Archives of Computational Methods in Engineering. 2021 May;28(3):1039-68.

102. A Gutkin, I Demirsahin, O Kjartansson, CE Rivera, and K Túbòsún, 2020. *Developing an open-source corpus of yoruba speech*.

103. Fréjus AA LAleye, Laurent Besacier, Eugène C Ezin, Cina Motamed. *First automatic fongbe continuous speech recognition system: Development of acoustic models and language models.* Federated Conference on Computer Science and Information Systems (FedCSIS), 477-482, 2016

104. RC *Solano. Explicit Tone Transcription Improves ASR Performance in Extremely Low-Resource Languages: A Case Study in Bribri.* In Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas 2021 Jun (pp. 173-184

105. U Sharma, H Om, AN Mishra. *HindiSpeech-Net: a deep learning based robust automatic speech recognition system for Hindi language*. Multimedia Tools and Applications. 2022 Oct 24:1-21.

106. MC Lee, SC Yeh, JW Chang, ZY Chen. *Research on Chinese speech emotion recognition based on deep neural network and acoustic features. Sensors.* 2022 Jun 23;22(13):4744.

107. T Hasija, V Kadyan, K Guleria. *Recognition of Children Punjabi Speech using Tonal Non-Tonal Classifier.* In2021 International Conference on Emerging Smart Computing and Informatics (ESCI) 2021 Mar 5 (pp. 702-706). IEEE.

108. YA Wubet, KY Lian. *Voice Conversion Based Augmentation and a Hybrid CNN-LSTM Model for Improving Speaker-Independent Keyword Recognition on Limited Datasets.* IEEE Access. 2022 Aug 19;10:89170-80.

109. R Coto-Solano. *Computational sociophonetics using automatic speech recognition. Language and Linguistics Compass*. 2022 Sep;16(9):e12474.

110. R Coto-Solano, Nicholas, S Datta, V Quint, P Wills, EN Powell, L Koka‘ua, S Tanveer, I Feldman. *Development of automatic speech recognition for the documentation of Cook Islands Māori.*

111. FA LAleye, L Besacier, EC Ezin, C Motamed. *First automatic fongbe continuous speech recognition system: Development of acoustic models and language models.* In2016 Federated Conference on Computer Science and Information Systems (FedCSIS) 2016 Sep 11 (pp. 477-482). IEEE.

112. K Radha, M Bansal, SM Shabber. *Accent Classification of Native and Non-Native Children using Harmonic Pitch.* In2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP) 2022 Feb 12 (pp. 1-6). IEEE.

113. K Radha, M Bansal. *Audio Augmentation for Non-Native Children's Speech Recognition through Discriminative Learning*. Entropy. 2022 Oct 19;24(10):1490.

114. LK Ajayi, A Azeta, I Odun-Ayo, ET Aniemeka. *Acoustic Nudging-Based Model for Vocabulary Reformulation in Continuous Yorùbá Speech Recognition*. International Conference on Computational Science and Its Applications 2022 (pp. 494-508). Springer, Cham.

115. A Scholar, 2020. *Development of a Robust Speech-to-Text Algorithm for Nigerian English Speakers* 1Mohammed M. Sulaiman, 2Yahya S. Hadi, 1Mohammed Katun and 1Shehu Yakubu.

116. RY Zakari, ZK Lawal, I Abdulmumin. *A systematic literature review of Hausa Natural Language Processing*. **International Journal of Computer and Information Technology** (2279-0764). 2021 Jul 31;10(4).

117. FO Oladipo, RA Habeeb, AE Musa. *Accent Identification of Ethnically Diverse Nigerian English Speakers.* Available at SSRN 3666815. 2020 Jul 24.

118. Schultz, EG Djomgang, DI Schlippe, DI Vu. *Hausa Large Vocabulary Continuous Speech Recognition.*

119. V Bhardwaj, V Kukreja. *Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions*. Applied Acoustics. 2021 Jun 1;177:107918.

120. A Mittal, M Dua. *Automatic speaker verification systems and spoof detection techniques: review and analysis.* **International Journal of Speech Technology.** 2022 Mar;25(1):105-34.

121. AB Nassif, I Shahin, M Lataifeh, A Elnagar, N Nemmour. *Empirical Comparison between Deep and Classical Classifiers for Speaker Verification in Emotional Talking Environments.* Information. 2022 Sep 27;13(10):456.

122. AR Ambili, RC Roy. *Multi Tasking Synthetic Speech Detection on Indian Languages.* In 2022 International Conference on Innovative Trends in Information Technology (ICITIIT) 2022 Feb 12 (pp. 1-6). IEEE.

123. J Zhou, X Hu, Q Ma. *A study of the emotional information acoustic characteristics of synthetic speech phoneme/ei.* In International Conference on Electronic Information Engineering and Computer Communication (EIECC 2021) 2022 May 4 (Vol. 12172, pp. 170-178). SPIE.

124. M Dua, V Kadyan, N Banthia, Bansal A, Agarwal T. *Spectral warping and data augmentation for low resource language ASR system under mismatched conditions.* Applied Acoustics. 2022 Mar 15;190:108643.

125. A Singh, N Kaur, V Kukreja, Kadyan V, Kumar M. *Computational intelligence in processing of speech acoustics: a survey*. Complex & Intelligent Systems. 2022 Feb 17:1-39.

126. S Singh, R Wang, F Hou, Z Ma. *Enhancing End-to-End Automatic Speech Recognition for Low-Resource Punjabi Language Using Synthesized Datasets.* Available at SSRN 4181844.

127. Y Lai. *Application of the Artificial Intelligence Algorithm in the Automatic Segmentation of Mandarin Dialect Accent.* Mobile Information Systems. 2022 Feb 24;2022.

128. I Hwang, JH Chang. *End-to-End Speech Endpoint Detection Utilizing Acoustic and Language Modeling Knowledge for Online Low-Latency Speech Recognition.* IEEE Access. 2020 Aug 31;8:161109-23..

Chapter Three

Methodology

3.1 Research Approach

There are a number of different approaches to the implementation of a speech recognition system, but this work considered the four major processing steps, namely: data preparation; feature extraction, training and testing which will be detailed in this chapter. Furthermore, the minimum hardware requirements and software needs will be covered.

3.2 System Design

The speakers of Hausa, Yoruba, and Igbo will be categorised according to the speech data collected from the three major indigenous languages in Nigeria. Figure 3.1 and Figure 3.2 depicts the conceptual model and flowchart for the proposed method of accent classification. In this research, a general speech data will be used i.e a general dialect for Hausa, Yoruba, and Igbo. Other dialects in each language (for example, Egba dialect in Yoruba) would not be considered.

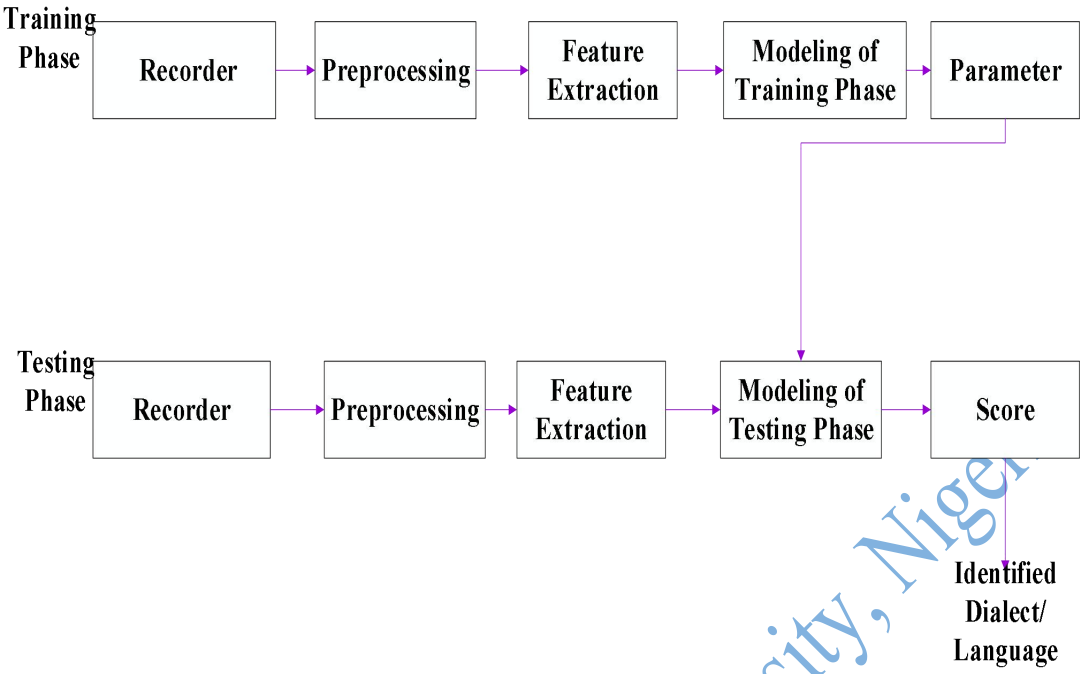


Fig. 3.1: Conceptual Model of the Proposed Design

3.3 Requirement Specification

Hardware Minimum Requirements: The minimum hardware requirements pertain to the physical features of the machine required to run the accent recognition program. The following are the features: at least 100-250 GB HDD, 2 GB RAM, and an Intel Pentium Dual-Core processor

Software Requirements: These are the computer programmes and procedures needed to put the system into action. The tools used include: Windows 10 operating system, MATLAB 2015, Python programming language and Audacity which will be also used for post-processing of all types of audio data

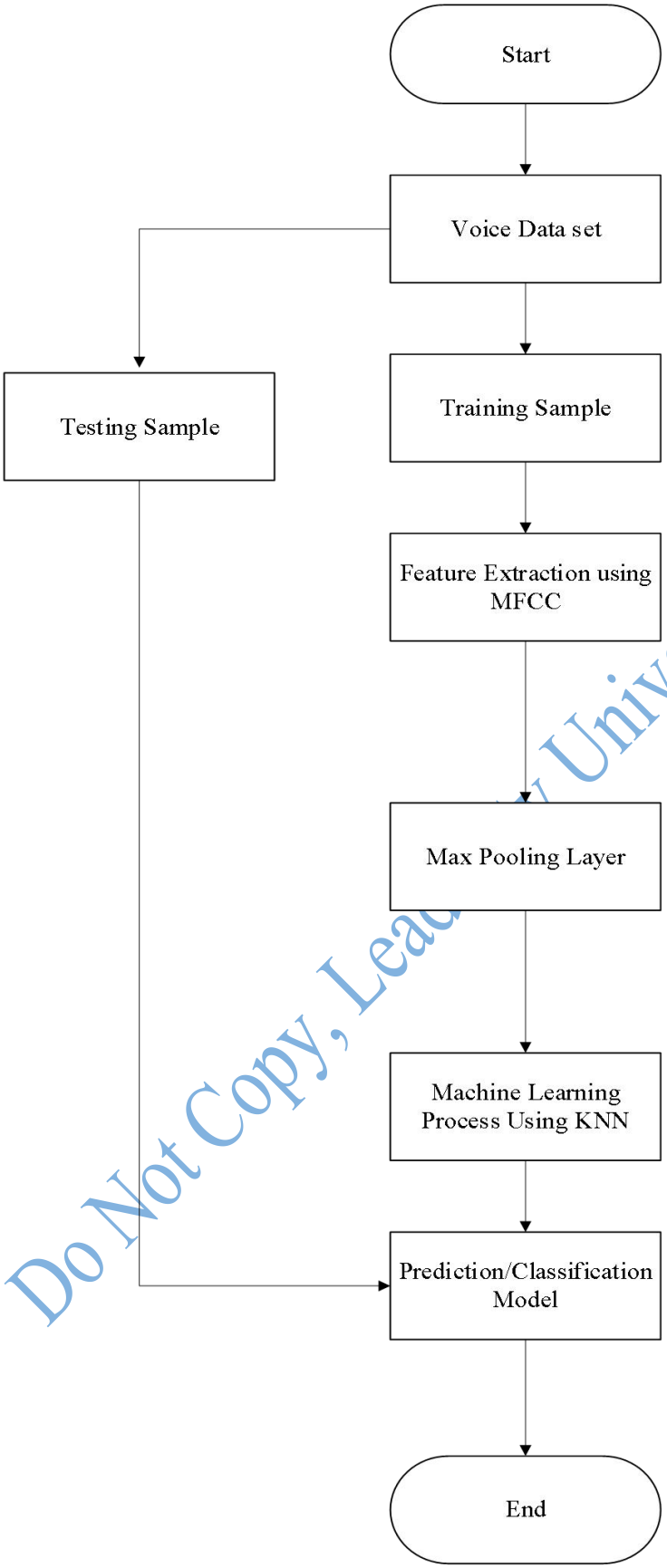


Fig. 3.2: Flowchart of the proposed Accent Classification Process

3.4 Research Method

3.4.1 Data Collection

Speech data is required not only for training but also for testing. The built system will collect audio data in the form of speech from 200 speakers from an open-source audio recording of each of the three major Nigerian indigenous languages, namely Hausa, Igbo, and Yoruba, recorded in an environment with no background noise. Also, voices were extracted from different samples from the news media (channels, NTA, government officials, radio presenters). which are passed through a low band filter to surpress the noice.

The ascent or speech was based on general tribal languages not various dialects in each language. For example, Egba dialect in Yoruba and other dialects apart from the general Ibo or Hausa will not be considered. Data on the speech was collected by using a mobile phone in the role of a recorder. The recordings was made in an atmosphere that is calm and has a low level of background noise, and the sampling rate will be 20,000 Hz. The acquired speech data was a continuous speech in English from the various languages stored in WAV format. As can be seen in Figure 3.1, the training dataset consisted of approximately 12000 sentences, and was recorded and labelled with the help of the audio editing software Audacity.

The recording device have the open-source Audacity software installed on it. This programme is a digital audio editor and recording application that is open source and compatible with Windows as well as other operating systems[1]. In addition to recording audio from a variety of sources, Audacity can also be used for the post-processing of any and all types of audio data that are brought into the programme. The noise reduction effect in Audacity can be utilised to eliminate distracting background noise, such as that caused by electric fans or hums[2]. Following the saving of all of the

acquired speech data in a folder, it was imported into the Audacity software in order to get rid of any background noise, and then the data was exported from the mp3 format into the.wav format. This format is an audio processing format that requires less compression than other formats.

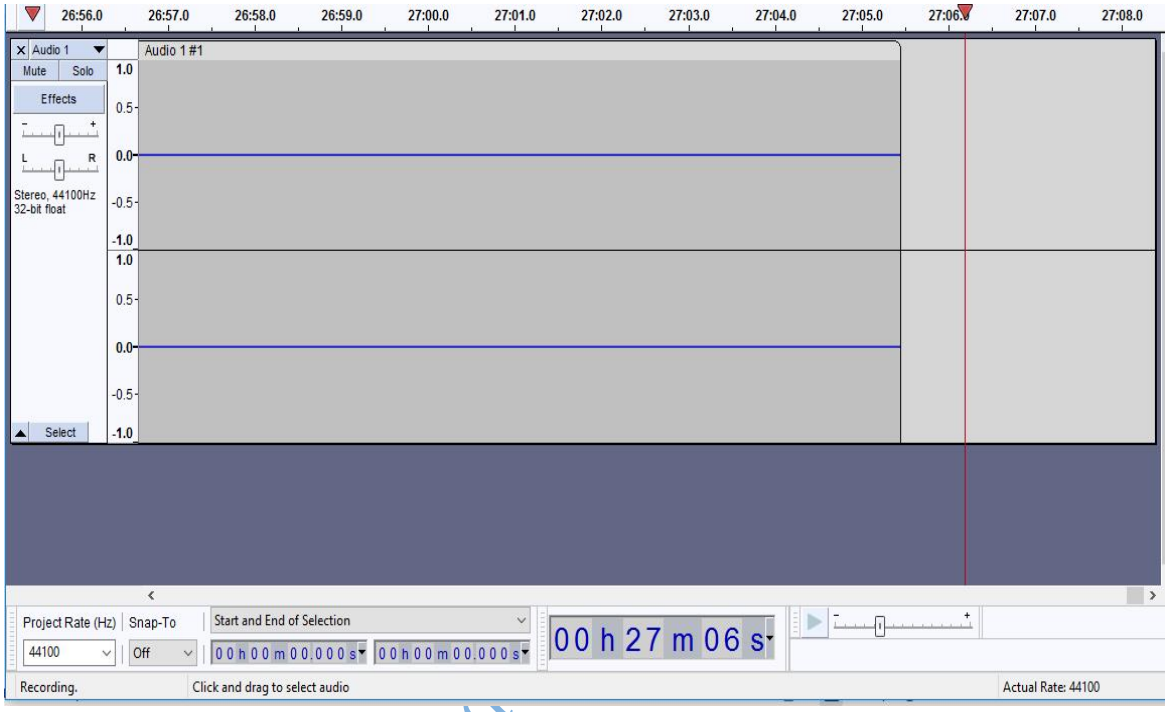


Fig. 3.3. Screenshot of the Speech Data Recording Process Using Audacity

Silence was removed from each sentence file using the Audacity audio editing software which enables the user to select the portion of speech file to save. All the speech data recorded were converted in order for it to be processed in MATLAB workspace. Each file was converted to 4800Hz 16bits wav file using the Audacity audio editing software.

3.4.2 Preprocessing Audio Data

Each audio (recorded and scrapped from news media source) was trimmed to the same length of sixty seconds in order to ensure uniformity in the inputs, which will be standardised. As a result, the dimensions of the audio data were shrunk into their mel-frequency cepstral coefficients (MFCCs), which is a representation of the short-term

spectrum of sounds. This occurred even after the trimming process has been completed. In order to accomplish this task, the audio clip was segmented into a few windows, and then the frequency information was extracted from each window.

3.4.3 Feature Extraction

This section provides a concise explanation of the procedures that were followed in order to extract features using MFCCs. The process of feature extraction will involve reading the audio speech into a numerical vector. This will assist in converting the speech data to the input required by the MFCC function, which will then be used to compute the MFCC vector.

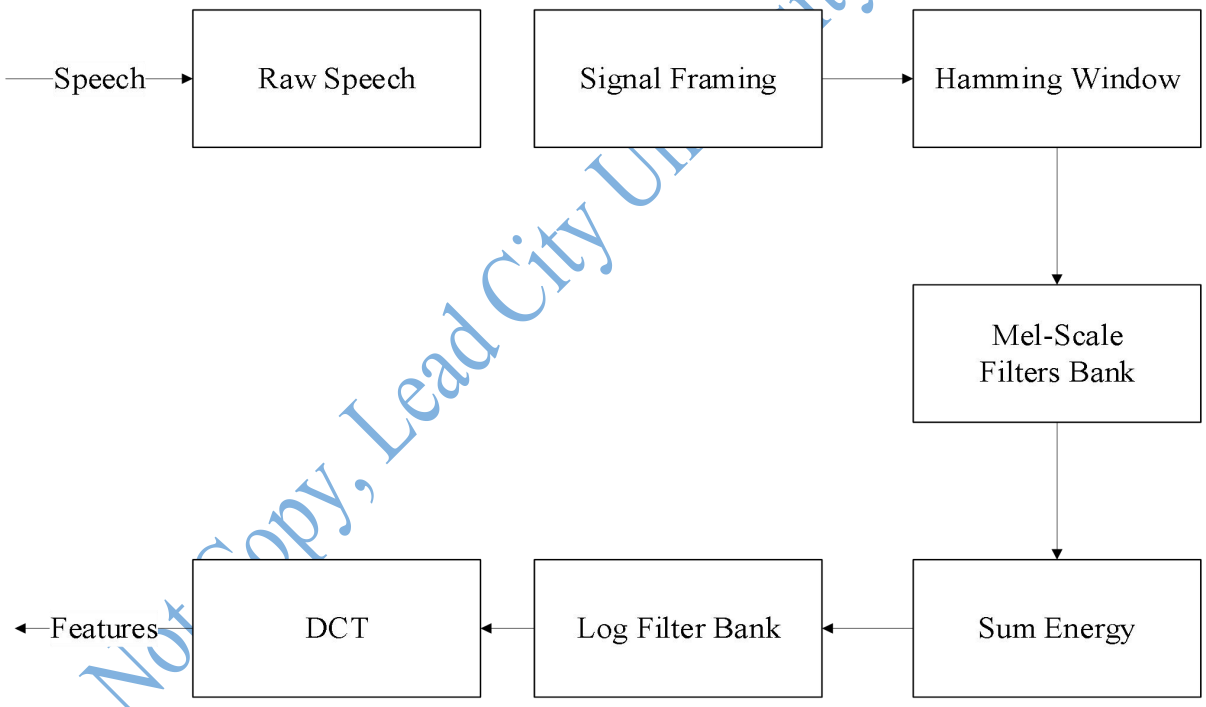


Fig. 3.4: MFCC Block Diagram[2].

$$f_{mel} = 2595 \log10(1 + f/700) \quad (3.1)$$

$$M^{-1}(m) = 700 (Exp(m/1125) - 1) \quad (3.2)$$

where f_{mel} and f are the mel and linear frequencies, respectively

The Mel scale is component of the MFCC, and it is used to establish a relationship between the perceived frequencies of pure tones and their actual measured

frequencies. The steps that are outlined below, which are depicted in Figure 3.2, will be used to accomplish this goal.

i. The signal data will be broken up into several smaller frames.
ii. An estimate of the power spectrum's periodogram will be computed for each of the frames individually.
iii. The mel filter bank will be utilised in order to process the power spectra, and the energy of each filter will be added.
iv. Determine the logarithm of all the energies in the filter bank.
v. Compute the discrete cosine transform of the logarithmic filter bank energies.

The conversions will be performed with Eqs. (3.1) and (3.2).

3.4.4 Training/Testing

The KNN algorithm was utilised both in the training phase and the testing phase. After the feature selection phase, features which represents which ascent the most and identify the percentage of closeness of each ascent to the sentences will be checked. The highest prediction percentage would be chosen as the preferred ascent. The K-nearest neighbours (KNN) prediction of an unknown pattern, also known as the query instance, is based on a very simple majority vote of the categories or classes of the nearest neighbours in the training space. This vote determines which neighbour is considered to be the most similar to the query instance. The fundamental concept is based on minimizing the distance between the unlabeled sample and the training samples in order to identify the samples that are the closest to being Kneighbors[3]. This calculation of the distance between one sample or pattern in the testing dataset, which contains the unknown patterns, and one sample in the training dataset, which contains samples with known class labels, is expressed as an equation and can be found below[3].

$$d_{ij}^2 = \sum_{m=1}^{M} \left[x_i\,(m) - x_j(m)\right]^2 \qquad (3.3)$$

where xi and xj are exemplars of the training and the testing datasets in the m[th] feature dimension, i.e., m=1, 2, …, M.

The next step will be to locate the class number to the unlabeled pattern based on the majority vote by simply summing up the class labels, assigned as c(xi)where xi is the class label of the selected NK (xj). The cardinality of NK (xj) is equal to K. Then, the subset of NN within the class set of l {1, 2,…L} is expressed mathematically as in Eq. (3.4)3.

$$N_k^l\,(x_j) = \{x_i \in NK\,(x_j) : c(x_i) = l\} \qquad (3.4)$$

Thus, the classification result l^* using majority vote is expressed mathematically as in Eq. (3.5).

$$l^* = \arg_l^{\max} \left| N_k^l(x_j) \right| \qquad (3.5)$$

The flowchart on how KNN works is shown in Figure 3.4. However, the K-parameter is determined by regression analysis.

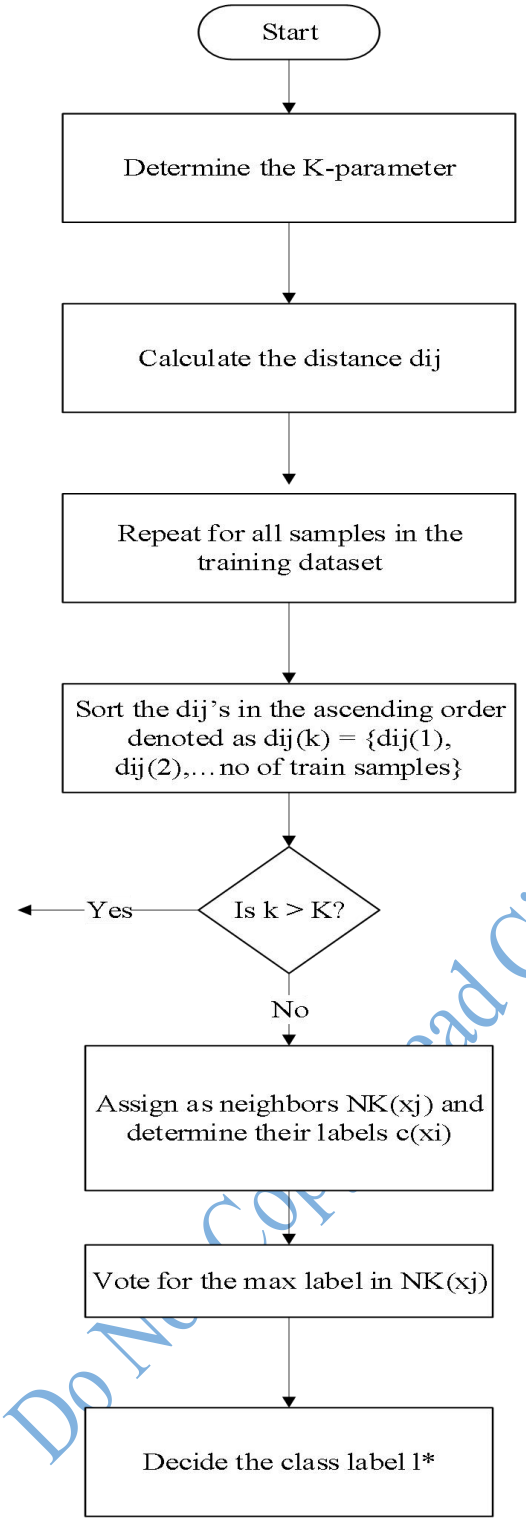


Fig. 3.5: Flowchart of K-Nearest Neighbors Algorithm[3].

Endnotes

1. Y.A Ibrahim , SA Faki, TI AbidemiI. *Automatic Speech Recognition Using Mfcc In Feature Extraction Based Hmm For Human Computer Interaction In Hausa*. Annals. Computer Science Series. 2019 Dec 1;17(2).

2. A.O Salau, T.D Olowoyo, S.O Akinola. *Accent classification of the three major nigerian indigenous languages using 1d cnn lstm network model*. InAdvances in Computational Intelligence Techniques 2020 (pp. 1-16). Springer, Singapore.

3. M.A Yusnita, M.P Paulraj, R.Y Sazali Yaacob, A.B Shahriman. A*nalysis of accent-sensitive words in multi-resolution mel-frequency cepstral coefficients for classification of accents in Malaysian English*. **International Journal of Automotive and Mechanical Engineering**. 2013;7:1053-73.

Chapter Four

Results and Discussion of Findings

4.1 Result on Acquiring Speech Data

Voiced signal for Nigerian from different tribes, mainly the 3 major tribes in Nigeria (Yoruba, Hausa and Igbo) speaking English was extracted (APIs (Application Programming Interfaces) and Web scrapping) from different platforms such as news media (Channels, NTA websites) where government officials, radio presenters from the 3 major tribes in Nigeria were on the news and other audio recordings. Once the inputs signal is known, next is to get the size. Also, the voice was in time domain, it was then converted to freq domain using fourier transform. This is done by Preprocessing (Optional) by applying filtering, windowing, or normalization techniques to ensure accurate frequency domain representation, Sampling the Signal at a regular interval to convert it into a discrete-time signal. The sampling rate is higher than twice the highest frequency present in the signal

FFT was applied calculating the Discrete Fourier Transform.

```
fft_result = np.fft.fft(signal)
```

The Magnitude Spectrum was calculated

```
magnitude_spectrum = np.abs(fft_result)
```

Finally, a corresponding frequency axis to represent the frequencies in the frequency domain was created. The frequency axis values are determined by the sampling rate and the length of the FFT.

```
# Length of the FFT result
fft_length = len(fft_result)
# Sampling rate of the signal
sampling_rate = ...
  # Frequency axis values
freq_axis = np.fft.fftfreq(fft_length, 1/sampling_rate)
```

4.2 Training the Dataset

The model was trained using Matlab R2015A. MATLAB R2015A is a popular programming language and interactive computing environment widely used for numerical computing, data analysis, visualization, and various other engineering and scientific applications.

Figure 4.1: Snapshot of the Training Model
Source: Research Design, 2023

4.2.1 Reading In the Voiced Input

The audio file was loaded into the model. Audio read function was used which allows reading of audio files in various formats and store the audio data in a variable for further processing. Below is the snippet code used, the full programming code is available at Appendix I

```
[St,Fs] = audioread(strcat('y','\',num2str(fn),'.aac'));
```

4.2.2 Define Window Size and Hop size

Window size refers to the length of the time window used to extract a segment of the signal for analysis, while The hop size (also known as the overlap size or step size) refers to the number of samples by which the window advances from one frame to the next. Window size used was 1024, while hopsize used was 512 (winSize = 1024;hopSize = 512)

4.2.3 Noise Reduction

Step 1: Parameters for the noise reduction filter was defined. Noise reduction filters are used to remove unwanted noise from audio or signals while preserving the desired underlying information. The snippet code used is given below. The full programming code is available at Appendix I

```
frame_len = round(fs*0.02); % frame length of 20ms
overlap_len = round(fs*0.01); % 50% overlap
freq_cutoff = 1000; % cutoff frequency for high-pass filter
noise_reduction = 10; % noise reduction level in dB
```

rame_len = round(fs*0.02);`**:
frame_len` is the length of each analysis frame or window in samples, fs` is the sampling frequency of the input signal in Hz. The value `0.02` represents the desired frame length in seconds (20 milliseconds in this case). The parameter `round` is used to round the result to the nearest integer, as the frame length needs to be an integer number of samples, overlap_len` is the length of the overlap between consecutive frames in samples, `fs` is the sampling frequency of the input signal in Hz. The value `0.01` represents the desired overlap length in seconds (10 milliseconds in this case). The parameter `round` is used to round the result to the nearest integer, as the overlap length needs to be an integer number of sample, freq_cutoff` is the cutoff frequency for a high-pass filter applied in the noise reduction algorithm.

The high-pass filter is used to attenuate low-frequency noise components, such as background noise or rumble, while preserving higher-frequency components associated with speech or desired signal information. The value `1000` represents the cutoff frequency in Hz. noise_reduction = 10;`noise_reduction` represents the desired noise reduction level in decibels (dB).

Step 2: A high-pass filter to remove low-frequency noise was created using the code below:

```
hp_filter = designfilt('highpassiir', 'FilterOrder', 8, 'PassbandFrequency', freq_cutoff, 'PassbandRipple', 0.2, 'SampleRate', fs);
```

Step 3: Apply the high-pass filter to the audio signal

```
x_filt = filtfilt(hp_filter, x);
```

Step 4: Perform noise reduction using spectral subtraction

```
[spec, f, t] = spectrogram(x_filt, hann(frame_len), overlap_len, [], fs);
[spec_noise, ~, ~] = spectrogram(x_filt, hann(frame_len), overlap_len, [], fs);
spec_noise = max(spec_noise, [], 2); % estimate the noise power spectrum
```

Step 5: Finally, the cleaned audio is written to a new file

```
audiowrite('audio_file_cleaned.wav', x_clean, fs);
```

4.2.4 Feature Extraction Using MFCC

Step 1: The clean audio file was read and pre-processed

Step 2: audio signal was divided into overlapping frames for further analysis.

```
% Frame the audio signals
frames_1 = buffer(audio_data_1, frame_length, frame_overlap, 'nodelay');
frames_2 = buffer(audio_data_2, frame_length, frame_overlap, 'nodelay');
% …
```

Step 3: Windowing function was applied to each frame to reduce spectral leakage.

```
% Compute MFCC coefficients for each frame
mfcc_coeffs_1 = computeMFCC(windowed_frames_1, sampling_rate,
num_mel_filters, lower_frequency, upper_frequency, num_mfcc_coeffs);
```

mfcc_coeffs_2 = computeMFCC(windowed_frames_2, sampling_rate,

Step 4: MFCC coefficients was combined from each frame to obtain a fixed-length feature representation for each dialect sample

```
% Feature aggregation (e.g., taking the mean of MFCC coefficients across frames)
mfcc_features_1 = mean(mfcc_coeffs_1, 2);
mfcc_features_2 = mean(mfcc_coeffs_2, 2);
% ...
```

4.3 Testing

Same procedure as above (training section) was used in the testing phase. Here, classification or prediction is done to ascertain the target class

4.3.1 Classification

The data are not in the same dimension. They have different length, speeches, timing etc. Hence the features were reshaped the into a row vector. The features were saved and labeled as 1, 2 or 3. Same thing is done for hausa and igbo. The features are now passed to a classifier based on different labels available and the features already generated. The Matlab interface thus displays training complete as shown in figure 4.2. This is when the training is complete.

Finally, the new input was passed in to the classifier to predict what class and after that, the prediction is done. If it is 1, it is Yoruba, if 2 Hausa and if 3, Igbo as shown in figure 4.3. Display of the classification result using the following snippet code

```
if class == 1
   disp('The new recording is in dialect A');
elseif class == 2
   disp('The new recording is in dialect B');
else
   disp('The new recording is in dialect C');
End
```

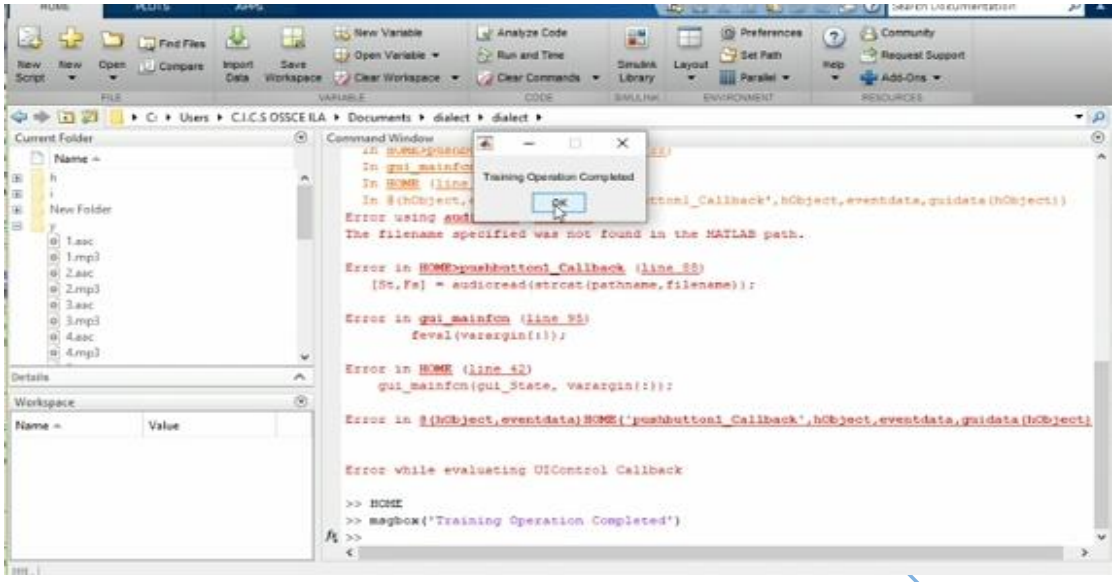


Figure 4.2: Snapshot of Matlab Interface Showing Training Completed

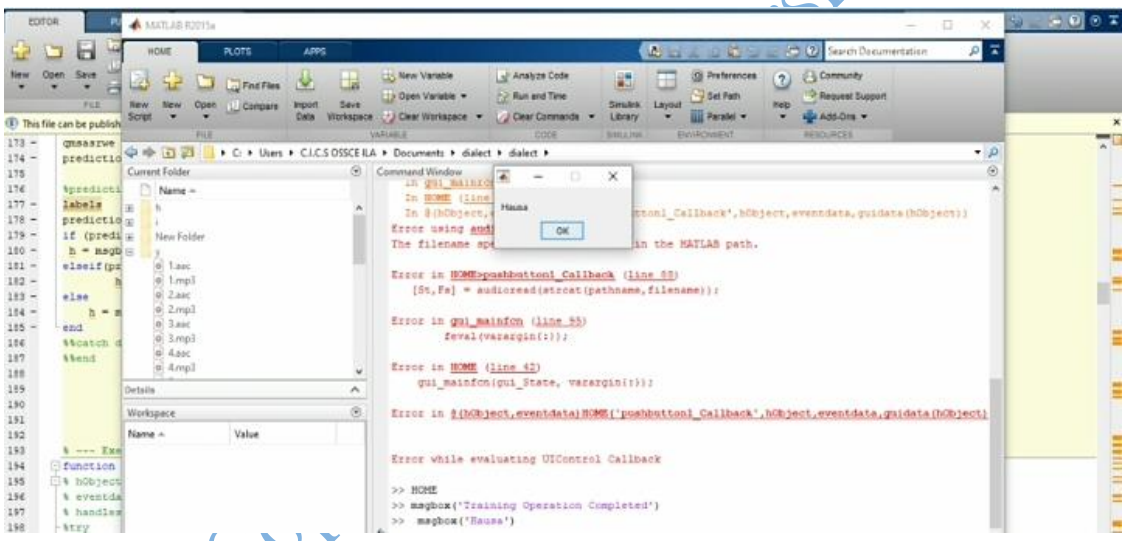


Figure 4.3: Snapshot of Matlab Interface Showing Hausa Dialect

4.4 Performance Evaluation

The performance of a model is evaluated using performance metrics like precision, recall, reject rate, and overall accuracy. The results presented apply to the data which is split into training (70% of the data) and testing (30% of the data).

The confusion matrix provides valuable insights into the performance of the dialect identification model. From the values in the matrix, various evaluation metrics such as accuracy, precision, recall (sensitivity), specificity, and F1 score can be calculated to

assess the model's overall performance and its performance for individual dialect classes.

Table 4.1: Performance Evaluation Table

	Predicted Yoruba	**Predicted Hausa**	**Predicted Igbo**
Actual Yoruba	TP (Yoruba)	FN (Yoruba)	FN (Yoruba)
Actual Hausa	FP (Hausa)	TP (Hausa)	FN (Hausa)
Actual Igbo	FP (Igbo)	FP (Igbo)	TP (Igbo)

TP (True Positive): The diagonal elements represent the number of instances where the model correctly predicted the corresponding dialect class. For example, TP(Yoruba) represents the number of instances correctly classified as Yoruba.

FN (False Negative): The elements in each row (excluding the diagonal) represent the number of instances of that actual dialect class that were incorrectly predicted as other dialect classes. For example, FN(Yoruba) represents the number of instances of Yoruba that were misclassified as Hausa or Igbo.

FP (False Positive): The elements in each column (excluding the diagonal) represent the number of instances of that predicted dialect class that belong to other actual dialect classes. For example, FP(Hausa) represents the number of instances predicted as Hausa but actually belong to Yoruba or Igbo.

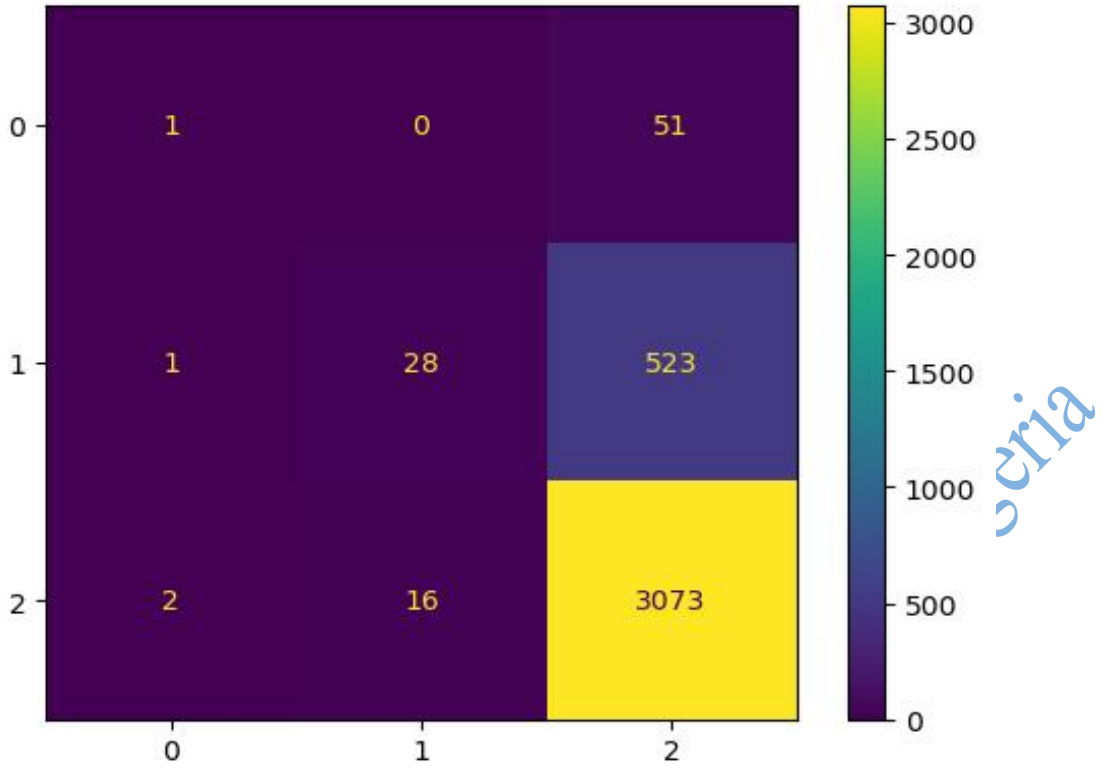


Figure 4.4: Confusion Matrix

Table 4.2: Precision, F - Score and Recall

	Precision	Recall	F1-score
0	0.25	0.02	0.04
1	0.64	0.05	0.09
2	0.84	0.99	0.91
Accuracy			**0.84**
Macro avg	0.58	0.35	0.35
Weighed avg	0.80	0.84	0.78

From the Table, Precision measures the accuracy of positive predictions made by the model. For class 0, the precision is low at 0.25, indicating that the model makes many false positive predictions for this class. For class 1, the precision is moderate at 0.64, indicating a higher accuracy in positive predictions. Class 2 has the highest precision at 0.84, meaning that the model's positive predictions for this class are more accurate. Recall measures the ability of the model to correctly identify positive instances from all actual positive instances. Class 0 has a low recall of 0.02, indicating that the model misses many actual positive instances for this class. Class 1 has a recall of 0.05,

indicating a similar issue. However, Class 2 has a high recall of 0.99, suggesting that the model is excellent at identifying positive instances for this class. he F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. Class 0 and Class 1 have low F1-scores at 0.04 and 0.09, respectively, indicating poor performance for these classes. Class 2 has a high F1-score of 0.91, indicating a strong balance between precision and recall for this class. The overall accuracy of the model is 0.84, indicating that the model correctly classifies 84% of all instances. However, as shown in the class-wise metrics, this accuracy is mainly driven by the high performance for Class 2, while Classes 0 and 1 have relatively poor performance.

The model gave a performance of 84%. This accuracy indicates how well the model was able to correctly identify the dialect of the audio samples it was tested on 84% performance implies that out of all the test instances (i.e., audio samples with known dialect labels) that the model was evaluated on, it correctly classified 84% of them with the correct dialect label. In other words, 84% of the audio samples were correctly identified with their respective dialects. The remaining 16% of the audio samples were misclassified, meaning the model incorrectly assigned them a dialect different from their true dialects.

4.5 Discussion of Results

The study focused on building a dialect detection model for the three major tribes in Nigeria: Yoruba, Hausa, and Igbo. The researchers collected voiced audio samples of speakers from these tribes speaking English from various platforms such as news media and radio recordings. The audio data was then preprocessed, and the voiced signal was converted from the time domain to the frequency domain using the Fourier transform. The training of the model was conducted using Matlab R2015A, a popular programming language and computing environment for numerical analysis and data

processing. The process involved reading in the voiced input and defining window size and hop size for analysis. Noise reduction techniques, such as applying a high-pass filter and spectral subtraction, were implemented to remove unwanted noise from the audio signals.

For feature extraction, Mel Frequency Cepstral Coefficients (MFCC) were calculated for each frame of the audio signals. The MFCC features were aggregated to obtain a fixed-length representation for each dialect sample. The model was then trained on this data using a classification algorithm. In the testing phase, the model was evaluated to ascertain its performance. The accuracy achieved by the model was 84%, indicating that it correctly identified the dialect of 84% of the test instances. The confusion matrix provided valuable insights into the model's performance, showing the number of true positives, false negatives, and false positives for each dialect class. Overall, the model's 84% accuracy suggests that it is capable of distinguishing between the three major dialects in Nigeria (Yoruba, Hausa, and Igbo) when speakers are speaking in English.

Comparing the result of this study with past literatures, in a work that proposed text-independent accent identification system using Gaussian mixture models (GMMs) for Kannada language reported a lesser accuracy to this research findings. The author conducted experiments using 32 speech samples from each region where each clip is of one minute duration spoken by native speakers. The baseline system implemented using MFCC features found to achieve 76.7 % accuracy[1].

Similarly, in in a study on automatic speech recognition and accent identification of ethnically diverse Nigerian English speakers which reported a lower accuracy than the result of this study which reported the accuracy of algorithm of the three major Nigerian languages (Yoruba, Igbo, and Hausa). Logistic regression emerged as the

best classifier in terms of accuracy (82%) ahead of K-Nearest Neighbor (75%) and Gaussian Mixture Model (50%)[2]. Also a work on accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features reported an accuracy is 51.92%, and the Unweighted Average Re-call (UAR) is 52.24% lesser than the studies findings[3].

Endnotes

1. R Soorajkumar, G.N Girish, P.B Ramteke, S.S Joshi & S.G Koolagudi. *Text-independent automatic accent identification system for Kannada language*. In *Proceedings of the International Conference on Data Engineering and Communication Technology: ICDECT 2017, Volume 2* (pp. 411-418). Springer Singapore.

2. F.O Oladipo, R.A Habeeb, A.E Musa, C Umezuruike, O.A Adeiza. *Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers.* **International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA** Volume 12– No.36, May 2021 – www.ijais.org

3. Y Jiao, M Tu, V Berisha, J.M Liss. *Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Feature*s. **InInterspeech** 2016 Sep (pp. 2388-2392).

Chapter Five

Conclusion

5.1 Summary of Results

The study aimed to acquire speech data from different Nigerian tribes speaking English. The researchers extracted voiced signals from the three major tribes in Nigeria, namely Yoruba, Hausa, and Igbo. To achieve this, they used APIs and web scraping from various platforms such as news media and audio recordings. Once the input signal was known, the next step was to determine its size. The voiced signal was in the time domain, so it was converted to the frequency domain using Fourier transform.

Preprocessing techniques such as filtering, windowing, or normalization were applied to ensure accurate frequency domain representation. The signal was then sampled at a regular interval to convert it into a discrete-time signal. The sampling rate was higher than twice the highest frequency present in the signal. After this, FFT was applied to calculate the Discrete Fourier Transform. The result was stored in an array called fft_result. The magnitude spectrum was calculated using the absolute value of the fft_result array. A corresponding frequency axis to represent the frequencies in the frequency domain was created.

The frequency axis values were determined by the sampling rate and the length of the FFT. The next step was to train the dataset. The researchers used Matlab R2015A, to ctreate a training model. The audio file was loaded into the model using the audioread function. This function allowed reading of audio files in various formats and storing of the audio data in a variable for further processing. Noise reduction was performed using spectral subtraction. Mel-frequency cepstral coefficients (MFCC) were used for

feature extraction. Classification was done using a classifier based on different labels available. The model gave a performance of 84%, indicating how well it was able to correctly identify the dialect of the audio samples tested.

5.2 Recommendation

The research has made significant progress in the development of a accent detection model for the three major tribes in Nigeria: Yoruba, Hausa, and Igbo, achieving an accuracy of 84%. This achievement is noteworthy as it demonstrates the model's ability to accurately distinguish between the dialects when speakers are speaking in English. The successful development of this model opens up exciting possibilities for various applications in the fields of speech recognition and language processing, particularly in Nigeria, where diverse dialects are spoken.

The potential applications of this dialect detection model are extensive. One of the primary areas where this technology can be utilized is in the field of speech recognition systems. Accurate identification of dialects can greatly enhance the performance and efficiency of speech recognition algorithms, enabling better comprehension and interaction with users from different regions of Nigeria. Additionally, this model can be integrated into language processing applications, facilitating automatic translation and language understanding tailored to specific dialects. The research contributes to the advancement of dialect identification technology in Nigeria, a country with a rich linguistic diversity. As dialects play a significant role in shaping communication and cultural identity, the successful development of this model can promote the preservation and recognition of various linguistic heritages within the country.

However, despite achieving an impressive accuracy of 84%, there is still room for improvement and further refinement of the model. As with any machine learning

model, exploring alternative classifiers or features can lead to enhanced performance and accuracy. Fine-tuning the parameters and conducting experiments with different feature extraction techniques may help in better representation and classification of dialects. Additionally, increasing the size and diversity of the dataset used for training can lead to a more robust and generalizable model.

The successful development of the dialect detection model and its high accuracy demonstrate promising prospects for its practical applications in speech recognition and language processing in Nigeria. The research contributes to the advancement of dialect identification technology, providing a foundation for further research and innovation in the field of linguistics. As the study highlights the importance of preserving and recognizing linguistic diversity, this model can foster a deeper appreciation of the rich cultural heritage embedded within Nigeria's diverse dialects. With continued refinement and exploration of alternative approaches, this research lays the groundwork for even more accurate and sophisticated dialect identification systems in the future.

Based on the findings from this study, the following recommendations were made:

i. Dataset Expansion: To further improve the accuracy and generalizability of the dialect detection model, it is recommended to expand the dataset used for training. Including more diverse and representative samples from different regions within each major tribe (Yoruba, Hausa, and Igbo) will allow the model to better capture the variations in dialects and accents. Moreover, considering additional languages spoken in Nigeria can help in developing a more comprehensive language processing system.

ii. Feature Engineering: Exploring alternative feature extraction techniques and linguistic features can be beneficial for enhancing the model's performance. Researchers can investigate the use of advanced techniques such as deep learning-based approaches, which can automatically learn relevant features from raw audio data. Additionally, incorporating prosodic features and linguistic information specific to each dialect may provide more discriminative characteristics for dialect identification.

iii. Hybrid Models: It is worth exploring the use of hybrid models that combine the strengths of different classifiers or machine learning techniques. Combining the outputs of multiple classifiers using ensemble methods like voting or stacking may result in improved performance and more reliable dialect predictions.

iv. Cross-Validation and Testing on Independent Data: To ensure the robustness of the model, it is essential to perform cross-validation using different subsets of the data during the training phase. Moreover, testing the model on an independent and larger dataset collected from diverse sources will provide a better assessment of its real-world performance and generalization ability.

v. Real-Time Implementation: Consideration should be given to developing a real-time implementation of the dialect detection model. This will enable its integration into various applications, such as voice assistants, call centers, and communication platforms, for seamless and accurate dialect recognition in real-world scenarios.

vi. Deployment in Local Languages: While the model has shown promising results for English-speaking speakers of different dialects, it is essential to extend its capabilities to recognize and identify local languages spoken in Nigeria. Developing dialect detection models for indigenous Nigerian languages can have

significant cultural and social impacts, promoting inclusivity and recognition of the nation's diverse linguistic heritage.

vii. Privacy and Ethical Considerations: When deploying the dialect detection model in real-world applications, it is crucial to address privacy and ethical concerns. Ensuring user consent, data security, and compliance with data protection regulations should be at the forefront of the model's deployment.

By considering these recommendations, future research can build upon the findings of this study and contribute to the advancement of dialect detection technology, language processing, and linguistics research in Nigeria. Implementing these suggestions will not only improve the accuracy and robustness of the model but also foster a deeper appreciation for Nigeria's linguistic diversity and cultural heritage.

5.3 Contribution to Knowledge

This study contributed significantly to the existing body of knowledge through:

i. Development of a Dialect Detection Model: The research successfully developed a dialect detection model specifically tailored for the three major tribes in Nigeria: Yoruba, Hausa, and Igbo. This model serves as a valuable addition to the field of speech recognition and language processing, as it addresses the unique linguistic characteristics and dialectal variations of these Nigerian tribes. By focusing on local languages, the model contributes to bridging the gap between traditional dialects and modern technological applications.

ii. Accurate Dialect Identification: The achieved accuracy of 84% demonstrates the model's effectiveness in accurately identifying the dialect of audio samples spoken in English. This level of accuracy is significant, considering the complexity of dialect variations in Nigeria and the challenges associated with dialect detection in multilingual and diverse linguistic contexts. The research

showcases the feasibility of building robust and efficient dialect identification systems for regional languages.

iii. Utilization of Real-World Data: The research gathered voiced audio samples from various real-world sources, including news media and radio recordings, capturing natural language usage in different contexts. By using authentic and diverse data, the model's training and evaluation reflect real-world scenarios, making its findings and applicability more relevant and practical.

iv. Exploration of Preprocessing Techniques: The research explored various preprocessing techniques, such as filtering, windowing, and normalization, to enhance the accuracy of the frequency domain representation. These techniques address challenges related to noise reduction and signal processing, contributing valuable insights into optimizing the preprocessing steps for dialect detection in similar contexts.

v. Cross-Cultural Linguistics: By focusing on dialects spoken in Nigeria, the research contributes to cross-cultural linguistics and sociolinguistics, shedding light on the diverse linguistic landscape of the country. The findings can foster a better understanding of language variation and cultural identity, promoting inclusivity and appreciation of linguistic diversity.

vi. Practical Applications: The developed dialect detection model has practical applications in various domains, including voice recognition, call center services, and multilingual communication platforms. Its accuracy and capability to distinguish between dialects open up opportunities for enhancing user experiences and language-specific services in Nigeria and other regions with linguistic diversity.

5.4 Suggestions for Further Research

Based on the findings of the current research, the following are the suggestions for further research.

i. Further research can extend the dialect identification model to include additional Nigerian languages to achieve a more comprehensive and representative language coverage.

ii. Deep Learning Approaches: Explore the application of deep learning techniques, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), for dialect identification. Deep learning models have shown promising results in various speech recognition tasks and may further enhance the accuracy and robustness of the dialect detection model.

iii. Dataset Augmentation: Augment the existing dataset with more diverse and representative speech samples from different regions and dialects. Dataset augmentation techniques, such as pitch variation, speed perturbation, and noise addition, can help improve the model's performance and generalization to different dialectal variations.

iv. Multilingual Dialect Identification: Investigate the development of a multilingual dialect identification model capable of identifying dialects from multiple languages simultaneously. This research can be particularly useful in regions with linguistic diversity and multiple official languages.

v. Real-Time Dialect Identification: Design and implement a real-time dialect identification system that can process and identify dialects in streaming audio or live speech. This application can be beneficial in call centers, language learning platforms, and voice-controlled devices.

vi. Dialectal Variation Analysis: Investigate the specific linguistic features that contribute to dialectal variations in different Nigerian languages. Understanding the linguistic characteristics of each dialect can lead to more informed and targeted approaches for dialect identification.

Biblography

Book

McArthur T, Lam-McArthur J, Fontaine L, editors. *Oxford companion to the English language*. Oxford University Press; 2018 May 14.

International Conference

Ajayi LK, Azeta A, Odun-Ayo I, Aniemeka ET. *Acoustic nudging-based model for vocabulary reformulation in continuous Yorùbá speech recognition*. International conference on computational science and its applications 2022 (pp. 494-508). Springer, Cham.

Ambili AR, Roy RC. *Multi tasking synthetic speech detection on Indian Languages.* In 2022 International conference on innovative trends in information technology (ICITIIT) 2022 Feb 12 (pp. 1-6). IEEE.

Atanda A, Yusof S, HariharanM. *Yorùbá automatic speech recognition: A review*. In rural ict development (RICTD) international conference 2013 (Vol. 1, No. 1, pp. 116-121).

Batzorig Z, Bukhtsooj O, Chensky AG, Galbaatar T. *Speech recognition in Mongolian language using a neural network with pre-processing technique*. In2020 international youth conference on radio electronics, electrical and power engineering (REEPE) 2020 Mar 12 (pp. 1-5). IEEE

Darshana S, Theivaprakasham H, Lal GJ, Premjith B, Sowmya V, Soman K. MARS: *A hy.brid deep cnn-based multi-accent recognition system for English Language*. In 2022 first international conference on artificial intelligence trends and pattern recognition (ICAITPR) 2022 Mar 10 (pp. 1-6). IEEE.)

Hasija T, Kadyan V, Guleria K. *Recognition of children Punjabi speech using tonal non-tonal classifier.* In2021 international conference on emerging smart computing and informatics (ESCI) 2021 Mar 5 (pp. 702-706). IEEE.

Laleye FA, Besacier L, Ezin EC, Motamed C. *First automatic fongbe continuous speech recognition system: development of acoustic models and language models.* In2016 federated conference on computer science and information systems (FedCSIS) 2016 Sep 11 (pp. 477-482). IEEE.

Mardhotillah R, Dirgantoro B, Setianingsih C. *Speaker recognition for digital forensic audio analysis using support vector machine.* In 2020 3rd international seminar on research of information technology and intelligent systems (ISRITI) 2020 Dec 10 (pp. 514-519). IEEE.

Muttaqi M, Degirmenci A, Karal O. *US accent recognition using machine learning methods.* In 2022 innovations in intelligent systems and a conference (ASYU) 2022 Sep 7 (pp. 1-6). IEEE

Nasrun M, Setianingsih C. *Human emotion detection with speech recognition using mel-frequency cepstral coefficient and support vector machine.* In 2021 international conference on artificial intelligence and mechatronics systems (AIMS) 2021 Apr 28 (pp. 1-6). IEEE

Prasangini N, Nagahamulla H. *Sinhala speech to sinhala unicode text conversion for disaster relief facilitation in Sri Lanka.* In 2018 IEEE International Conference on Information and Automation for Sustainability (ICIAfS) 2018 Dec 21 (pp. 1-6). IEEE.

Purwar A, Sharma H, Sharma Y, Gupta H, Kaur A. *Accent classification using machine learning and deep learning models.* In2022 1st International Conference on Informatics (ICI) 2022 Apr 14 (pp. 13-18). IEEE.

Radha K, Bansal M, Shabber SM. *Accent classification of native and non-native children using harmonic pitch.* In2022 2nd International Conference on Artificial Intelligence and Signal Processing (AISP) 2022 Feb 12 (pp. 1-6). IEEE.

Rajadnya V & Joshi K. *Raga Classification Based on MFCC and Variants.* In 2021 IEEE 2nd International conference on technology, engineering, management for societal impact using marketing, entrepreneurship and talent (TEMSMET) (pp. 1-6). IEEE. (2021, December)

Solano RC. *Explicit tone transcription improves asr performance in extremely low-resource languages: A case study in Bribri*. In Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas 2021 Jun (pp. 173-184

Sönmez YÜ, Varol A. *In-Depth analysis of speech production, auditory system, emotion theories and emotion recognition.* In 2020 8th International Symposium on Digital Forensics and Security (ISDFS) 2020 Jun 1 (pp. 1-8). IEEE

Stefanenko OS, Lipinskiy LV, PolyakoJ AS, Khudonogova JA, Semenkin ES. *An intelligent voice recognition system based on fuzzy logic and the bag-of-words*

technique. InIOP Conference Series: Materials science and engineering 2022 Mar 1 (Vol. 1230, No. 1, p. 012020). IOP Publishing.

Yusof SA, Atanda AF, Hariharan M. *A review of Yoruba automatic speech recognition.* In2013 IEEE 3rd international conference on system engineering and technology 2013 Aug 19 (pp. 242-247). IEEE.).

Zhou J, Hu X, Ma Q. *A study of the emotional information acoustic characteristics of synthetic speech phoneme/ei.* In international conference on electronic information engineering and computer communication (EIECC 2021) 2022 May 4 (Vol. 12172, pp. 170-178). SPIE.

Journals

Abdou SM, Moussa AM. *Arabic speech recognition: challenges and state of the art. Computational linguistics, speech and image processing for Arabic language.* 2019:1-27

Abdulsatar AA, Davydov VV, Yushkova VV, Glinushkin AP, Rud VY. *Age and gender recognition from speech signals*. **In Journal of Physics: Conference Series 2019** Dec 1 (Vol. 1410, No. 1, p. 012073). IOP Publishing.

Abdulwahab AF, Mohd Yusof SA, Husni H. *Acoustic comparison of Malaysian and Nigerian English accents.* **Journal of Telecommunication, Electronic and Computer Engineering (JTEC)** 9 (3-5), 141-146, 2017

Abdusalomov AB, Safarov F, Rakhimov M, Turaev B, Whangbo TK. *Improved feature parameter extraction from speech signals using machine learning algorithm*. Sensors. 2022 Oct 24;22(21):8122.

Adam EE. *Deep learning based NLP techniques in text to speech synthesis for communication recognition.* **Journal of Soft Computing Paradigm (JSCP).** 2020 Dec 18;2(04):209-15

Adetunmbi OA, Obe OO, Iyanda JN. *Development of standard Yorùbá speech-to-text system using HTK*. **International Journal of Speech Technology**. 2016 Dec;19(4):929-44.

Akinwonmi AE. *Development of a prosodic read speech syllabic corpus of the Yoruba language.* Development. 2021 Jun;7(36).

Al-Anzi FS, AbuZeina D. *The capacity of mel frequency cepstral coefficients for speech recognition.* **International Journal of Computer and Information Engineering.** 2017 Sep 1;11(10):1149-53.

Alyafeai Z, Al-shaibani MS, Ghaleb M, Ahmad I. *Evaluating various tokenizers for Arabic text classification.* Neural Processing Letters. 2022 Aug 18:1-23

Ambrogio S, Narayanan P, Tsai H, Shelby RM, Boybat I, Di Nolfo C, Sidler S, Giordano M, Bodini M, Farinha NC, Killeen B. *Equivalent-accuracy accelerated neural-network training using analogue memory.* Nature. 2018 Jun;558(7708):60-7.

Anwar MZ, Kaleem Z, Jamalipour A. *Machine learning inspired sound-based amateur drone detection for public safety applications.* IEEE Transactions on Vehicular Technology. 2019 Jan 17;68(3):2526-34.

Aouragh SL, Yousfi A, Laaroussi S, Gueddah H, Nejja M. *A new estimate of the n-gram language model.* Procedia Computer Science. 2021 Jan 1;189:211-5.

Araújo Alves J, Neto Paiva F, Torres Silva L, Remoaldo P. *Low-frequency noise and its main effects on human health—a review of the literature between 2016 and 2019.* Applied Sciences. 2020 Jul 28;10(15):5205.

Avci U. *A pattern mining approach for improving speech emotion recognition.* **International Journal of Pattern Recognition and Artificial Intelligence.** 2022 Nov 24:2250045.

Babii H, Janes A, Robbes R. *Modeling vocabulary for big code machine learning.* arXiv preprint arXiv:1904.01873. 2019 Apr 3.

Becerra A, De La Rosa JI, González E. *Speech recognition in a dialog system: from conventional to deep processing*. Multimedia Tools and Applications. 2018 Jun;77(12):15875-911.

Bernstein JG, Stakhovskaya OA, Jensen KK, Goupell MJ. *Acoustic hearing can interfere with single-sided deafness cochlear-implant speech perception*. Ear and hearing. 2020 Jul;41(4):747.

Bezzam E, Kashani S, Hurley P, Simeon Mi. *pyFFS: A python library for fast fourier series computation.* arXiv preprint arXiv:2110.00262. 2021 Oct 1.

Bhardwaj V, Ben Othman MT, Kukreja V, Belkhier Y, Bajaj M, Goud BS, Rehman AU, Shafiq M, Hamam H. *Automatic speech recognition (asr) systems for children: a systematic literature review.* Applied Sciences. 2022 Apr 27;12(9):4419.

Bhardwaj V, Kukreja V. *Effect of pitch enhancement in Punjabi children's speech recognition system under disparate acoustic conditions.* Applied Acoustics. 2021 Jun 1;177:107918.

Bhatt S, Jain A, Dev A. *Feature extraction techniques with analysis of confusing words for speech recognition in the Hindi language.* Wireless Personal Communications. 2021 Jun;118(4):3303-33.

Bigi B, Abiola OS, B Caron. *Resources and tools for automated speech segmentation of the African language naija (Nigerian Pidgin).* InLanguage and Technology Conference 2020 (pp. 164-173). Springer, Cham.

Bloomfield L, Lane E, Mangalam M, Kelty-Stephen DG. *Perceiving and remembering speech depend on multifractal nonlinearity in movements producing and exploring speech*. **Journal of the Royal Society Interface**. 2021 Aug 4;18(181):20210272.

Borisagar KR, Thanki RM, Sedani BS. *Generation of speech signal and its characteristics.* In speech enhancement techniques for digital hearing aids 2019 (pp. 13-27). Springer, Cham

Bostrom K, Durrett G. *Byte pair encoding is suboptimal for language model pretraining.* arXiv preprint arXiv:2004.03720. 2020 Apr 7.

Carr CT. *Computer-mediated communication: a theoretical and practical introduction to online human communication*. Rowman & Littlefield; 2021 Apr 29.

Choutri K, Lagha M, Meshoul S, Batouche M, Kacel Y, Mebarkia N. *A Multi-Lingual Speech Recognition-Based Framework to Human-Drone Interaction*. Electronics. 2022 Jun 9;11(12):1829

Chowdhary K. *Natural language processing. Fundamentals of artificial intelligence*. 2020:603-49.

Chrupała G. *Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques.* **Journal of Artificial Intelligence Research.** 2022 Feb 18;73:673-707.

Contreras R, Ayala A, Cruz F. *Unmanned aerial vehicle control through domain-based automatic speech recognition. Computers*. 2020 Sep 19;9(3):75.

Coto-Solano R. *Computational sociophonetics using automatic speech recognition*. Language and Linguistics Compass 2022;16. https://doi.org/10.1111/lnc3.12474.

Coto-Solano R. *Computational sociophonetics using automatic speech recognition. Language and Linguistics Compass*. 2022 Sep;16(9):e12474.

Dargan S, Kumar M, Ayyagari MR, Kumar G. *A survey of deep learning and its applications: a new paradigm to machine learning.* Archives of Computational Methods in Engineering. 2020 Sep;27(4):1071-92.

Das N, Chakraborty S, Chaki J, Padhy N, Dey N. *Fundamentals, present and future perspectives of speech enhancement.* **International Journal of Speech Technology.** 2021 Dec;24(4):883-901.

Debnath S, Roy P. *Automatic speech recognition based on clustering technique.* In Emerging Technology in Modelling and Graphics 2020 (pp. 679-688). Springer, Singapore.

Deshwal D, Sangwan P, Kumar D. *Feature extraction methods in language identification: a survey.* Wireless Personal Communications. 2019 Aug;107(4):2071-103.

Ding ZW, Li XF, Huang X, Wang MB, Tang QB, Jia JD. *Feature extraction, recognition, and classification of acoustic emission waveform signal of coal rock sample under uniaxial compression.* **International Journal of Rock Mechanics and Mining Sciences**. 2022 Dec 1;160:105262.

Dua M, Kadyan V, Banthia N, Bansal A, Agarwal T. *Spectral warping and data augmentation for low resource language ASR system under mismatched conditions.* Applied Acoustics. 2022 Mar 15;190:108643

ElBedwehy MN, Behery GM, Elbarougy R. *Emotional speech recognition based on weighted distance optimization system*. **International Journal of Pattern Recognition and Artificial Intelligence.** 2020 Oct 19;34(11):2050027

Elelu K, Le T, Le C. *Collision hazard detection for construction worker safety using audio surveillance.* **Journal of Construction Engineering and Management.** 2023 Jan 1;149(1):04022159

Fréjus AA LAleye, Laurent Besacier, Eugène C Ezin, Cina Motamed. *First automatic fongbe continuous speech recognition system: Development of acoustic models and language models.* Federated Conference on Computer Science and Information Systems (FedCSIS), 477-482, 2016

Haeb-Umbach R, Watanabe S, Nakatani T, Bacchiani M, Hoffmeister B, Seltzer ML, Zen H, Souden M. *Speech processing for digital home assistants: Combining signal processing with deep-learning techniques.* IEEE Signal processing magazine. 2019 Oct 30;36(6):111-24

Haq AS, Nasrun M, Setianingsih C, Murti MA. *Speech recognition implementation using MFCC and DTW algorithm for home automation.* Proceeding of the Electrical Engineering Computer Science and Informatics. 2020 Oct;7(2):78-85

Hedger SC, Johnsrude IS. *Speech perception under adverse listening conditions*. In Speech Perception 2022 (pp. 141-171). Springer, Cham.

Hixon TJ, Weismer G, Hoit JD. *Preclinical speech science: anatomy, physiology, acoustics, and perception*. Plural Publishing; 2018 Aug 31.

Hwang I, Chang JH. *End-to-end speech endpoint detection utilizing acoustic and language modeling knowledge for online low-latency speech recognition.* IEEE Access. 2020 Aug 31;8:161109-23

Ibrahim YA, Faki SA, Abidemi TI. *Automatic speech recognition using mfcc in feature extraction based hmm for human computer interaction in Hausa. annals.* Computer Science Series. 2019 Dec 1;17(2).

Jagadeeshwar K, Sreenivasarao T, Pulicherla P, Satyanarayana KN, Mohana Lakshmi K, Kumar PM. *ASERNet: Automatic speech emotion recognition system using MFCC-based LPC approach with deep learning CNN.* **International Journal of Modeling, Simulation, and Scientific Computing.** 2022 Nov 30:2341029.

Jain P, Kasture NR, Kumar T. *Comparative study of speaker recognition techniques in iot devices for text independent negative recognition.* Scalable Computing: Practice and Experience. 2020 Aug 1;21(3):359-68.

Jiao Y, Tu M, Berisha V, Liss JM. *Accent identification by combining deep neural networks and recurrent neural networks trained on long and short term features*. **InInterspeech** 2016 Sep (pp. 2388-2392).

Kapil P, Ekbal A. *A deep neural network based multi-task learning approach to hate speech detection.* Knowledge-Based Systems. 2020 Dec 27;210:106458.

Kaur J, Singh A, Kadyan V. *Automatic speech recognition system for tonal languages: State-of-the-art survey.* Archives of Computational Methods in Engineering. 2021 May;28(3):1039-68.

Koenecke A, Nam A, Lake E, Nudell J, Quartey M, Mengesha Z, Toups C, Rickford JR, Jurafsky D, Goel S. *Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences*. 2020 Apr 7;117(14):7684-9.

Krizhevsky A, Sutskever I, Hinton GE. *Image net classification with deep convolutional neural networks.* Communications of the ACM. 2017 May 24;60(6):84-90.

Kumar T, Mahrishi M, Meena G. *A comprehensive review of recent automatic speech summarization and keyword identification techniques*. Artificial Intelligence in Industrial Applications. 2022:111-26

Lai Y. *Application of the artificial intelligence algorithm in the automatic segmentation of Mandarin dialect accent.* Mobile Information Systems. 2022 Feb 24;2022

Lee MC, Yeh SC, Chang JW, Chen ZY. *Research on Chinese speech emotion recognition based on deep neural network and acoustic features. Sensors.* 2022 Jun 23;22(13):4744

Lesnichaia M, Mikhailava V, Bogach N, Lezhenin I, Blake L, Pyshkin E. *Classification of accented english using cnn model trained on amplitude mel-spectrograms.* Proc. Interspeech 2022. 2022:3669-73

Liu AT, Li SW, Lee HY. Tera: *Self-supervised learning of transformer encoder representation for speech. IEEE/ACM transactions on audio, speech, and language processing.* 2021 Jul 8;29:2351-66.

Liu B, Ding X, Cai H, Zhu W, WangZ, Liu W, Yang J. *Precision adaptive MFCC based on R2SDF-FFT and approximate computing for low-power speech*

keywords recognition. IEEE Circuits and Systems Magazine. 2021 Nov 15;21(4):24-39.

Liu H, Lang B, Liu M, Yan H. *CNN and RNN based payload classification methods for attack detection. Knowledge-Based Systems*. 2019 Jan 1;163:332-41.

Liu W, Liao Q, Qiao F, Xia W, Wang C, Lombardi F. *Approximate designs for fast Fourier transform (FFT) with application to speech recognition.* IEEE Transactions on Circuits and Systems I: Regular Papers. 2019 Aug 23;66(12):4727-39

Liu Y, Qian Y, Chen N, Fu T, Zhang Y, Yu K. *Deep feature for text-dependent speaker verification.* Speech Communication. 2015 Oct 1;73:1-3

Lokesh S, Devi MR. *Speech recognition system using enhanced mel frequency cepstral coefficient with windowing and framing method.* Cluster Computing. 2019 Sep;22(5):11669-79.

Martinez A, Sudoh K, Matsumoto Y. *Sub-subword N-Gram features for subword-level neural machine translation.* **Journal of Natural Language Processing.** 2021;28(1):82-103

Mazzei D, Chiarello F, Fantoni G. *Analyzing social robotics research with natural language processing techniques*. Cognitive Computation. 2021 Mar;13(2):308-21.

Mikhailava V, Lesnichaia M, Bogach N, Lezhenin I, Blake J, Pyshkin E. *Language accent detection with cnn using sparse data from a crowd-sourced speech archive. Mathematics. 2022 Aug 13;10(16):2913.*

Mittal A, Dua M. *Automatic speaker verification systems and spoof detection techniques: review and analysis.* **International Journal of Speech Technology.** 2022 Mar;25(1):105-34

Mohammadpour L, Ling TC, Liew CS, Aryanfar A. *A survey of CNN-based network intrusion detection.* Applied Sciences. 2022 Aug 15;12(16):8162.

Mourad T. *Arabic speech recognition by stationary bionic wavelet transform and mfcc using a multi-layer perceptron for voice control*. In the stationary bionic wavelet transform and its applications for ECG and speech processing 2022 (pp. 69-81). Springer, Cham.)

Narendra NP, Schuller B, Alku P. *The detection of Parkinson's disease from speech using voice source information.* IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021 May 7;29:1925-36.

Nassif AB, Shahin I, Lataifeh M, Elnagar A, Nemmour N. *Empirical comparison between deep and classical classifiers for speaker verification in emotional talking environments.* Information. 2022 Sep 27;13(10):456.

Oladipo F, Habeeb RA, Musa AE. *Accent identification of ethnically diverse Nigerian English speakers*. Available at SSRN 3666815. 2020 Jul 24.

Oladipo FO, Habeeb RA, Musa AE, C, Adeiza OA. *Automatic speech recognition and accent identification of ethnically diverse Nigerian English speakers.* **International Journal of Applied Information Systems (IJAIS) –** ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 12– No.36, May 2021

Oladipo F, Habeeb RA, Musa AE. *Accent identification of ethnically diverse Nigerian English speakers*. SSRN Electronic Journal 2020. https://doi.org/10.2139/ssrn.3666815.

Priscilla SJ, Vanithalakshmi M. *Aggression monitoring in speech using semantics and pitch.* **Global Journal of Pure and Applied Mathematics.** 2017;13(9):5437-45.

Radha K, Bansal M. *Audio augmentation for non-native children's speech recognition through discriminative learning*. Entropy. 2022 Oct 19;24(10):1490.

Radzikowski K, Nowak R, Wang L, Yoshie O. *Dual supervised learning for non-native speech recognition*. **EURASIP Journal on Audio, Speech, and Music Processing**. 2019 Dec;2019(1):1-0.

Radzikowski K, Wang L, Yoshie O, Nowak R. *Accent modification for speech recognition of non-native speakers using neural style transfer*. **EURASIP Journal on Audio, Speech, and Music Processing**. 2021 Dec;2021(1):1-0.

Räsänen O, Seshadri S, Lavechin M, Cristia A, Casillas A. ALICE: *An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings.* Behavior Research Methods. 2021 Apr;53(2):818-35.

Salau AO, Olowoyo TD, Akinola SO. *Accent classification of the three major nigerian indigenous languages using 1d cnn lstm network model.* In Advances in Computational Intelligence Techniques 2020 (pp. 1-16). Springer, Singapore

Salau AO, Olowoyo TD, Akinola SO. *Accent classification of the three major nigerian indigenous languages using 1d cnn lstm network model*. InAdvances in Computational Intelligence Techniques 2020 (pp. 1-16). Springer, Singapore.

Sharma U, Om H, Mishra AN. *HindiSpeech-Net: a deep learning based robust automatic speech recognition system for Hindi language.* Multimedia Tools and Applications. 2022 Oct 24:1-21

Simha KP. *Improving automatic speech recognition on endangered languages*. Rochester Institute of Technology; 2019.

Singh A, Kaur N, Kukreja V, Kadyan V, Kumar M. *Computational intelligence in processing of speech acoustics: a survey*. Complex & Intelligent Systems. 2022 Feb 17:1-39.

Song Z. *English speech recognition based on deep learning with multiple features.* Computing. 2020 Mar;102(3):663-82.

Soorajkumar R, Girish GN, Ramteke PB, Joshi SS & Koolagudi SG. *Text-independent automatic accent identification system for Kannada language*. In Proceedings of the International Conference on Data Engineering and Communication Technology: ICDECT 2017, Volume 2 (pp. 411-418). Springer Singapore.

Stampfl AP, Liu Z, Hu J, Sawada K, Takano H, Kohmura Y, Ishikawa T, Lim JH, Je JH, Low CM, Teo A. Synapse: a*n international roadmap to large brain imaging*. Physics Reports. 2023 Feb 9;999:1-60.

Stenman M, 2015. *Automatic speech recognition An evaluation of Google Speech.*

Wang D, Wang X, Lv S. *An overview of end-to-end automatic speech recognition.* Symmetry. 2019 Aug 7;11(8):1018.

Wang G, Wang X, Zhao C. *An iterative hybrid harmonics detection method based on discrete wavelet transform and bartlett–hann window*. Applied Sciences 2020;10:3922. https://doi.org/10.3390/app10113922.

Wang H, Wang D. *Towards robust speech super-resolution.* IEEE/ACM transactions on audio, speech, and language processing. 2021 Jan 25;29:2058-66.

Wang J, Li B, Zhang J. Use Brain-Like *audio features to improve speech recognition performance*. **Journal of Sensors** 2022;2022:1–12. https://doi.org/10.1155/2022/6742474.

Wubet YA, Lian KY. *Voice conversion based augmentation and a hybrid CNN-LSTM model for improving speaker-independent keyword recognition on limited datasets.* IEEE Access. 2022 Aug 19;10:89170-80.

Yang R , Cheng G, Miao H, Li T, Zhang P, Yan Y. *Keyword search using attention-based end-to-end asr and frame-synchronous phoneme alignments.* IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2021 Oct 15;29:3202-15.

Yusnita MA, Paulraj MP, Sazali Yaacob RY, Shahriman AB. A*nalysis of accent-sensitive words in multi-resolution mel-frequency cepstral coefficients for classification of accents in Malaysian English*. **International Journal of Automotive and Mechanical Engineering**. 2013;7:1053-73

Zaidi BF, Selouani SA, Boudraa, Sidi Yakoub M. *Deep neural network architectures for dysarthric speech analysis and recognition.* Neural Computing and Applications. 2021 Aug;33(15):9089-108

Zakari RY, Lawal ZK, Abdulmumin I. *A systematic literature review of Hausa natural language processing*. **International Journal of Computer and Information Technology** (2279-0764). 2021 Jul 31;10(4)

Zhang C, Lu Y. *Study on artificial intelligence: The state of the art and future prospects*. **Journal of Industrial Information Integration**. 2021 Sep 1;23:100224.

Zhang F, Han S, Gao H, Wang T. *A Gaussian mixture based hidden Markov model for motion recognition with 3D vision device*. Computers & Electrical Engineering. 2020 May 1;83:106603

Zielonka M, Piastowski A, Czyżewski A, Nadachowski P, Operlejn M, Kaczor K. *Recognition of emotions in speech using convolutional neural networks on different datasets.* Electronics. 2022 Jan;11(22):3831.

Zonooz B, Arani E, Körding KP, Aalbers PA, Celikel T, Van Opstal AJ. *Spectral weighting underlies perceived sound elevation.* Scientific reports. 2019 Feb 7;9(1):1-2

Thesis

Bensch C. C*ontinuous learning in automatic speech recognition* (Doctoral dissertation, Maastricht University).2021

Bhabad SSS. *Speech recognition & rectification for articulatory handicapped people* (Doctoral Dissertation, Savitribai Phule Pune University). 2019

Ibrahim SA. *Speech recognition based on convolutional neural networks* (Doctoral dissertation, University of Gezira).2020

Kruthika P S. *Improving automatic speech recognition on endanger omatic speech recognition on endangered languages.* Thesis. Rochester Institute of Technology, 2019

Mikuöová I. *Estimating vocal tract resonances of synthesized high-pitched vowels using CNN* (Doctoral dissertation, Technische Universität Wien)

Muhammad UG. *A comparative phonological analysis of varieties of english spoken by native speakers of Nigerian languages (Hausa, Igbo, Kanuri and Yoruba) for the determination of speakers' origins* (Doctoral dissertation, University of York)

Onyenwe IE. *Developing methods and resources for automated processing of the african language igbo* (Doctoral dissertation, University of Sheffield).2017

Appendices

Appendix I: Source Code

```
% Load audio file
[y,fs] = audioread('audio_file.wav');
 11111111
% Define window size and hop size
winSize = 1024;
hopSize = 512;

% Compute spectrogram
spectrogram = stft(y, winSize, hopSize, winSize, fs);

% Compute noise spectrum using first 5 frames of spectrogram
noiseSpect = mean(abs(spectrogram(:,1:5)).^2, 2);

% Initialize output matrix
spectrogramClean = zeros(size(spectrogram));

% Spectral subtraction
for i = 1:size(spectrogram,2)
    % Compute magnitude spectrum of current frame
    magnSpec = abs(spectrogram(:,i)).^2;
```

```
    % Subtract noise spectrum from magnitude spectrum
    magnSpecClean = max(magnSpec - noiseSpect, 0);

    % Reconstruct complex spectrum
    spectrogramClean(:,i) = sqrt(magnSpecClean) .* exp(1i*angle(spectrogram(:,i)));
end

% Inverse STFT to obtain cleaned signal
yClean = istft(spectrogramClean, winSize, hopSize, winSize, fs);

% Play cleaned audio
sound(yClean, fs);

% Write cleaned audio to file
audiowrite('audio_file_clean.wav', yClean, fs);
```

```
% Read the audio file
[x, fs] = audioread('test.aac');

% Define the parameters for the noise reduction filter
frame_len = round(fs*0.02); % frame length of 20ms
overlap_len = round(fs*0.01); % 50% overlap
freq_cutoff = 1000; % cutoff frequency for high-pass filter
noise_reduction = 10; % noise reduction level in dB

% Create a high-pass filter to remove low-frequency noise
hp_filter = designfilt('highpassiir', 'FilterOrder', 8, 'PassbandFrequency', freq_cutoff,
'PassbandRipple', 0.2, 'SampleRate', fs);

% Apply the high-pass filter to the audio signal
x_filt = filtfilt(hp_filter, x);

% Perform noise reduction using spectral subtraction
spectrogram = stft(x_filt, frame_len, overlap_len, hann(frame_len));
spectrogram_noise = stft(x_filt, frame_len, overlap_len, hann(frame_len));
spectrogram_noise = max(spectrogram_noise, [], 2); % estimate the noise power
spectrum
spectrogram_clean = max(spectrogram - noise_reduction, 0); % subtract noise power
spectrum from original spectrogram
x_clean = istft(spectrogram_clean, frame_len, overlap_len, hann(frame_len)); %
convert back to time domain

% Write the cleaned audio to a new file
audiowrite('audio_file_cleaned.wav', x_clean, fs);
```

This code uses a high-pass filter to remove low-frequency noise and then applies spectral subtraction to further reduce the noise.
The resulting cleaned audio is saved as a new file.
You may need to adjust the filter parameters and noise reduction level based on your specific audio file and noise characteristics.

updated version

```
% Read the audio file
[x, fs] = audioread('test.aac');

% Define the parameters for the noise reduction filter
frame_len = round(fs*0.02); % frame length of 20ms
overlap_len = round(fs*0.01); % 50% overlap
freq_cutoff = 1000; % cutoff frequency for high-pass filter
noise_reduction = 10; % noise reduction level in dB

% Create a high-pass filter to remove low-frequency noise
hp_filter = designfilt('highpassiir', 'FilterOrder', 8, 'PassbandFrequency', freq_cutoff, 'PassbandRipple', 0.2, 'SampleRate', fs);

% Apply the high-pass filter to the audio signal
x_filt = filtfilt(hp_filter, x);

% Perform noise reduction using spectral subtraction
[spec, f, t] = spectrogram(x_filt, hann(frame_len), overlap_len, [], fs);
[spec_noise, ~, ~] = spectrogram(x_filt, hann(frame_len), overlap_len, [], fs);
spec_noise = max(spec_noise, [], 2); % estimate the noise power spectrum
spec_clean = max(spec - noise_reduction, 0); % subtract noise power spectrum from original spectrogram
x_clean = istft(spec_clean, hann(frame_len), overlap_len, length(x), fs); % convert back to time domain

% Write the cleaned audio to a new file
audiowrite('audio_file_cleaned.wav', x_clean, fs);
```

Noice reduction module

```
L = size(St,1);
T = 1/Fs;
t = (0:L-1)*T;
% Input signal chart S(t)
figure
subplot(2,3,1);
plot(t,St);grid;
title('Signal S(t)')
xlabel('Time,s')
ylabel('Signal amplitude S(t)')
% Fourier transform of the input signal S(t)
```

```
% NFFT = 2^nextpow2(L);                    % Next power of 2 from
length of L
NFFT=2^16;
Sf = fft(St,NFFT)/NFFT;
f =Fs/2*linspace(0,1,NFFT/2);
Z=2*abs(Sf(1:NFFT/2));
% Input signal spectrum chart Sf(f)
subplot(2,3,2);
plot(f(1:NFFT/2),20*log10(Z(1:NFFT/2)));grid;
title('Signal spectrum S(t)')
xlabel('Frequency (Hz)')
ylabel('Signal magnitude |Sf(f)|, dB')
% Band filter
[b,a]=ellip(4,0.001,30,[300 3400]*2/Fs);
[H,w]=freqz(b,a,NFFT);
subplot(2,3,3);
plot(w(1:NFFT)*Fs/(2*pi),abs(H(1:NFFT)));grid;          % Frequency
response of the filter
title('Frequency response')
xlabel('Frequency (Hz)')
ylabel('Response factor')
SL=filter(b,a,St);
% Output signal chart SL(t)
subplot(2,3,4);
plot(t,SL);grid;
title('Signal SL(t)')
xlabel('Time,s')
ylabel('Signal amplitude SL(t)')
% Fourier transform of the output signal (after a filtering)
SLf = fft(SL,NFFT)/NFFT;
% ff =Fs/2*linspace(0,1,NFFT/2);
ZZ=2*abs(SLf(1:NFFT/2));
% Output (filtered) signal spectrum chart SLf(f)
subplot(2,3,5);
plot(f(1:NFFT/2),20*log10(ZZ(1:NFFT/2)));grid;
title('Signal spectrum SL(t)')
xlabel('Frequency (Hz)')
ylabel('Signal magnitude |SLf(f)|, dB')
% Write to disk
%wavwrite(SL,Fs,'output.wav');
Powf1 = trapz(f,Z.^2)
Powf2 = trapz(f,ZZ.^2)
Sp1 = trapz(t,St.^2)
Sp2 = trapz(t,SL.^2)
Ratio = Powf2/Powf1
```

1. Feature extraction

```
[ceps,freqresp,fb,fbrecon,freqrecon] = mfcc(SL,8000);
melC= imresize(ceps,[100 100]);
as = reshape(melC,[],1);
```

```
            features = [features  as];
            labels = [labels,1];
```

2. Step 1-4 is done for the training
3. For the testing step 1-4 is repeated
4. Classification

```
    trainedClassifier2 = fitcknn( ...
   qrwe', ...
    labels, ...
   'Distance','euclidean', ...
   'NumNeighbors',3, ...
   'DistanceWeight','squaredinverse', ...
   'Standardize',false, ...
   'ClassNames',unique(labels));
%%%%%%%% change for between R2 here [13]
qmsasrwe =imresize(msas,[100 34]);
```

5. predictionn = predict(trainedClassifier2,qmsasrwe');

6. the output of the predictor is now the used to determine the dialect of the speaker

7. labels
8. predictionn = int16(predictionn);
9. if (predictionn == 1)
10. h = msgbox('Yoruba');
11. elseif(predictionn == 2)
12. h = msgbox('hausa');
13. else
14. h = msgbox('igbo');
15. end
16.

```
function varargout = HOME(varargin)
% HOME M-file for HOME.fig
%      HOME, by itself, creates a new HOME or raises the existing
%      singleton*.
%
%      H = HOME returns the handle to a new HOME or the handle to
%      the existing singleton*.
%
%      HOME('CALLBACK',hObject,eventData,handles,...) calls the local
%      function named CALLBACK in HOME.M with the given input arguments.
%
%      HOME('Property','Value',...) creates a new HOME or raises the
%      existing singleton*.  Starting from the left, property value pairs are
%      applied to the GUI before HOME_OpeningFcn gets called.  An
%      unrecognized property name or invalid value makes property application
%      stop.  All inputs are passed to HOME_OpeningFcn via varargin.
%
%      *See GUI Options on GUIDE's Tools menu.  Choose "GUI allows only one
```

```
%      instance to run (singleton)".
%
% See also: GUIDE, GUIDATA, GUIHANDLES

% Edit the above text to modify the response to help HOME

% Last Modified by GUIDE v2.5 15-May-2013 23:55:01

% Begin initialization code - DO NOT EDIT
gui_Singleton = 1;
gui_State = struct('gui_Name',       mfilename, ...
                   'gui_Singleton',  gui_Singleton, ...
                   'gui_OpeningFcn', @HOME_OpeningFcn, ...
                   'gui_OutputFcn',  @HOME_OutputFcn, ...
                   'gui_LayoutFcn',  [] , ...
                   'gui_Callback',   []);
if nargin && ischar(varargin{1})
    gui_State.gui_Callback = str2func(varargin{1});
end

if nargout
    [varargout{1:nargout}] = gui_mainfcn(gui_State, varargin{:});
else
    gui_mainfcn(gui_State, varargin{:});
end
% End initialization code - DO NOT EDIT


% --- Executes just before HOME is made visible.
function HOME_OpeningFcn(hObject, eventdata, handles, varargin)
% This function has no output args, see OutputFcn.
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
% varargin   command line arguments to HOME (see VARARGIN)

% Choose default command line output for HOME
handles.output = hObject;

% Update handles structure
guidata(hObject, handles);
i = imread('hospital1.jpg');
axes(handles.axes1);
imshow(i);

% UIWAIT makes HOME wait for user response (see UIRESUME)
% uiwait(handles.figure1);


% --- Outputs from this function are returned to the command line.
```

```
function varargout = HOME_OutputFcn(hObject, eventdata, handles)
% varargout  cell array for returning output args (see VARARGOUT);
% hObject    handle to figure
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

% Get default command line output from handles structure
varargout{1} = handles.output;


% --- Executes on button press in pushbutton1.
function pushbutton1_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)

%try

h = waitbar(0,'Recording your pass voice...');
steps = 4;

%for step = 1:steps
   % computations take place here

% Record your voice for 5 seconds.
recObj = audiorecorder;
waitbar(0.01);
disp('Start speaking.')
waitbar(0.2);
recordblocking(recObj,2);
waitbar(0.4);
disp('End of Recording.');
waitbar(0.4);
% Play back the recording.
%play(recObj);
waitbar(0.8);
% Store data in double-precision array.
global myRecording  as
myRecording = getaudiodata(recObj);
%disp(myRecording);
[m,n] = size(myRecording);
disp(m);
waitbar(0.9)
close(h)
% v = MFCCProcessor(myRecording, 8000,12)
% %%
% [ceps,freqresp,fb,fbrecon,freqrecon] = ...
% mfcc(myRecording,8000,100);
% rastaout = rasta(ceps,100);
% mfccDCTMatrix = 1/sqrt(40/2)*cos((0:(13-1))' * ...
```

```
% (2*(0:(40-1))+1) * pi/2/40);
% mfccDCTMatrix(1,:) = mfccDCTMatrix(1,:)*sqrt(2)/2;
% rastarecon = 0*fbrecon;
% for i=1:size(rastaout,2)
%  rastarecon(:,i) = mfccDCTMatrix' * ...
% rastaout(:,i);
% end

% s = vqlbg(rastaout,1);
%global voiceprint
%voiceprint =s;


% ff = []
% for fn = 1:5
%     ff(fn)=fn
%
% end
%disp(s)

%[m,n] = size(s)
%figure
%plot(myRecording);
%end

fs = 16000;
       windowLength = round(0.03*fs);
       overlapLength = round(0.025*fs);
       win = hann(1024,"periodic");
       S = stft(myRecording,"Window",win,"OverlapLength",512,"Centered",false);
       melC = mfcc(S,fs);
       as = reshape(melC,[],1)
      %                                    f0                                    =
pitch(myRecording,fs,'WindowLength',windowLength,'OverlapLength',overlapLengt
h)
      %feat = [melC,f0]
%        size(melC)
%        size(f0)
       %voicedSpeech                                                             =
isVoicedSpeech(myRecording,fs,windowLength,overlapLength);

       %feat(~voicedSpeech,:) = [];
       %label = repelem(dsInfo.Label,size(feat,1));
       M = mean(melC,1);
       St = std(melC,[],1);
       nfeatures = (melC-M)./St
%        label = repelem("me",size(features,1))
    %   helperExtractAuditoryFeatures(myRecording,fs)
%try
 load ('voice.mat');
```

```
 sds = exist('voicecode');
 if sds < 1
    disp('No user registered yet');
 end
 i = size(voicecode,1);
 features = [];
 features2 =[];
labels = [];
labels2 = [];
 for fn = 1:i
    ff= voicecode{fn,:};
   ass = reshape(ff,[],1)
    label =fn;
    M = mean(ff,1);
    St = std(ff,[],1);
    feat = (ff-M)./St;
    label = repelem(label,size(ff,1))
    features = [features;feat];
    features2 = [features2;ass'];
    labels = [labels,label];
    labels2 = [labels2,fn];
 end
 features2
 labels2
 trainedClassifier = fitcknn( ...
   features2, ...
    labels2, ...
   'Distance','euclidean', ...
   'NumNeighbors',1, ...
   'DistanceWeight','squaredinverse', ...
   'Standardize',false, ...
   'ClassNames',unique(labels));
%   trainedClassifier = fitcknn( ...
%    features2, ...
%     labels2,'OptimizeHyperparameters','auto',...
%    'HyperparameterOptimizationOptions',...
%    struct('AcquisitionFunctionName','expected-improvement-plus'))
%
%  k = 2;
%                     group = labels2;
%                             c = cvpartition(group,'KFold',k); % 5-fold stratified cross
validation
%                     partitionedModel = crossval(trainedClassifier,'CVPartition',c);
%                                                         validationAccuracy = 1 -
kfoldLoss(partitionedModel,'LossFun','ClassifError');
%                                         fprintf('\nValidation accuracy = %.2f%%\n',
validationAccuracy*100);
%
%                     validationPredictions = kfoldPredict(partitionedModel);
%                     figure
```

```
%                                                                    cm    =
confusionchart(labels2,validationPredictions,'title','Validation Accuracy')
%                       cm.ColumnSummary = 'column-normalized';
%                       cm.RowSummary = 'row-normalized';

prediction = predict(trainedClassifier,as')
%prediction = categorical(string(prediction));
prediction = int16(prediction)

loginid =voicecode{prediction,2}
  save('login.mat','loginid');
Hosipital;
%  trainedClassifier = fitcknn( ...
%     features, ...
%      labels,'OptimizeHyperparameters','auto',...
%     'HyperparameterOptimizationOptions',...
%     struct('AcquisitionFunctionName','expected-improvement-plus'))

                        % k = 5;
                        % group = labels2;
                        % c = cvpartition(group,'KFold',k); % 5-fold stratified cross
validation
                        % partitionedModel = crossval(trainedClassifier,'CVPartition',c);
                        %                 validationAccuracy                 =                 1
-
kfoldLoss(partitionedModel,'LossFun','ClassifError');
                        %         fprintf('\nValidation         accuracy         =         %.2f%%\n',
validationAccuracy*100);
                        %
                        % validationPredictions = kfoldPredict(partitionedModel);
                        % figure
                        % cm = confusionchart(labels,validationPredictions,'title','Validation
Accuracy')
                        % cm.ColumnSummary = 'column-normalized';
                        % cm.RowSummary = 'row-normalized';

% prediction = predict(trainedClassifier,nfeatures)
% prediction = categorical(string(prediction));
% %prediction = int16(prediction)

% figure('Units','normalized','Position',[0.4 0.4 0.4 0.4])
% cm = confusionchart(int16(labels),prediction,'title','Test Accuracy (Per Frame)');
% cm.ColumnSummary = 'column-normalized';
% cm.RowSummary = 'row-normalized';

% figure('Units','normalized','Position',[0.4 0.4 0.4 0.4])
% cm = confusionchart(labels,prediction,'title','Test Accuracy (Per Frame)');
% cm.ColumnSummary = 'column-normalized';
% cm.RowSummary = 'row-normalized';
%
% r2 = prediction(1:numel(adsTest.Files));
```

```
% idx = 1;
% for ii = 1:numel(adsTest.Files)
%     r2(ii) = mode(prediction(idx:idx+numVectorsPerFile(ii)-1));
%     idx = idx + numVectorsPerFile(ii);
% end
%
% figure('Units','normalized','Position',[0.4 0.4 0.4 0.4])
% cm = confusionchart(adsTest.Labels,r2,'title','Test Accuracy (Per File)');
% cm.ColumnSummary = 'column-normalized';
% cm.RowSummary = 'row-normalized';

% % catch vv
% %     msgbox(vv);
% % end
% i = size(voicecode,1);
% ff = [];
% ter =[];
%
%
% for fn = 1:i
%     fn;
%     ff(:,1)= voicecode{fn,:};
%     %[Dist,D,k,w]=dtw(as,ff,0);
%    ter(fn) = num2str(fn);
%     %ff(fn)=Dist;
% end
%xc= mindist_classifier_type_final(as,ff,0.2)
% k=1;
% [dist, iidx]=pdist2(ff,as,'mahalanobis','smallest',k)
% matching_class=ceil(iidx/fn)
% dg = sqrt(dist)
%
% eas=CreateAndTrainANN(ff,as)
% kts = gmdistribution(ff,1)
%  d2 = mahal(ff,as)
%  dg2 = sqrt(d2)
%  [m Ind] = min(d2)
%  Mdl = fitcknn(ff,jj,'NumNeighbors',1)
%  flwr = mean(as);
% flwrClass = predict(Mdl,as)
% for fn = 1:i
%     fn;
%     dff= voicecode{fn,1};
%     [Dist,D,k,w]=dtw(as,dff,0);
%     ff= gmdistribution(dff,1)
%   % d2 = mahal(ff,as)
%     %ff(fn)=Dist;
% end
```

```
% ds=[0;1;2;3;4;5;6;];
%  d2 = mahal(ff,as)
%  [m Ind] = min(d2)

 %cAlpha(Ind)

% [c,dd] = min(ff)
%
%         if c <= 0.4    % u can increase the value 0.4 if u want to increase vunerability
or decrease to 0.2 0r 0.3 increase accuracy
%             got = voicecode{dd,2}
% %             dg = cd;
% %             str = strcat(cd,'\',got);
% %             cd(str);
% %             %from
% %             data = load ('info.mat');
% %             ima = imread(strcat(got,'.jpg'));
% %
% %
% %             cd(dg);
% %             %to
% %
% %              save('login.mat','data');
% %              imwrite(ima,'login.jpg');
% %                Hosipital;
% %
%         else
%          msgbox('Please, try to login again with the specific voice word you used or in
a less noisy environment');
%         end

% targetD=categorical([1;2;3;4;5;6;7])
%
targetD=categorical([0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;0;1;1;1;1;1;1;1;
1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1;1]);
% for fn = 1:i
%     fn;
%     ff(:,fn)= gmdistribution(voicecode{fn,1},1,);
%     %[Dist,D,k,w]=dtw(as,ff,0);
%    ter(fn) = num2str(fn);
%     %ff(fn)=Dist;
% end
%ff
% [coeff,scoreTrain,~,~,explained,mu]=pca(ff)
% idx = find(cumsum(explained)>95,1)
% scoreTrain95 = scoreTrain(:,1:idx);
% mdl = fitctree(scoreTrain95,as)
% %scoreTest95 = (XTest-mu)*coeff(:,1:idx);
% YTest_predicted = predict(mdl,as)
%[predict1] = classifyANNoffline(ff', ter, as)
```

```
%[predict1] =classifyLDAoffline(ff, ter, as)
%Mdl = fitcknn(ff,as)
% [label,score] = predict(Mdl,as)
%targetD=categorical(ter)
%[c,dd] = min(ff)
%         if c <= 0.4    % u can increase the value 0.4 if u want to increase vunerability
or decrease to 0.2 0r 0.3 increase accuracy
%            got = voicecode{dd,2}
%           % dg = cd;
% %             str = strcat(cd,'\',got);
% %             cd(str);
% %             %from
% %             data = load ('info.mat');
% %             ima = imread(strcat(got,'.jpg'));
% %
% %
% %             cd(dg);
% %             %to
% %
% %              save('login.mat','data');
% %              imwrite(ima,'login.jpg');
% %                Hosipital;
%
%        else
%         msgbox('Please, try to login again with the specific voice word you used or in
a less noisy environment');
%        end
% catch exception
%    msgbox(exception)
% end

% %% Define Network Architecture
% % Define the convolutional neural network architecture.
% layers = [
%    imageInputLayer([52 1 1]) % 22X1X1 refers to number of features per sample
%    convolution2dLayer(3,16,'Padding','same')
%    reluLayer
%    fullyConnectedLayer(384) % 384 refers to number of neurons in next FC hidden
layer
%    fullyConnectedLayer(384) % 384 refers to number of neurons in next FC hidden
layer
%      fullyConnectedLayer(2) % 2 refers to number of neurons in next output layer
(number of output classes)
%    softmaxLayer
%    classificationLayer];
%
% options = trainingOptions('sgdm',...
%    'MaxEpochs',500, ...
%    'Verbose',false,...
%    'Plots','training-progress');
```

```
%net = trainNetwork(ff,targetD,layers,options);

%predictedLabels = classify(net,trainD)'



% --- Executes on button press in pushbutton2.
function pushbutton2_Callback(hObject, eventdata, handles)
% hObject    handle to pushbutton2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
try
delete(HOME);
register;
catch dd
end

% --------------------------------------------------------------------
function Untitled_1_Callback(hObject, eventdata, handles)
% hObject    handle to Untitled_1 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)


% --------------------------------------------------------------------
function Untitled_2_Callback(hObject, eventdata, handles)
% hObject    handle to Untitled_2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)


% --------------------------------------------------------------------
function Untitled_3_Callback(hObject, eventdata, handles)
% hObject    handle to Untitled_3 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
try
delete(HOME);
Main;
catch
end


% --------------------------------------------------------------------
function Untitled_4_Callback(hObject, eventdata, handles)
% hObject    handle to Untitled_4 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
try
```

```
h = waitbar(0,'Recording your pass voice...');
steps = 4;

%for step = 1:steps
   % computations take place here

% Record your voice for 5 seconds.
recObj = audiorecorder;
waitbar(0.01);
disp('Start speaking.')
waitbar(0.2);
recordblocking(recObj,2);
waitbar(0.4);
disp('End of Recording.');
waitbar(0.4);
% Play back the recording.
%play(recObj);
waitbar(0.8);
% Store data in double-precision array.
global myRecording
myRecording = getaudiodata(recObj);
disp(myRecording);
[m,n] = size(myRecording);
disp(m);
waitbar(0.9)
close(h)
%%
[ceps,freqresp,fb,fbrecon,freqrecon] = ...
mfcc(myRecording,8000,100);
rastaout = rasta(ceps,100);
mfccDCTMatrix = 1/sqrt(40/2)*cos((0:(13-1))' * ...
(2*(0:(40-1))+1) * pi/2/40);
mfccDCTMatrix(1,:) = mfccDCTMatrix(1,:)*sqrt(2)/2;
rastarecon = 0*fbrecon;
for i=1:size(rastaout,2)
 rastarecon(:,i) = mfccDCTMatrix' * ...
rastaout(:,i);
end

s = vqlbg(rastaout,1);
%global voiceprint
%voiceprint =s;


% ff = []
% for fn = 1:5
%     ff(fn)=fn
%
% end
```

```
%disp(s)

%[m,n] = size(s)
figure
plot(myRecording);
%end
 load ('voice.mat');
i = size(voicecode,1);
ff = [];
for fn = 1:i
    fn;
    dff= voicecode{fn,1};
    [Dist,D,k,w]=dtw(s,dff,1);
    ff(fn)=Dist;
end
[c,dd] = min(ff);
       if c <= 0.3    % u can increase the value 0.4 if u want to increase vunerability or
decrease to 0.2 0r 0.3 increase accuracy
           got = voicecode{dd,2};
           dg = cd;
           str = strcat(cd,'\',got);
           cd(str);
           %from
           data = load ('info.mat');
           ima = imread(strcat(got,'.jpg'));


           cd(dg);
           %to

            save('login.mat','data');
            imwrite(ima,'login.jpg');
              Hosipital;

       else
       msgbox('Please, try to login again with the specific voice word you used or in a
less noisy environment');
       end
catch exception
  % msgbox(exception)
end



% --- Executes on button press in pushbutton2.
% --- Executes on button press in checkbox2.
function checkbox2_Callback(hObject, eventdata, handles)
% hObject    handle to checkbox2 (see GCBO)
% eventdata  reserved - to be defined in a future version of MATLAB
% handles    structure with handles and user data (see GUIDATA)
```

```
% Hint: get(hObject,'Value') returns toggle state of checkbox2
% Load the feature matrix


load('features.mat');

% Create the target vector
targets = [ones(10,1); 2*ones(10,1); 3*ones(10,1)];

% Set up the neural network
hiddenLayerSize = 10;
net = patternnet(hiddenLayerSize);

% Set up the training parameters
net.trainParam.showWindow = false;
net.trainParam.showCommandLine = true;
net.trainParam.epochs = 100;

% Train the neural network
[net,tr] = train(net,X',targets');

% Test the neural network
outputs = net(X');
classes = vec2ind(outputs);

% Display the confusion matrix
confusionmat(targets,classes)

% Load the new recording
[y,fs] = audioread('new_recording.wav');

% Extract the features using MFCCs
mfccs = mfcc(y,fs);

% Reshape the features into a row vector
x = reshape(mfccs',1,[]);

% Classify the new recording using the neural network
output = net(x');
class = vec2ind(output);

% Display the classification result
if class == 1
    disp('The new recording is in dialect A');
elseif class == 2
    disp('The new recording is in dialect B');
else
    disp('The new recording is in dialect C');
end
```

Bio-data

1. **Personal Data:**

Full name:	Taiwo Mauyon Kuponu
Address:	St. Anne's Anglican Church, Vicarage, Molete, Ibadan.
E-mail:	taiwokuponu59@gmail.com
Phone No:	08103765114
Date of Birth:	12th February, 1995
Nationality:	Nigerian

2. **Educational Background:**

Educational Institutions Attended with Dates and Qualifications:

2012-2016: B.SC Computer Science, Caleb University, Imota, Ikorodu.

2012: (WAEC) ST. Alphonsus De Liguori Private Secondary School, Aboru, Iyana Ipaja.

2006-2011: Babington Macaulay Junior Seminary (BMJS)Secondary School, Agunfoye-Lugbusi Village, Ikorodu.

2004-2005: Mate Nursery & Primary School, Alapere, Ketu, Lagos.

3. Working Experience with Dates.

January 2022 – Till Date -Network Marketing Professional, Secure Your Future Group, Bodija, Ibadan

June 2020 – Quality Assurance Analyst (QA), Netow Solutions, Lagos.

2018-2019 – Customer Service Representative (Proctor & Gamble(P&G), British American Tobacco (BAT), Consol Limited, Oshodi, Lagos

August 2016 - 2017 – NYSC. Baptist High School (Teacher), Osogbo, Osun State.

2014-2015 (SIWES – IT) WAEC examiner assistant, JAMB Marker and Invigilator for TOEFL exams, Receptionist. Data Science Nigeria Limited (DSNL)

4. Conferences Attended and Publications

Crime Reduction System through Bryne Criminal Justice Innovation, INTERNATIONAL JOURNAL OF CREATIVE THOUGHTS, ISSN:2320 – 2882(August,2022)

The University Compliance Certification

This is to certify that this thesis written by Taiwo Mauyon Kuponu with the matriculation number LCU/PG/002632 in the **Department of Computer Science**, **Faculty of Natural and Applied Sciences**, Lead City University, Ibadan, Oyo State, Nigeria is in full compliance with the approved University Format and Style.

_____ _____

Signature **Date**