**NATURAL LANGUAGE TEXT PROCESSING USING MACHINE LEARNING TECHNIQUES**

**BY**

**SURNAME, OTHER**

**00/000ABCD/001**

**ABSTRACT**

The rapidly developing field of machine-to-man, or M2M, communication of machines communicating with humans in a more natural and straightforward way changes how we engage with technology. Ranging from the management of smart homes and gadgets with the help of voice assistants customer services using Chabots and voice auto, more complex and seamless M2M interactions is developing as more increase is anticipated. The main objective of this research is to improve the natural language processing (NLP) of a trained data set of texts and speech synthesis using an intersection of machine learning algorithms by applying novel techniques to improve the accuracy of NLP systems as regards text pronunciation errors of lambdacism resulting from the inters witch of the of the /L/ and /R/ phonemes by the Igbo native English speaker. This approach applies the pre-processing of text with Feature extraction using the bag of words model and Mel-Coefficient Frequency Cepstral Coefficients MFCC, Tokenization, parts of speech (POS) tagging, Text Extraction using Named Entity Recognition (NER), Text and audio normalization for resolving lexical, semantics and syntactic ambiguities, word and character detection, audio conversions using the Fast Fourier Transform (FFT) models. Speech Recognition via Automatic Speech Recognition (ASR) is performed on the text after which sorting and replacement is applied using NLP process by the WordPunctTokenizer.

**CHAPTER ONE**

**INTRODUCTION**

**1.1 Background of Study**

Sound, words, and sentences are the fundamental components of speech, for humans to communicate effectively among themselves, conversations and dialogues should be expressed using methods including languages comprehensible to all parties involved in the communication process (Guzman & Lewis, 2020). This is to ensure that information passed is not misinterpreted or misunderstood in any form due to mistakes that could result from the wrong pronunciation of phonemes.

The study offers a unique approach to processing natural language text using machine learning methods since communication between humans and machines requires an interface, therefore, human-to-machine interaction with dynamic technical systems such as industrial plants must be accurate due to its significance in guaranteeing efficiency, quality, and safety of all factors involved. The solution this study provides focuses primarily on the problem of textual lambdacism faults, which are found and fixed using a combination of error detection, conversion, and replacement methods (Leng et al., 2022). These errors can bring about irredeemable damage to such systems if not prevented from occurring.

A speech condition known as lambdacism is characterized by excessive use of the "/L/" sound. When producing other consonant sounds (typically the /R/) accurately, people with an accent lallation of lambdacism end up replacing them with the "/L/" sound. As a result, occurrences of lambdacism in speech will lead to communication and social as well as linguistic difficulties (Grigorova et al., 2020).

For people who are natives and speakers of Igbo, lambdacism is a major challenge as regards spoken words of the English Language, this is due to the difficulty that results from the mother tongue’s accent influencing and overshadowing the correct pronunciation of certain English words (Nwigwe & Izuagba, 2017).

The study focuses on how ML algorithms can be used to enhance the pronunciation of these natural language texts using Artificial Intelligence (AI) for grapheme-to-phoneme (G2P) and Phoneme to Phoneme (P2P) in Text to Speech Synthesis (T2S), then Speech-to-Speech (S2S) in Automated Speech Recognition (ASR) modules. A combination of these modules will serve as a standard basis for the design of an improvement in human-computer interaction (HCI). (Guzman, 2018) describes the HCI as "the creation of meaning among humans and machines", the idea to cease communication from being a human-only process brought about the technology to develop models, algorithms, and end-to-end systems and is described as a sum up of the numerous methods that people use to communicate with different technology.

Other authors have indeed worked on Machine Learning (ML) and Natural Language Processing (NLP) using several algorithms and models to achieve various possible results some of which are discussed below. However, the study seeks to focus majorly on the error that occurs when an Igbo native English speaker interchanges L and R while pronouncing certain words due to accentuation differences using a concatenating algorithm of the Hough transform (Ploß et al., 2012), vertical and horizontal profile projection and cross-correlation measures of the trained input data text. The study applies the Linear Support Vector Machine (LSVM) algorithm for text classification. This classification is made based on past observations by using the pre-labeled examples of trained data and entails leaving the different associations between pieces of text that is to say that for a particular output, a particular input data needs to be trained. This trained data is called a “tag”.

**1.2 Statement of the Problem**

English as the universal language for the world community is known to come with a level of difficulty for the Igbo people majorly found in the Eastern region of Nigeria, Africa as they are popularly known to have an accent that enables the conflicting shift in pronunciation of the /L/ and /R/ alphabets in certain English words as these two letters are usually interchanged, this is typically known as lambdacism (Oyeka, 2017).

The inability of a machine to recognize, interpret, and replace words of lambdacism instances has resulted in multiple errors thus reducing the quality of word processing in text and speech translation. Seeing that data originally in human language is highly unstructured thereby making it an ambiguous task for machines to understand and interpret, there is a need for a systemic end-to-end design of a comprehensive algorithm that would proffer a long-lasting solution to this issue thereby improving clarity and elimination of accent in the pronunciation of certain words (Alkatheiri, 2022). The study applies a hybrid approach that involves implementing a combination of various machine learning models and algorithms to analyze the method of conversion of the natural language text data into comprehensible sounds.

**1.3 Aim and Objectives of Study**

This research aims to use machine learning to improve the accuracy of pronunciation of some English language phonemes affected by lambdacism.

The objectives are to:

1. Identify the features of cluttering of speech caused by mis-articulation of the L and R lambdacism using Part of speech (POS) Tagging and dictionary-based Named Entity Recognition (NER).
2. Solve the ambiguities resulting from lallation in the lexical, syntactic, and semantic levels of the natural language text using machine learning algorithm Stemming and Lemmatization techniques.
3. Convert and replace words affected by lambdacism using Fast Fourier Transform (FFT) and Digital Signal Processing (DSP) involving a spectrogram.
4. Convert spoken words to text accurately using ASR technique and MFCC for Audio Feature Extraction.
5. Evaluate the results and compare with existing results to determine accuracy using word error reduction (WER).

**1.4 Research Questions**

The research questions used in this study are:

1. What are the features of cluttering of speech caused by mis-articulation of the L and R lambdacism mispronunciation?
2. How can a machine learning algorithm be used to resolve ambiguities in lexical, syntactic, and semantic levels of the natural language text?
3. How can machine learning be used to convert and replace spoken words affected by lambdacism?
4. To use machine learning techniques to convert spoken words to text accurately.

**1.5 Significance of the Study**

This study is advantageous to the communication industry since it is advantageous for basic error identification and correction coding methods which is necessary for optimizing cyclic redundancy checks (Kumari et al., 2022). Some programs allow you to review and correct each page in turn and they instantly process the entire page then use a built-in spellchecker to highlight any misspelled words that may indicate misrecognition, so mistakes can be automatically corrected. An example is when Griol & Molina, (2016) designed a framework that estimated the probability of the existence of errors using a combination of user behaviour and error modeling in the user utterance of specific speech.

The study will help Audio forensic experts with their research as it aids in the proofreading process to removetext paragraphs with harmful materials and images can be identified where the hidden text is retrieved and categorized using machine learning methods. A typical example of where this method is applied is in spam detection, cancer classification, Neural Machine Translation (NMT), and license plate detection, (Imam et al., 2022).

For the braille reading in educational development centres, the study is used to enhance effortless comprehension of accents in word pronunciation, errors that occur as a result of the difficulty in pronunciation of certain vowels and consonants due to dialectal or accentual speech disparities would be improved on. To identify poorly uttered speech sounds, it is first necessary to extract the spoken language from other elements in the signal.

**1.6 Scope of Study**

The proposed research will focus on how man-to-machine-to-man communication in the English Language can be improved by employing some machine learning techniques namely Support Vector Machine and Linear Kernelling algorithms on some trained data sets.

**1.7 Limitations of the Study**

The research study is limited to how the computer system can recognize, interpret, and convert some specifically trained data set of words from the English language text that is made up of lambdacism (/L/ and /R/) interchange, converting some specific characters of these words from texts into sounds and back to text.

**1.8 Expected Research Contributions**

This research will contribute to the improvement of language expressions in Consumer Robotic Interactive Languages through the proper articulation of speech by machines. This improves the quality of digital communication in the telecommunication industry and the world at large.

**1.9 Motivation for the Research**

With an increase in the use of automated devices globally, there is a rise in the popularity of speech recognition systems such as Chat GPT (Liu et al., 2023), Alexa of Amazon, and Siri from the iPhone thereby making interactions between machines and humans the next big thing. These systems are therefore applicable in the home front, for robotics, improvement of reading and use of social media for the visually and hearing impaired, enhancement of transcription, ease the control of automated devices, etc. This brings more value by fine-tuning the NLP process for machine recognizing and reading texts effectively as well as providing ease for data analysis and input.

**1.10 Dissertation Organisation**

The report starts with the first chapter discussing an introduction and a background to the proposed study, a statement of the problem, research objectives, motivation, scope, limitations, significance of the study, expected research contributions, and definition of terms. The second chapter gives a brief literature review on the history of natural language processing, machine learning, speech synthesis, lambdacism, a review of related studies, and a research gap. The third chapter explains the various steps in which the study proposes to achieve the design of an algorithm that enhances natural language processing using machine learning techniques.

**1.11 Definition of Terms**

**Advanced Research Projects Agency (ARPAbet):** a set of phonetic transcription codes that represent phonemes and allophones of general American English developed as part of a speech understanding research project by the Advanced Research Projects Agency in the 1970s (Nakatsuka et al., 2020).

**Automated Speech Recognition (ASR):** The technology that enables humans to use their voices to speak with a computer interface in a way that, in its most advanced variants, resembles typical human conversation (Shadiev & Liu, 2023).

**L2-ARCTIC:** A speech corpus of non-native English (Zhao et al., 2018).

**Lallation:** A defect of speech consisting of the pronunciation of (r) as (l) (Lallation, 2019).

**Lambdacism:** A speech defect in which l is pronounced instead of r (Lambdacism, 2010).

**Linguistics:** The scientific study and analysis of language and its communication tool

(Shaykhislamov & Makhmudov, 2020).

**Machine Translation (MT):** The automatic translation of texts from one natural language to another while maintaining their meaning and creating fluent content in the target language using computers (Haifeng et al., 2022).

**Morpheme:** The area of linguistics that studies words, their internal structure, and how they are generated (Aronoff et al., 2022).

**Morphology**: The study of morpheme (Kolanchery, 2015).

**Natural language Generation (NLG):** The process of creating linguistic output from underlying non-linguistic input and is sometimes referred to as the "poor sister" of work in natural language understanding (NLU) (Dale, 2020).

**Natural Language Processing (NLP):** This is most times used for the interpretation of raw (handwritten) and structure-less texts to make them more interpretable and understandable to the communicators involved (Chowdhary, 2020).

**Neural Machine Translation (NMT):** The use of end-to-end systems on neural network models to learn a statistical model for machine translation (Stahlberg, 2020).

**Phonemes:** The smallest unit of speech that distinguishes one word from the other. For example, the English language letters b and p (in the words tab and tap). In this research, we will be making use of phonemes (Davenport & Hannahs, 2013).

**Phonology**: The study of sound patterns within languages, their meanings, and the way they come together to form speech and words. This science helps to deal with patterns present in the sound and speeches related to the sound as a physical entity (Davenport & Hannahs, 2013).

**Pragmatics:** A subfield of linguistics that describes the unspoken intentions of a writer or speaker in connection to the joint creation of a linguistic form. This science studies the different uses of language and how context contributes to significance (Razzakberdiyevna, 2023).

**Syntax**: The arranging of words and phrases to produce complete sentences in a language (Mariani, Mu’in, & Al Arief, 2019, p. 10).

**Semantics**: The area of linguistics and logic focused on the meaning of words and how they are employed in sentences. This research will apply the science involved with the literal meaning of the words, phrases and sentences (Ataboyev & Turgunova, 2022).

**CHAPTER TWO**

**LITERATURE REVIEW**

**2.0 Introduction**

This chapter seeks to provide the reader with an overview of what natural language processing, speech recognition, machine learning, and lambdacism are all about. The second part of this chapter gives a brief description of current and past works, existing models integrating NLP and ML, and the research gap.

**2.1 History of Natural Language Processing (NLP)**

Natural language processing can be traced back to the Second World War when the need for Machine Translation (MT) came into existence. The whole idea at that time was to translate from one human language to another using the brain of the computer for example English to Russia and back. The first patent for a translation machine was in the mid-1930s by Georges Artsrouni (Jakubková & Warner, 2023), which involved the use of a bilingual dictionary in mapping the words of one language directly to another on paper tape, its limitation was not being able to handle grammatical aspects of languages.

The second patent of Machine Translation was in the 1950s by Peter Troyanskii a Russian; he gave detailed strategies that tackled the grammar of a given language. He achieved this by using two disruptive approaches that is the use of a bilingual dictionary alongside another method to deal with the grammar of the language established over the international language known as Esperanto.

Johri et al. (2021) wrote that the first attempt at using NLP was by the Germans in World War II, a machine called Enigma was built and it was used for the secret message of Germans into the secret code and was also used to transfer the message to the field commanders and military units of Germans placed in Europe. It was instrumental in arranging the attacks thereby meeting the requirements of the army. The enigma is one of the great achievements of Germans enhancing the ability to communicate secretly even in the most precarious situations. Later in 1946, Britain came up with, Colossus, a machine capable of successfully decrypting the secret codes generated by Tunny, a code name given to Enigma by the British.

Natural Language Processing (NLP) is the branch of computer sciences that enables machines to understand, decode, and manipulate human languages just as humans do (Chowdhary, 2020). The study also referred to Chomsky’s 1959 book on syntactic structures expressed in NLP as bringing about a fundamental reorientation in language speaking whereby large annotated corpora of text were used to offer gold standards for the evaluation of speech as machine learning algorithms were being trained. The evaluation method grew more exacting as deep analysis was substituted with simple reliable approximations which resulted in the birth of statistical natural language parsing. Natural language processing includes many different techniques for interpreting human language, ranging from statistical and machine-learning methods to rules-based and algorithmic approaches. Natural language text could be processed orally or written provided that these texts are understandable and used by humans for interaction and communication.

Thus Natural Language Processing is defined by Liddy (2001) as a theoretically motivated range of computational techniques for analysing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications. There exist seven levels of NLP which include: phonology, morphology, lexicon, syntactic, semantic, speech, and pragmatic (Ribeiro, 2021).

The ability of the computer to understand and compose the natural language humans speak and write amidst pronunciation, spelling, accent, dialect, word skip errors and every other possible word ambiguity is what processing is all about and is also referred to Text Analysis. This aspect of computer science is also referred to as Computational Linguistics basically because it deals with the structures and compositions of languages as regards to how the computer relates with them, the language this research seeks to work with is the English Language with special scrutiny to gliding vowels (Tsujii, 2021). There are three (3) Natural Language Processing approaches to text processing which are: Rule-Based NLP, “Traditional” - Machine Learning NLP & Deep Learning/Neural Network NLP. This research focuses on the Machine Learning (ML) Approach for English Language text analysis.

Intelligent Character Recognition (ICR) is an advanced technology of that allows for the digital capturing of handwritten texts and translating them into machine-readable ASCI codes (Ptucha et al., 2019). For this to be achieved, lines of text are first extracted and processed at once, image data could consist of at least alpha characters (a, b, c…), numeric characters (0, 1, 2...) special characters ($, %, &...), or a “reject” also known as void - meaning that the algorithm was not able to identify the image as a particular character.

Automatic Speech Recognition (ASR) is a machine learning technology that identifies, decodes, and converts spoken words to written texts popularly called Speech-To-Text technology (Panda, 2017). It involves the use of modern computer techniques (both hardware and software) to identify, sort, and process the human voice, it is also used to authenticate the identity of an already programmed person speaking into the system. Some ASR systems are either speaker-dependent or speaker-independent, speaker-dependent systems are trained to recognize particular words and speech patterns while speaker-independent systems do not require any training whatsoever and can recognize spoken words regardless of speaker. The study applies speaker-dependent systems to train certain words to be used for input data as it refines its performance by fine-tuning shift parameters and the size of the spectrum using the Fourier Transform (Marini et al., 2021).

Major components of the ASR systems include the lexicon, acoustic models and then the language model. The lexical design is the building block of acoustic models for every vocal input and is paramount to the accuracy of the speech recognizer, this is due to the instance that some words such as diagraphs, homophones, and homonyms need to be pronounced properly. One of the most widely used sets of lexicons is the Advanced Research Projects Agency (ARPAbet) which represents phonemes of the English language. The acoustic model separates and analyses each audio signal into small frames and then predicts which sound is spoken into each frame, this is achieved through the use of a deep learning algorithm (Bhatt & Jain, 2020). This process establishes statistical representation for the feature vector sequences for a particular sound unit so that a classifier for the entire sound unit used in the ASR system can be designed.

An analysis of a machine learning-based algorithm for text classification using different machine techniques for natural language text processing was carried out thus eliminating the need for manual data classification. Two different datasets were compared using the Support Vector Machine (SVM), K-nearest Neighbour (KNN), Logistics Regression (LR), Multinomial Naïve Bayes (MNB), and Random Forest (RF) with accuracy, precision, and recall as performance matrix (Sayar et al., 2022).

Balyan, McCarty, & McNamara (2017) carried out a study on the combination of the use of Machine Learning and Natural Language Processing to Assess Literary Text Comprehension with the use of Naive Bayes, Maximum Entropy, Neural Networks, Support Vector Machine, Bagging, Boosting, Stacking, Random Forests and Elaborative N-gram Algorithms. The study indicated that ensemble classification algorithms were, generally, more accurate than single classifiers.

Wang et al., (2017) designed an end-to-end generative text-to-speech model that synthesizes speech directly from characters called Tacotron from scratch with the given text, and audio pairs by random initialization. It took a character sequence as input and output the corresponding spectrogram using simple text normalization which was substantially faster than sample-level autoregressive methods and achieved a 3.82 5-scale mean opinion score on the United States English language thereby outperforming a production parametric system in terms of naturalness. It was however found to require an improvement on its audible artifacts.

Gupta (2019) focused on two Natural Language Processing tasks which are Relation Extraction and Topic Modelling. Relation Extraction is aimed at identifying semantic relationships between entities or nominal within a sentence or document, while Topic Modelling is aimed at understanding the thematic structures underlying a collection of documents. Logistic Regression; a supervised machine learning approach was employed for this task for a semantic relationship needed as a text-mining tool to build a structured knowledge base for the NLP that would be used to analyse the texts.

Another important line of work is a single feed-forward neural network, trained end-to-end in a single stage which produced waveforms given character or phoneme sequences and learned to align without additional supervision from auxiliary sources or teacher forcing (Donahue et al., 2021). Speech is synthesized from normalized text or phonemes in an end-to-end manner, resulting in models that operate directly on character or phoneme input sequences and produce raw speech audio outputs. This implies that the training process is considerably fine-tuned end-to-end in the same fashion, but requires a pre-training stage with vocoder features used for intermediate supervision.

According to Canuma (2019), Alan Turing in 1950 developed the Turing test which is the test of a machine’s ability to exhibit intelligent behaviour equivalent to, or indistinguishable from, that of a human.

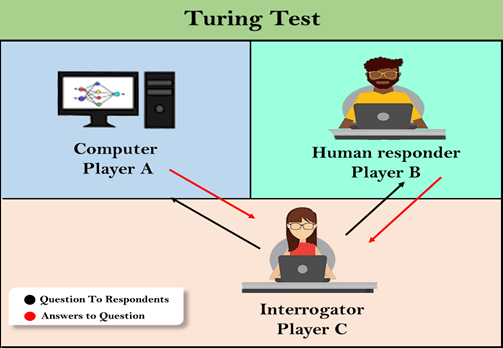


Figure 2.1: How Turing Test Worked. (Javapoint, 2021)

Any language can be understood as a group of symbols or rules. The symbols are re-integrated and used to transmit and broadcast the information while the rules are applied to suppress the symbols. The natural language text is disintegrated into various levels by the machine such as phonology, morphology, lexicon, syntactic, semantics, speech, and pragmatics. Phonology - the study of sound patterns within languages, their meanings, and the way they come together to form speech and words. This science helps to deal with patterns present in the sound and speeches related to the sound as a physical entity, Phonemes - the smallest unit of speech that distinguishes one word from the other. For example, the English language letters l and r (in the words lap and rap). The study will focus on the use of phonemes, Pragmatics - this is the study of how context contributes to meaning, the ability to understand another speaker's intended meaning is called pragmatic competence (Schlechtweg, 2021). This science studies the different uses of language. Morphology - This science deals with the structure of words and the systematic relations between them. Syntax - the study of how words and morphemes combine to form larger units such as phrases and sentences, the study will apply the science involved in the structure of sentences. Semantics – includes the logical and lexical semantics that deal with the presupposition and analysis of words respectively. This research applies the science involved with the literal meaning of the words, phrases, and sentences (Bender, 2022).

A typical example of this in a study is the prosodic domain of phonological encoding for speech error detection by Beirne & Croot (2018) where spoken words were recorded, and speech errors were transcribed after training in a repetitive format and then analyzed using Poisson Regression.

A word2vec text pre-processing model which transforms texts into a row of numbers approach for metaphor comprehension was used for figurative language processing to successfully differentiate metaphors rated as good (> 1.5z) from metaphors rated as bad (< −1.5z; Cohen’s d = 0.72) (Harati et al., 2021) and was able to successfully classify good metaphors with an accuracy of 82.9%. However, it achieved a true negative rate below the chance at 36.3% and had a resultantly low kappa of 0.037. The model could not distinguish unselected random metaphors from those selected by humans as having metaphorical potential. In the K-nearest neighbor and logistic regression algorithm (supervised learning used to solve binary classification tasks), words with similar meanings but different spellings (homophones) had similar vector representations, the logistic function also called the sigmoid function was used to turn regression lines into decision boundaries for binary classification, (Shah et al., 2020)

A good NLP approach would strike a balance between theory and practice as depicted in Figure 2.2.

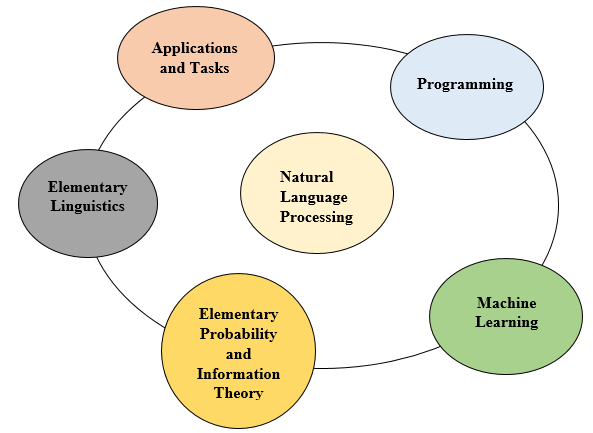


Figure 2.2: Phases of a standard NLP approach. (Kennington, 2021)

Arthur Samuel, an American IBMer and pioneer in the field of computer gaming and artificial intelligence coined Machine learning in 1959. He defined machine learning as a “Field of study that gives computers the ability to learn without being explicitly programmed” (Nazly et al., 2023)

**2.1.1 Speech Recognition**

Speech recognition is the digital process of the machine to convert speech into text (Abidin et al., 2022), speech in this form could be in the form of audio recording, voice recording, sounds from languages spoken, etc. It is categorized into two namely speaker-dependent and speaker-independent, speaker-dependent speech recognition software is developed to aid the improvement of digital dictations while the latter is developed to enhance telephone communications, although speaker-independent recognition systems do not require any training phase before being used, speaker-dependent recognition systems perform better by multiple factors so long as training is done thoroughly, the study will focus on speaker-dependent speech recognition, (Tandel et al., 2020). There are stages to achieving speech recognition which include Speech/Voice Capture for Audio Signals: speech captured using a microphone and recorder by reading selected words from data to be trained, reading should be done continuously irrespective of mistakes made or interference from external noise; Feature Extraction; Transcription and Transformation of Audio Frequencies done to rid the recorded words of noise, vibrations, and mistakes by trimming the recording to desired length capturing only the required data; Language Modelling, and Testing (Gaikwad et al., 2010).

**2.1.2 Machine Learning**

Machine Learning (ML) is a subset of artificial intelligence that involves programming a computer to optimize a performance criterion using sample data or past experiences (Alpaydin, 2020). Machine learning is categorized into three mainly supervised, unsupervised, and semi-supervised learning. Supervised machine learning according to Singh et al., (2016) is defined as the construction of algorithms that can produce general patterns and hypotheses by using externally supplied instances to predict the fate of future instances. Supervised machine learning classification algorithms aim at categorizing data from prior information. This means that labelled data sets are used to train algorithms in order to classify data and predict outcomes accurately. Supervised learning algorithms are trained in a way that supports functions which best describes the input data to be picked. This function states that for a given X, the best estimation of y (X -> y) is made (Fumo, 2017). Examples of supervised learning algorithms include; Nearest Neighbour, Naive Bayes, Decision Trees Linear Regression, Support Vector Machines (SVM), and Neural Networks.

On the other hand, unsupervised machine learning algorithms discover hidden patterns of data groups without the need of any human intervention or interference, this is achieved by the use of machine learning algorithms to analyse and cluster unlabelled datasets. Its ability to discover similarities and differences in information makes it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition. Unsupervised machine learning is where you only have input data (X) and no corresponding output variables (Brownlee, 2023). Examples of unsupervised learning algorithms are; clustering (k-means), Association Rules, etc. Semi-supervised learning is concerned with using labeled as well as unlabelled data to perform certain learning tasks (Engelen & Hoos, 2020). It is situated in between supervised and unsupervised learning and permits harnessing large amounts of unlabelled data available in many use cases in combination with typically smaller sets of labelled data. Semi-supervised machine learning seeks to understand how a combination of both labeled and unlabelled data changes the learning behavior of machines and then designs algorithms to harness the strengths of such combination. Examples of semi-supervised learning include Transductive SVM, Generative Model, Self-Training, Boosting, Ensemble learning, etc. (Mahesh, 2019). Machine learning is used for the classification and prediction of data, however in the proposed research work, machine learning would be used to classify the English language text and we will be using a combination of supervised and unsupervised learning techniques to achieve this via Intelligent Character Recognition, and Automatic Speech Recognition

Machine learning architecture deals with the various layers and stages of the machine learning cycle. This also includes the major steps raw data undergoes to be transformed to the trained data set, Machine learning is based upon the different algorithms that are used on the training data and is categorized into three types that is; Supervised Learning, Unsupervised learning, and Semi-supervised/Reinforcement Learning. The process involved in this architecture is; Data acquisition, Data Processing, Model Engineering, Execution, and Deployment (Washizaki et al., 2020).

**2.1.3 Speech Synthesis**

Speech synthesis is the generation of a human voice artificially through the use of a computer while the technology used to achieve this is called Text-To-Speech Synthesis.

The first mechanical synthesizer called the speech organ was produced in 1846, it could sing God Save the Queen followed by the first electrical synthesis device (buzzer and 2 resonant circuit for the first two formants) was developed in 1922. 1923 added the third format of the electrical synthesizer then 1938-1939, Homer Dudley manufactured the first electrical speech synthesizer called Bells Telephone Laboratory Voder, which was human-controlled by a keyboard and pedal. Although the quality of speech generated on it was limited, it however still demonstrated a synthesis of the human voice (Davis, 2014). Later on in 1961 the first phonemic-synthesis-by-rule program for digital computers was created, it was then followed almost immediately in 1968 by the first full text-to-speech system. The 1980s was the beginning of commercial Text-To-Speech as prosodic modification was carried out on existing systems between 1985 to the early 1990s, a speech synthesizer was used to translate written information into oral. Voice Conversion takes the speech of the source speaker as input and generates speech that sounds from a target speaker while maintaining the linguistic content, Zhang et al., (2021). Present-day machine-to-man communication has gone further due to the sporadic improvement that keeps taking place to include deterministic and stochastic models, large databases, automatically labeled databases, telephone applications such as helplines, call centers, and so on. These have been integrated into dialog systems, banks, and also navigation systems in cars, etc.

The three stages of the TTS include; Text to words, Words to Phonemes, and Phonemes to Voice by Woodford (2021).

1. Text to Words: this is the process by which ambiguity is reduced by narrowing down the many different ways a piece of text could be read most appropriately. This process is called Pre-processing and Text normalization. Pre-processing also tackles homographs (words spelled the same way but pronounced in different ways according to what they mean) for example the word "bow" can be pronounced either "/bow/" or "/bəʊ/".
2. Phoneme-to-phoneme (P2P): one fundamental phase of speech recognition and natural language processing is the P2P conversion as written by (Yang et al., 2006) deals with the generation of how a word is pronounced from its written form.
3. Words/Graphemes to Phonemes:A demonstration is the grapheme-to-phoneme (G2P) conversion for central Kurdish by Mahmudi & Veisi (2021) using optimality theory – a constraint-based theory that has an input-output structure. This G2P method is used to find the best pronunciation for written words. Here the speech synthesizer generates the sounds that make up the words that is to say that for each word, phonemes generated and stored are picked from the database. For example, the word cat is broken into three (3) - /c/, /a/, and /t/.
4. Phonemes to Voice: This is the process where the computer gets the phonemes together and then reads them out loud and can be achieved in three ways (a) to use recordings of humans saying the phonemes called the concatenative approach (b) the computer generates the phonemes itself by generating basic sound frequencies (a bit like a music synthesizer) called the formant approach (c) to mimic the mechanism of the human voice called the articulatory approach.

**2.1.4 Lambdacism**

This is as an articulation disorder that occurs as a result of the effect of the excessive pronunciation of the consonant /L/ multiple times in place of the consonant /R/ and vice versa (Georgievska-Jancheska, 2019). A typical example is when “Lorry” is pronounced as “Rorry” and “Laid” is pronounced as “Raid”, the condition is also referred to as a lallation.

**2.2 Review of Related Studies**

Relevant work on some machine learning approaches and algorithms used in diverse natural language text processing is briefly reviewed in this section.

Guo et al. (2024) proposed an unsupervised framework for ASR Error Correction called UCorrect, it consisted of three parts which included a generator; for candidate character generation, a selector to select the most confident character for output, and a detector to detect the position of the character. The UCorrect did not depend on trained data and had a high word error reduction (WER) rate. When the proposed UCorrect was not fine-tuned, it outperformed existing standards by 6.83% and when it was fine-tuned, it surpassed the non-linear auto aggressive (NAR) correlation model with a low latency and by a more significant margin of 14.29%.

The design of a text-to-speech, text to audio and image to audio, and video to the audio system by (Thanneru et al., 2023) described a prototype that allowed users to hear the content of text images, thereby reducing the problem of linguistic bias among communicators. This entailed taking the text out of the picture using the user’s phone and speaking it out in the user's chosen tongue majorly advantageous to individuals with visual problems when visuals are being displayed. The designed system enabled the user to capture a picture, which the application scans and interprets to read the English text. After that, the data is transformed into speech, making it possible for the visually impaired user to comprehend the text's substance. To enable access to the content of the document, the output was spoken aloud. Natural Language Processing (NLP) techniques such as the Naïve Bayes, and Convolutional Neural Network (CNN), were employed by the system to guarantee improved accuracy and performance. A Graphical User Interface (GUI) was incorporated into the system design to enhance precision and user-friendliness.

The development of an end-to-end controlled speech synthesis system that made use of prosodic representations at discrete phoneme level based on F0 and duration clustering (Ellinas et al., 2023) established alignments between the prosody encodings and the decoder hidden state, a corresponding attention module and an extra encoder for the discrete prosodic representations were incorporated. For every speaker's recording, the study made use of augmentation, feature normalization, and a universal clustering technique. Phonemes were generated from the input text by a front-end pre-processing module and then used as linguistic inputs by the acoustic model. A forced-alignment approach based on an HMM monophone acoustic model was utilized to get precise alignments between the speech and its phonetic transcription.

In a journal published by (Tan et al., 2022) a fully automated speech system for speech correction and accent reduction named CorrectSpeech was developed to correct inappropriate words and mispronunciations that need to be corrected. This was achieved by recognizing errors in speech first and then using a time stamp to determine the location of the errors to generate the corrected speech. This was possible to achieve using the Centre for Speech Technology Voice Cloning Toolkit (VCTK) and L2-ARTIC datasets. Speech-to-text (S2T) F-align and CTC-align modules were deployed with the F-align performing better by 43% with a 95% confidence interval. The speech was discovered to need modifications at the word level for better naturalness.

To detect fake news from news articles, Sharifani et al (2022) summarized some common NLP tasks and common machine learning algorithms. Classifiers such as Count Vectorizer, Naive Bayes, Support Vector Machine, and Passive aggressive were applied to achieve a 93% success rate for fake news detection. These classifiers focused on extracting and incorporating the various features of text into differentiating between fake and non-fake news based on the availability of corpora. Texts were tokenized and words were converted into their base form for a clearer understanding. Due to its higher accuracy rate, the Porter algorithm was used for stemming. An automated fact-checking ML system that combines data and knowledge to help non-experts check the content of the news thoroughly in real time via improved NLP processes after known and existing facts have been compared was later recommended to further improve the algorithm.

Ali et al., (2021) designed a voice recognition system to detect impostors by combining the most effective machine learning approaches for classification and speaker recognition implemented. Various methods of audio pre-processing techniques which included vocal enhancements and noise reduction were applied in this study to enhance raw audio. The study explored two .WAV audio data sets (Voice Data sets and Speaker Recognition Audio Data sets) on the Naive Bayes, JRip, J48, PART, Random Forest, and K-Nearest Neighbour algorithms while the differential and acceleration of each audio was extracted using Mel Frequency Cepstral. The pre-processing technique applied included noise removal and hamming, smoothing, and signal domain transformation. The accuracy extracted from the audio for each classifier showed that the Random Forest and Naive Bayes presented the highest performance while the PART presented the lowest performance. The study demonstrated that using a machine learning classifier for the classification process improved accuracy by 97.9%.

Research was carried out by (Oladipo et al., 2021) to deploy a supervised learning algorithm that provides support for the identification of an accent-dependent automatic speech recognition among the three (3) major ethnic groups in Nigeria (Hausa, Igbo, and Yoruba). This was made possible by training and classifying acoustic features of Mel-Coefficient Frequency Cepstral Coefficients (MFCC) on Logistic Regression, K-nearest neighbour, and Gaussian Mixture Models with Logistic regression emerging as the best with 82% accuracy in accent identification for the three languages. The research only identified accents in digital speech recognition but did not correct the errors.

To close the performance gap between native speaker (L1) and non-native (L2) English speakers using ASR models, (Shibano, et al., 2021) carried out research that involved the fine-tuning of (i) pre-trained wav2vec 2.0 models and (ii) zero-shot model on a non-native L2-ARCTIC English speech corpus using various training conditions. Models trained with diverse accents (accent-independent) were compared to models trained with a single accent (accent-dependent) model. The result showed that on the word-2-vec model, with a small amount of pre-trained labelled data, there was an improvement on L2 speakers for both multi-accent and single-accent models, while on the zero-shot model, improvement occurred on only single-accent models.

A system that enabled users to carry out operations such as Exit, Open, and Save by converting input audio into text and Speech Recognition was designed by Vashisht et al., (2021) to aid information sharing, bridge the language barrier, and boost accurate translation between English and Hindi speakers with its output being in text form. The study used the Neural Machine Translation (NMT) algorithm to construct an encoder-decoder architecture for the translation, this encoder-decoder text used the Gated Recurrent Unit (GRU) instead of the Long-Term-Short-Term (LSTM) memory alongside a bidirectional forward and backward input sequence then concatenated before passing on to a decoder. The input was placed within an encoder model that provided the concealed shape state of the form encoder's output. The study presented a real-time voice detection and translation device that made use of a multimodal vector source to source to function. The system designed successfully converted English to Hindi and Hindi back to English.

A text normalizer that transcribed Vietnamese speech into text was developed by (Tran & Bui, 2021) to automatically convert proper nouns and format transcriptions like date, time, and numbers into applicable expressions. Text normalization was carried out on 13 sentences using deep neural networks that had manually designed rules, these rules helped to recognize, design, and convert the text sequence. After a pre-processing stage, proper nouns were recognized through the design of labels and data tagging to develop a corpus of Vietnamese texts used as a data set after which a Convolutional Neural Network (CNN) was used to encode the characters and further concatenated them into bi-LSTM encoders, CRF was then used to predict the labels for each word. The experiment was carried out in two ways: first by using rich features to integrate the text normalizer into the real speech-to-text system to measure its performance on real outputs and secondly, by evaluating the effectiveness of the baseline of the normalizer designed without using rich features. This resulted in a 90.67% F1 score against the rule-based approach with an F1 score of 84.07% in text recognition and conversion sequence.

A Framework to separate error tendency and error detection in Computer-Assisted Pronunciation Training (CSAPT) Computer Aided Language Learning (CALL) into accent and lexical errors by Kyriakopoulos et al., (2020) using three (3) corpora: the Business Language Testing Service (BULATS), SELL-CORPUS, Foreign Language (LeaP) Project. While accent errors occur when pronunciation is not done correctly, the research generated lexical errors by passing the spelling through a grapheme-to-phoneme (G2P) system trained to pronounce and delete canonical pronunciations from the output. The framework was able to successfully recognize word-level accents and lexical errors on the LeaP project but not on the BULATS and SELL-CORPUS.

When a necessity arises for people to learn a second language (L2), the major challenge faced is usually the wrong pronunciation of highly stressed words while speaking. (Korzekwa et al., 2020) designed an attention-based multi-speaker learning model that would automatically derive an optimal syllable-level representation of audio features from phoneme and frame level. This was achieved by generating Neural Text-To-Speech with datasets undergoing Feature Extraction and Attention-Based Classification Model training between word-level attention-based model and lexical stressed models. This boosted the performance of the attention base by 14.8% while maintaining close to 50% recall. This model was however not able to classify lexical stress correctly when there is a link between two words.

Automatic Speech Recognition (ASR) technique was used to analyse five classifications of disordered speech which include: stuttering, dysarthria, apraxia, cluttering, and lisping in Arabic language (Arpitha et al., 2020). Speech signals were detected using the Mel-Frequency Cepstral Coefficients (MFCC), Gaussian Mixture Model (GMM), and Hidden Markov Model (HMM) to differentiate between disordered and normal speech by correlation. One (1) normal speaker was compared to fifty (50) speech articulation-affected speakers. The disordered speech was recognized using a voice-input voice-output aid (VIVOCA) giving an 82.50% efficiency from 56% when HMM and MFCC were used to train its acoustic models as against 96% recognition accuracy from normal speech. The study showed that disorganized speech can truly affect the accuracy of automated speech recognition.

In a book review of a machine learning-based speech synthesis, Shiga et al. (2020) explored the fundamentals of the research and development of the multi-lingualization of speech-to-speech translation technology of speech signals among some languages including English, Japanese, Thai, etc. Texts and speech samples of each of the languages were taken to form phonemes to phonemes and graphemes to phoneme grouping of words, its features extracted, and then TTS was performed to present a speech corpus. To generate the speech signal, the multi-lingualization process involved creating a speech corpus for the new language by, formatting text, recording a certain amount of human speech in that language, and training the models using the speech corpus of word segmentation and pronunciation dictionary. POS tagging was also introduced in the process. The HMMs and DNN models were applied as generative models to train and synthesize speech.

In an analysis written by Ranja & Thakur (2019) on various feature extraction techniques for speech recognition systems, audio signals were filtered via the Mel-frequency cepstral coefficient feature extraction (MFCC) method. The MFCC method was used to perform speaker identification, noise reduction in audio signals as well as voice classification. Types of MFCC feature extraction methods including the MFCC, Delta MFCC, and Delta-Delta MFCCs were compared with the Delta-Delta MFCC with minimum standard deviation being the best feature extraction technique for audio signals.

An article written by Razno (2019) expounds how machine learning and natural language processing techniques apply to word processing tasks written in human language, these tasks are furthermore utilized with the Python programming language and its frameworks by using it to illustrate the idea of ML as a whole while working on text data The study used over 200 Wikipedia articles introduced the text processing-based machine learning classification model algorithm such as natural language tool kit (NLTK), Scikit-learn and Artificial Intelligence, by pre-processing the developed text corpus from those articles, identifying objects and subjects of expressions: verbs inclusive, identifying and extracting semantics from those articles. Automatic text extraction and analysis tools that operated on a higher level and used technologies for text mining were developed to aid the optimization of search engines, query extensions, and oncology constructions.

A review that observed the various algorithms and techniques by which Speech-To-Text (STT) and text-to-speech (TTS) recognitions are achieved and how they can be applied to assist the disabled (Trivedi et al., 2018) using the Indian language as a basis discussed and compared some machine translation models such as Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), Example-Based Machine Translation (EBMT), and Hybrid Machine Translation (HMT) for feature extraction, Linear Predictive Coding (LPC), Mel-Frequency Cestrum Co-efficient (MFCC), and Dynamic Time Warping (DTW) techniques were measured; for pattern matching, the study measured Knowledge based, Neural Based, Statistical Based, and Hidden Markov Model (HMM) techniques then for the speech to text conversion, the study applied Artificial Neural Network based Cuckoo Search Optimization technique. The study showed how the HMM served as a better model for Speech signals in the speech-to-text conversions after undergoing feature recognition, pattern matching, and speech-to-text conversions respectively.

A study on lambdacism (Oyeka, 2017) focused on how to help secondary students of some schools in Agbor, an Alor Igbo Language speaking community who are learning English as their second language pronounce English works more accurately and fluently. The study observed that the students were faced with an error of lambdacism which occurred as a result of interference from the mother tongue even after the students had been taught about the variation between the /L/ and /R/ alphabets. In the learned language, English, the phoneme /R/ is erroneously interchanged by the pre-existing phoneme /L/ and vice versa. This error in speaking led to the need for the schools to employ more language instructors alongside procuring a variety of teaching aids to facilitate learning the English language. Students who were affected by lambdacism were advised to speak slowly to overcome these struggles or hide the errors in cases that were not manageable.

Nchena & Larsson (2017) in reviewing data structures and sort algorithms, employed two data structures (arrays and linked lists) to compare five sort algorithms: bubble sort, selection sort, insertion sort, merge sort, and quick sorts. The research found that the various algorithms performed differently with various sample sizes when a reverse order was inserted, three test cases were used for each of the two data structures by populating both the array and lists with over 1000 records after which the program generated some random numbers to be sorted using a timer. The research showed that with a small sample size on an array, the insertion sort and selection sort algorithms performed more efficiently than the rest.

A review of learning paradigms used in ASR for both space exploration and home automation on dictated speech using ML techniques such as Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Support Vector Machines (SVM), Artificial Neural Networks (ANN) by (Padmanabhan, & Premkumar, 2015) showed resolution of issues such as multi-modal and multi-lingual recognition as well as noise interference from the environment. DNN to HMM hybrid models proved to have better performance than the ANN-based speech recognizer although its learning technique was more difficult than the rest.

The technique in natural language engineering that develops natural language outputs from non-linguistic inputs is known as natural language generation (NLG), (Semaan, 2012). The study showed how an NLG system performed when an algorithm was applied to it by comparing types of NLG systems such as canned text: simple domains such as generators, horoscope machines, and template filling: natural statements generated via entering data into slots and fields. The study showed that while canned texts are easily implemented, they are hardly adaptable to changes without a programmer making modulations. On the other hand, template filling although easy to implement, cannot handle variations in applications.

Trilla (2009) reviewed the use of NLP techniques for end-to-end speech production (TTS) and inversely for ASR. The research went further to reflect the significance of NLP techniques while processing synthesized speech from input text. Some Techniques studied include sentence segmentation, tokenization, Parts of speech tagging, Grapheme to Phoneme (G2P) Conversion, Word Stress for Text Normalization Then Context-Free Grammar (CFG), N-gram Language Models, and Word Lattices for Speech Recognition. The review stated that although with more transcription accuracy and excellence, the traditional rule-based approach for phonetic transcriptions is more rigorous and time-constraining to develop and difficult to maintain than the computer learning approach.

The invention of dynamic computing devices to assist people in both work and leisure is accomplished with significant human–machine communication (Juang & Furui, 2000). Naturalistic human–machine contact started with the most important stage of automatic speech detection and comprehension which is Artificial Intelligence. The study provided an overview of the evolution of spoken language technology from two approaches: the horizontal spectrum of technical approach and the vertical perspective of chronology, it furthermore drew attention to the advent of statistical techniques for solving language-related issues since they signify a standard change in the field of spoken language processing research. Statistical techniques were utilized to enable automatic speech identification and understanding by teaching the machine directly from data the structure, irregularities, and regularities in the speech signal. The study performed speech-to-text conversion by formulation and word decoding using the Naive Bayes decision theory with the observation sequence consisting of multiple classes. Signal analysis was used to extract salient features from the speech waveform, the post-processing included the use of the Legendre Polynomial and Slepian filter noise and error filtering for optimal pattern matching. The Hidden Markov Model (HMM) was used for characterization while the N-gram Module was used for a language structure. The study provided a speech modeling that was not just easy to implement, but also highly effective.

Table 2.1: Review of the related studies

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S/N | AUTHOR(S) | TITLE | RESEARCH OUTCOME | RESEARCH LIMITATION | RECOMMENDATIONS |
| 1 | Guo et al (2024) | UCorrect: An unsupervised framework for automatic speech recognition error correction | Over 14.2% improvement on WER when the characters were fine-tuned. | It was limited to the use of pre-trained data sets | Recommended that untrained datasets be tried with the proposed UCorrect. |
| 2 | Thanneru et al (2023) | Image to audio, text to audio, text to speech, video to text conversion using, NLP techniques | The system design improved the precision and user-friendliness of text, speech, audio, and video conversions for the visually impaired. | The research was limited to texts from pictures captured using mobile phone | Accuracy can be increased with the use of multiple algorithms |
| 3 | Ellinas et al (2023). | Controllable speech synthesis by learning discrete phoneme-level prosodic representations. | Development of an end-to-end controllable speech system. | It is limited to the use of intuitive discrete labels. | Further study recommended |
| 4 | Tan et al (2022) | Correct Speech: A Fully Automated System for Speech Correction and Accent Reduction. | 43% improvement in accent reduction | Limited to the use of VCTK & L2-ARCTIC data sets as against primary data. | Worded data, ASR model with a high phone accuracy and a higher frame rate. |
| 5 | Sharifani et al (2022) | Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model. International Journal of Science and Information System | Machine learning classifiers were trained on NLP and used to sieve out fake news from newspapers via the use of Text Classification modules with 93% accuracy. | It is limited to datasets drawn from a public domain sorted with qualitatively with fake, non-fake, and clear labels. | An automated fact-checking ML system that combines data and knowledge to help non-experts and checks the content of the news thoroughly via improved NLP processes after known and existing facts must have been compared is recommended. |
| 6 | Ali et al (2021). | Voice recognition system using machine learning techniques | Highest audio accuracy for Naïve Bayes and Random Forest algorithms. Lowest Audio accuracy for PART algorithm. |  | the study demonstrated that using a machine learning classifier to the classification process improved accuracy; 97.9% was achieved by (RF) classifiers. |
| 7 | Oladipo et al (2021). | Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers. | 82% accuracy on Logistic Regression  75% accuracy on K-Nearest Neighbour  50% accuracy on the Gaussian Mixture Model | Accent Identification | Error correction and accuracy percentage should be increased in the future. |
| 8 | Vashisht et al (2021). | Speech recognition using machine learning. | Developed an Encoder-decoder architecture that converted English to Hindi and vice versa | Limited to the use of a multimodal source of vector | Recommended the further study of the use of other sources of vectors. |
| 9 | Tran et al (2021). | Neural text normalization in speech-to-text systems with rich features | 90.67% F1 score on rich features in text recognition and extraction sequence. | The study was limited to proper nouns, numbers, date, and time. | More research should be carried out in ways proper nouns can be more easily identified just like numbers are. |
| 10 | Arpitha et al (2020). | Diagnosis of Disordered Speech using Automatic Speech Recognition. International Journal of Engineering Research & Technology | Normal Speech had 96% Accuracy. Disorganized speech had 82.5% accuracy after ASR  was performed from an initial 56% accuracy | Smooth conversion of English to Hindi and Hindi back to English. The system designed could perform operations like Save, Exit, and Open through audio only input. | The software design did not accommodate large samples, further works should explore algorithms that could broaden the sample size of audio datasets. |
| 11 | Kyriakopoulos et al (2020) | Automatic detection of accent and lexical pronunciation errors in spontaneous non-nativeEnglish speech. ISCA. | Able to detect word-level accent and lexical errors in Leap Corpus data sets | Not able to detect errors in BULTAS and SELL\_CORPUS datasets | Recommended for further study on the detection of errors in sell\_corpus |
| 12 | Korzekwa et al (2020) | Detection of lexical stress errors in non-native (L2) English with data augmentation and attention | 94.8% precision & 49.2% recall in detecting incorrectly stressed words in L2 Baltic-English speakers. | Neural Text-To-Speech (TTS) using correctly and incorrectly stressed words | Recommended for further study on the processing of large data set at a time |
| 13 | Razno (2019). | Machine Learning Text Classification Model with NLP Approach | Automatic text extraction and Optimization of search engines, query extensions & oncology constructions using python language. | NLTK, Scikit-learn, Artificial intelligence | Recommend the use of another programming language apart from Python to implement the search engine optimization |
| 14 | Ranjan & Thakur (2019) | Analysis of Feature Extraction Techniques for Speech Recognition System | Delta-Delta turned out best for MFCC feature extraction. | Limited to MFCC only | Other Feature Extraction Techniques such as Perceptual Linear Prediction (PLP) and Linear Predictive Codes (LPC) can be explored. |
| 15 | Trivedi et al (2018). | Speech-to-text and text-to-speech recognition systems-A. Review. | The production of speech signals with the HMM technique performs best for text conversion. | Limited to generating speech signals for text conversion only | Further studies to explore other TTS and STT techniques. |
| 16 | Oyeka (2017) | Causes And Effects Of Lambdacism In The Speech Of Igbo-English Bilinguals In Alor Dialect | More teachers were employed to train students | Limited to Alor Igbo and English languages only. | Training should start at the elementary level and a device should developed to aid in speech translations. |
| 17 | Nchena & Larsson (2017). | Sort Algorithms and Data Structure: An Overview and Comparison. | The insertion sort algorithm performed more efficiently on an array with a small sample size | Bubble sort, Selection sort, Insertion sort, Merge sort, Quick sorts, Arrays, Linked lists | Other algorithms apart from selection sort and bubble sort should be explored. |
| 18 | Padmanabhan & Premkumar (2015) | Machine learning in automatic speech recognition: A survey. | DNN/HMM systems showed to have obtained more significant performance gains as compared to HMM/ANN-based speech recognizers | Automatic Speech Recognition, Gaussian mixture models, Hidden Markov models, Machine learning, & Support vector machines. | The further design of a more scalable, robust, and efficient deep learning architecture for uncertain and untrained data is recommended |
| 19 | Semaan (2012) | Natural language generation (NLG): An overview. | Review of types of NLG | Canned text, template filling. | Further research on other types of NLG techniques should be carried out |
| 20 | Trilla (2009) | Natural language processing techniques in text-to-speech synthesis and automatic speech recognition | Speech produced by the  Signal-processing modules were tightly bound to the performance of the previous text-processing modules. | Limited to the production of voice from an input text via an inverse process | The use of NLP for a more realistic and natural interface is recommended. |
| 21 | Juang & Furui (2000). | Automatic recognition and understanding of spoken language first step toward natural human-machine communication | Development of a highly effective and easy-to-implement speech model | Limited to machine learning the languages via the use of statistical methods | The development of a model that enables the machine to store, retrieve, and represent knowledge is recommended for future studies |

**2.3 Research Gap**

Irrespective of how broad and detailed natural language processing has advanced both in research and innovations, there are still a lot of lapses and areas uncovered within the scope of speech synthesis. This is due to some unavoidable factors such as complexity with text pre-processing methods; getting the machine to recognize abbreviations, special characters and symbols, fractions and dates, accent pronunciations, and fluctuation while reading texts for native speakers of the English language. The proposed research seeks to solve the problem identified but not resolved by (Oladipo et al., 2021) where a supervised learning approach was used to design an algorithm that recognizes and distinguishes the three (3) Nigerian ethnic groups - Hausa, Igbo, Yoruba based by constructing sequential Mel-Frequency Cepstral Coefficient (MFCC) on their features of accents from the audio samples of text read. The study seeks to further resolve an accent-related issue of lambdacism lallation (Bahromovna, 2023) in one of the Nigerian languages (Igbo) studied. The study not only seeks to identify the texts affected by the accents, it also goes further to select and replace affected words with the correct text pronunciation using a concatenating algorithm of both supervised and unsupervised machine learning models. The supervised ML model will be used to classify labelled text data (Jiang et al., 2020) while the unsupervised ML model will be used to categorize unlabelled text data as well as evaluate and validate the models used (Naeem et al., 2023) thereby improving the pre-existing machine performance to accent readability for non-native English speakers of the Igbo Language. For unsupervised machine learning, the model remains untrained and hence finds similarities that allow objects to complete tasks, clustering is majorly used in unsupervised ML techniques alongside bootstrapping (Khan et al., 2016).

**CHAPTER THREE**

**RESEARCH METHODOLOGY**

**3.0 Introduction**

This chapter discusses the proposed methods and techniques to be used within the study including the research design, methodology, data collection techniques, the necessary tools employed to collect data, sample sizes and techniques, methods of data analysis, performance metrics, and evaluation as well as ethical considerations of the algorithm to be designed.

**3.1 Research Design**

The study will follow a True-experimental design method which will rely on the statistical analysis of results to prove or disprove the hypothesis that machine learning can improve the nature of language process using the algorithm. The design pattern would be explorative and explanatory to seek more insight on elements of the research such as the lallation of lambdacism evident in English speakers of Igbo natives.

The design follows a four-stage process of data collection and training, speech recognition (text-to-speech), natural language processing, and then speech synthesis (text-to-speech).

Data collection and training involves the selection of the data sets data cleaning, normalization, and standardization. Normalization is used to equalize the scale of each data point thus giving equal weight to each attribute of the data set. The data cleaned is normalized using the z-score normalization technique also called standardization.

The pre-processing stage in the NLP involves Part of speech (POS) Tagging, Tokenization, Lemmatization, and stemming. The effect of (POS) tagging in machine learning is evident in machine translation, question-answering parsing, word sense disambiguation, etc., it is a grammatical classification that assigns each word of any given text to its proper syntactic tag in its context of appearance which could be verbs, adjectives, adverbs, nouns. It is also the automatic assignment of part-of-speech tags to words in a sentence (Chiche & Yitagesu, 2022).

Tokenization is used to divide the textual information into individual words. The study applies the Python Natural Language Tool Kit (NLTK) for Word Tokenization (Vijayarani & Janani, 2016).

Lemmatization is the process of putting a word's inflected pieces together and then identifying it as a single unit, often known as the word's vocabulary form or lemma, stemming creates variations of a root or base word by lowering the base word to its stem term thus making for a shorter look-up process and the phrases easier to comprehend (Khyani et al., 2021).

Named Entity Recognition (NER) is the identification and classification of desired entities in text and is principal for information extraction in NLP (Sun et al., 2018). Natural Language Generation (NLG) employs the multi-stage process of content analysis, data understanding, document structuring, sentence aggregation, grammatical structuring, and language presentation from a data set to produce written or spoken words (Wigmore, 2023).

The selection sort algorithm is applied to single out words affected by errors, executed by swapping the front element in an array with the smallest element in the array thereby moving it to the beginning of the array (Naz et al., 2021), while the insertion sort algorithm is applied to replace the correct words already removed as it arranges each item in the final sorted array or list sequentially due to the small sample size (Al-Kharabsheh et al., 2013).

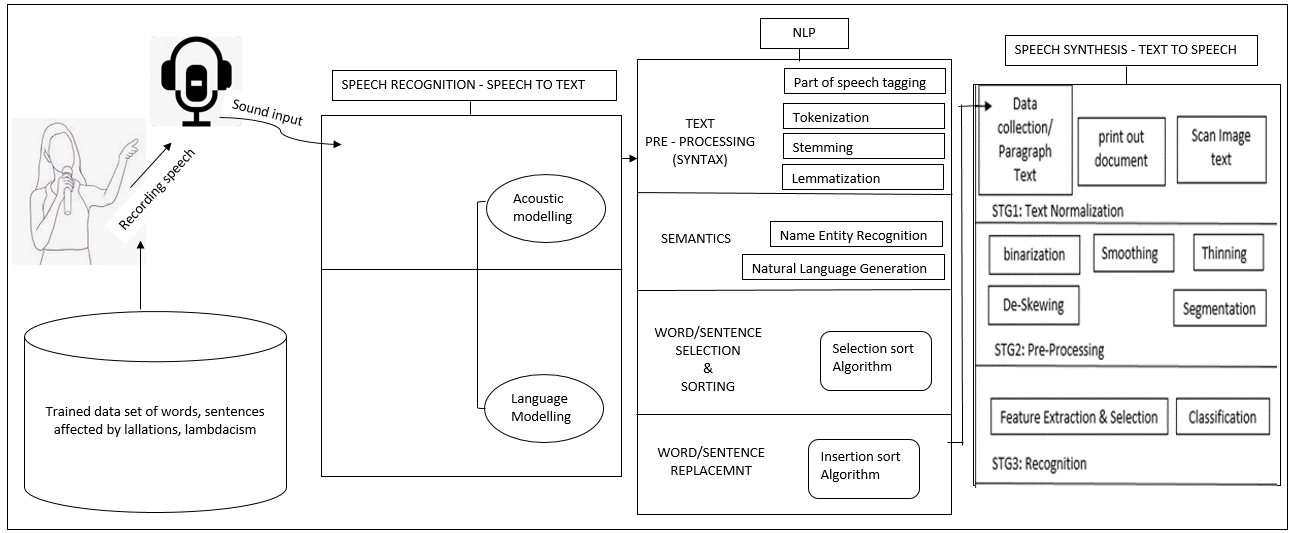
Below is the design framework of the proposed study.

Fig 3.1 Research Design

**3.2 Research Methodology**

The study follows a quantitative and qualitative case study research method. The quantitative research method made use of numeric data while the qualitative method made use of non-numeric data which will enable the contextualization and interpretation of data in a detailed and complete description of observations and findings, the qualitative method is used because the sample size is small and the population is finite.

The first step of the study is to carry out Text Extraction on the input data to be used, after which Speech Recognition will be done on the audio data, text from the recognized speech will undergo an NLP process for language modelling after which speech synthesis will be used to convert speech back to the text and then the algorithm developed will be used for sorting and replacement of corrected speech.

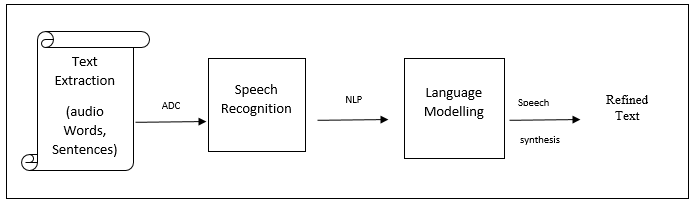


Figure 3.2: Pictorial representation of the method stages for the study

Step 1: Text Extraction

This process is carried out using a dictionary-based Named Entity Recognition (NER) system (Cook & Jensen, 2019) where a tagger is combined with an English dictionary which allows for the extraction of entities of words that could be affected by the lallation. Named Entity Recognition is the task that involves the extraction of names from natural language text through learning (Mansouri et al., 2008). The text mining results are then normalized and maintained by manually curating a block list and regular expressions removing common words in text like and, was, a, etc. to extract the needed texts. The Dictionary-based NER is applied in this study because of its extremely high speed, recall, and precision rate. Texts extracted are then spoken into a microphone and recorded using a voice recorder to serve as audio input for the ADC to convert.

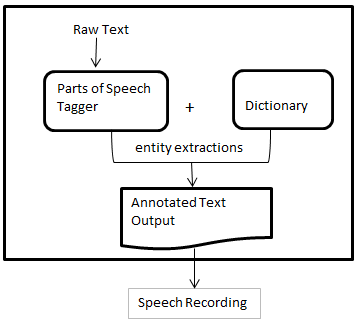


Figure 3.3: Dictionary-Based NER Architecture for the Text Extraction

Table 3.1: Sample Dataset of Texts Extracted from the Dictionary-Based NER

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Allow | Alive | lactose | Laid | Latrine | Laugh | Lavatory | laxative |
| Lead | Led | Leech | Legal | Leprosy | Lexical | Liberal | liberation |
| Liberate | Literal | Litany | Locust | Locative | Locum | local | loaf |
| Loot | Light | Leap | Love | Luxury | Loyal | Labarum | lability |
| Blank | Bleat | Bleach | Bleed | Blind | Clime | Clamp | Cloak |
| Plural | Reflex | Flower | Flood | Glow | Glide | Blow | Delay |

Step 2: Recording speech

The process is done with the use of a microphone, recording equipment, and computer device and saved in an acceptable format. The study saves the recorded speech in a .wav audio format

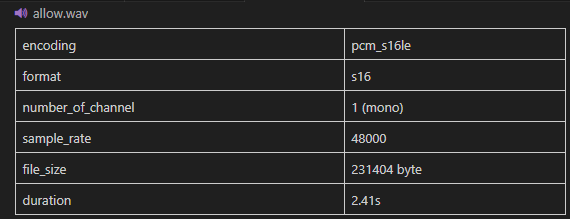


Figure 3.4: Recorded Speech saved in .wav format

Step 3: Audio Normalization:

All words are made of a distinct vowel sound which has various frequencies recorded on the device. The various frequency periods are pre-programmed on the computer allowing it to recognize when a spoken sound matches the vowel sounds known as phonemes. The process of normalization enables the machine to understand, remove unnecessary noise, and normalize the sound and speed of speech as different people speak differently, measuring an audio file's aspect and modifying it to meet a predetermined target is known as audio normalization. This study uses the instance normalization process to convert the audio from stereo track to mono track (Chang et al., 2021).

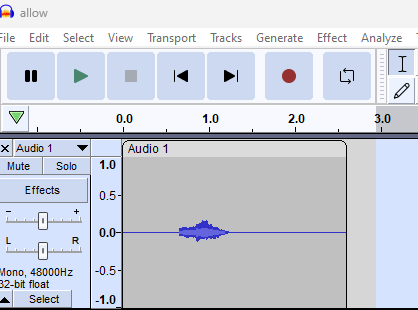


Figure 3.5: Normalization: converting audio from stereo to mono track

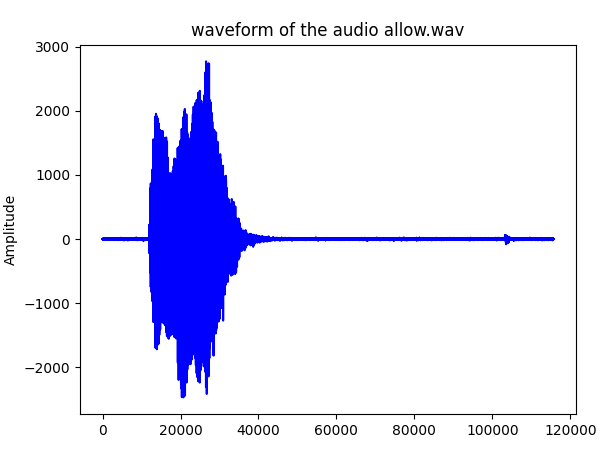
Step 4: Convert the sounds from recorded speech into digital formats of 0 and 1

The study read the audio file by performing digital signal processing (DSP) and returning the results in a NumPy array where nchannels is the number of channels, sampwidth is the number of bytes per sample, sampling\_rate is the sampling rate, nframes is the total number of samples as seen below.



Fig 3.6: Digital form of the audio after conversion.

The values from the conversion above are then plotted into a waveform with Amplitude on the y-axis and time on the x-axis in seconds.



x =5.84e+04, y =4.8e+02

Fig 3.7: waveform of audio file

Step 5: Fast Fourier Transform (FFT)

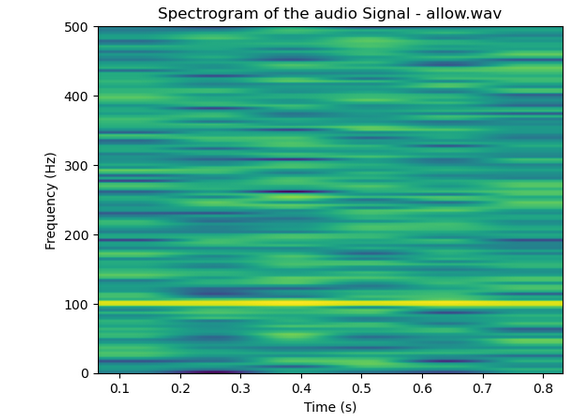
The study applies the Fourier Transform (Saddam, 2022) on each framed and windowed segment of the audio to convert the time-domain signal into the frequency domain resulting in to creation of a spectrogram that represents the distribution of frequency components within that frame, FFT is applied to process the signals in the frequency and spectral domain. 

Figure 3. 8: FFT Spectrogram for the audio file

Time is on the x-axis and Frequencies are on the y-axis. The intensity of the different colours shows the amount of energy that is how loud the sound is, at different frequencies, at different times.

Step 6: Feature Extraction:

The study uses the Mel-Frequency Cepstral Coefficients MFCC (Chang et al., 2021) to map the power spectrum obtained from the FFT onto a mel-frequency scale of  (Hui, 2019)

and then applies a discrete cosine transform (DCT) to capture the most relevant coefficients.

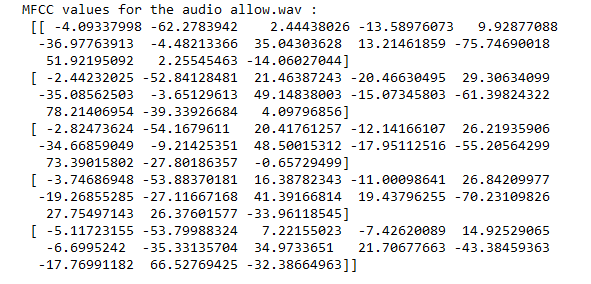


Figure 3.9: Array of MFCCs feature extraction for the audio file.

The study implements the extracted features of the MFCCs above as inputs for machine learning algorithms such as Hidden Markov Models (HMMs), Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), or Convolutional Neural Networks (CNNs) to perform speech recognition tasks. This study will make use of the HMM for the recognition process.

Step 7: Speech Recognition Process:

The audio file undergoes a speech detection process using Automated Speech Recognition is performed on the .wav file to show that the machine is affected by the error of lallation hence converting “Allow” to “Arrow” as seen in the figure below.

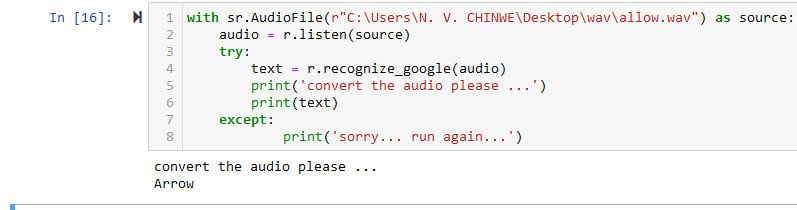


Figure 3.10 Speech (word) converted to text with ASR showing errors

The word allow is inculcated into a short sentence “Allow me take a nap” and read by an English speaker of Igbo native, after undergoing ASR, the machine read the sentence as “Around me take a nap” showing the availability and un-avoidance of machine reading error due to lallation.

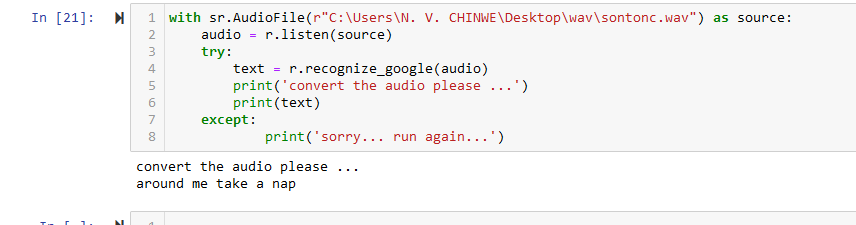


Figure 3.11 Speech (sentence) converted to text with ASR showing errors

The text converted by the machine will undergo an NLP process in readiness for sorting and replacement of the corrected speech.

Step 8: The NLP process

Text Pre-processing:

The WordPunctTokenizer tokenization (Pactpub, 2020) method is used to tokenize the text because of its efficiency as it captures punctuation by splitting them into separate tokens, unlike the WhiteSpaceTokenizer which tokenizes punctuations alongside the words.

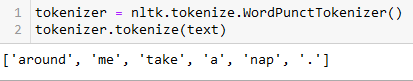


Figure 3.12 Text tokenization using WordPunctTokenizer

Text is normalized such that the machine can recognize it in both uppercase and lowercase letters.

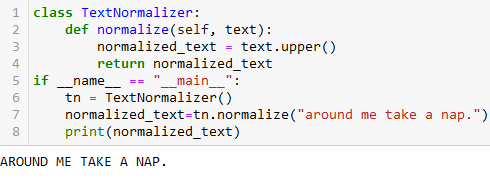


Figure 3.13 Text Normalized

Feature Extraction: the study applies the bag of words model (Bhuiyan et al., 2021) to extract features of the text, each token has a feature column called text vectorization because text is replaced with a huge vector of numbers and each dimension of the vector corresponds to the token on the database. Seeing that the tokens are not ordered in the bag of words, we employ the n-gram frequency count of tokens in pairs or triplets. The counters are replaced with Term Frequency-Inverse Document Frequency - TF-IDF values and calculated by taking the term frequency (tf) of our term (t) and multiplying it with the inverse document frequency (idf) of the corpus.

Equation 3.1

 (Zhao et al., 2022)

While (w) represents the i-th word in the vocabulary based on the construction.

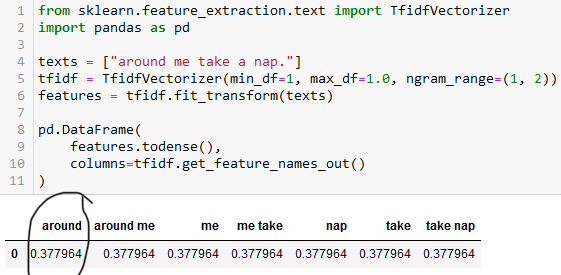


Figure 3.14 Feature extraction using TF-IDF

TF-IDF statistically evaluates the significance of a word in a text and is reached when we have a high frequency in the given document and a low document frequency of the term in the whole collection of documents.

The n-gram range tells the TF-IDF vectorizer what n-grams to use in the bag-of-words representation between 1-gram and 2-gram. The features extracted from this process will be used to train our model.

Text Classification and Sentiment Analysis: This process is used to classify unstructured text data to understand user sentiments and user behaviour surrounding the text. Support Vector Machine, (Camastra & Razi, 2020) a supervised machine learning method is applied to the classification process. A linear kernel is used to find the hyperplane that separates the data points in the training set with the farthest distance. To achieve this, texts have to first be turned into vectors, these vectors are lists of numbers that represent a list of coordinates in some space. Its primary variants make use of the most widely used Python tools and packages for handling text data. The corresponding results to give a Mean Square Error (MSE) of 0.5422430057681027.

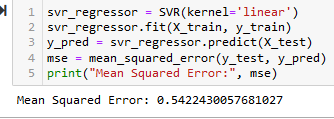


Figure 3.15 MSE of the linear kernel

**3.3 Procedure for Data Collection**

Primary data is collected experimentally and manually using words formed with letters from existing old data of the English Language alphabet (Ten-Hacken, 2002). These words will serve as the input text data to be made into sentences and recorded using a microphone and recording machine. This recorded speech serves as the audio input data for the research and is paired with a transcriptor. The table below gives an overview of how letters from the English Language alphabet are grouped into phonemes for possible use in word and sentence formation for the audio and text dataset.

Table 3.2: consonants, vowels and phoneme grouping for the R-L word/sentence structure

|  |  |
| --- | --- |
| **Consonant phonemes** | **Long Vowel phonemes** |
| b, d, f, g, h, j, k, l, m, n, p, r, s, t, v, w, y, z | a, e, i, o, u, and oo |
| **Consonants Blends of L and R** | **R-controlled vowels** |
| cl:, fl:, gl:, pl:, sl:,  br:, cr:, dr:, fr:, gr:, pr:, tr:, sk:, sp:, st:, sw:, spr:, str: | ar, er, ir, or, and ur |

**3.4 Tools and Instruments Used for Data Collection**

Tools and instruments used for data collection in this study include texts – from the Oxford English Dictionary (Oxford Dictionary, 2023) for text data and mechanical devices such as human speech, (Igbo native English speakers), a microphone, a recording machine, a computer device (laptop), power supply, internet, camera, software applications: python – (pyaudio, audacity, NLTK), a soundproof system for audio data, etc.

**3.5 Validity and Reliability of the Instruments Used for Data Collection**

The reliability of the proposed study will be measured using the Longe (2024) Inter-Rater Reliability measure, this method ensures the elimination of experimenter bias as it helps to judge the outcome of the proposed model from multiple perspectives of observers while the viability of this proposed study is to be measured using the Concurrent Validity measure which would require the test of several variables at a time.

**3.6 Population and Sample Size of the Study**

The target population for this study is the non-native English speakers of the Igbo language who are affected by the lallation of lambdacism. With a finite population size of 650 words of audio data, a population percentage of 5, a confidence interval of ± 0.01, and a 99% confidence level, the proposed sample size used is calculated to be 45 using the simple sample size formula for finite population.

Equation 3.2

 (Chaudhuri & Dutta, 2018).

Given that Y¯ is the sample size, n is the percentage of the population, N is the population size, and y¯ is the sample mean.

**3.7 Sampling Technique**

The sampling technique to be used for the proposed study is probability sampling (Etikan & Bala, 2017) which gives the best accuracy for qualitative research. Simple random sampling with replacement (SRSWR) (Chaudhuri & Dutta, 2018) allows every case of the sample population to be included hence it will be used in this study.

**3.8 Methods of Data Analysis**

Given the study works with text, audio, and image data, the Grounded Theory Analysis (Clarke, 2021) is applied to transform qualitative input into quantitative data as it involves the systematic collection and analysis of primary data, repetitive ideas, and codes giving room for new theories to be formed.

**3.9 Performance Metrics**

Squared Error regression metrics, are used to measure the average squared difference between the actual and predicted values of the regression model which is the Gaussian distribution. In a given dataset of x observations, if a is the actual value, and b is the predicted value for the nth observation, The Mean Squared Error (M2E), (Shah, 2023) is calculated using the formula

Equation 3.3

.

**3.10 Evaluation of Model Performance**

The study makes use of the model monitoring evaluation technique of precision and recall for the classification metrics and weighted mean percentage error for the regression metrics to evaluate the model to be developed (Erickson & Kitamura, 2021). Some advantages of model monitoring include cost-effectiveness, efficiency gains, and time-friendliness.

**3.11 Ethical Consideration**

The ethical consideration to be implored in the study is truth versus loyalty. Data will is gathered and stored securely, participants of the research will be encouraged to take part willingly and not forced, applications used are obtained legally and all data involved are secured efficiently, the integrity of the researcher and participants are confirmed to be of optimal standards.

**3.12: Dissertation Conclusion**

This chapter begins with an introduction of the chapter, then further describes the research design to be used as well as the methodology, the methodology process was also discussed. The instruments and tools used for data collection, validity and reliability of the instruments used were elaborated, the proposed population and sample size were discussed as well as the method of data analysis, performance metrics, evaluation models and ethical considerations to be observed at the course of the study.

**REFERENCES**