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

Automatic Speech Recognition and Accent Identification of Ethnically Diverse Nigerian English Speakers

Francisca O. Oladipo
Federal University Lokoja
Department of Computer Science

Rahmon A. Habeeb
Federal University Lokoja
Department of Computer Science

Abraham E. Musa
Federal University Lokoja
Department of Computer Science

Chinecherem Umezuruike
Kampala International University
School of Mathematics and Computing

Ohieku Andrew Adeiza
Department of Mathematics
Capital Science Academy, Kuje FCT – Abuja

ABSTRACT
It is imperative to improve the speech recognition system as human-machine interfaces are advancing in the growing global market of technologies. There are quite a number of Nigerian English speakers' accents to which the speech recognition systems are not sufficiently exposed. Accents may suggest a lot of information about someone's whereabouts, for example, their native language, place of origin, or ethnic groups and accent classification. Given the importance of accents, efficiency and accuracy of speech recognition systems can be improved with training data of diverse accents. This research provides support for accent-dependent automatic speech recognition by deploying a supervised learning algorithm to the task of recognizing three Nigerian ethnic groups (Yoruba, Igbo, and Hausa) and distinguish them based on their accents by constructing sequential Mel-Frequency Cepstral Coefficients (MFCC) features from the frames of the audio sample. Our results show that concatenating the MFCC features sequentially and applying a supervised learning technique to provide a solution to the problem of identifying and classifying accents works efficiently and accurately.

General Terms
Speech Recognition

Keywords
Acoustic modeling, non-native speaker, speech recognition, supervised learning

1. INTRODUCTION
A person's accent can be a pointer to his/her first language or mother tongue. The ability to recognize different types of accents can improve the quality of transcription in a text language; allowing for specific preprocessing of recordings [6]. Communication plays a vital role in our daily life activities, and interactions between fellow humans and machines. Among the various means of communication viz. writing, gesture, posture, eye contact etc., speech is the most famous, convenient and understandable. Communication via speech includes verbal pronunciation, expression, and fluency. Due to differences in the articulation of speech (sound), English; the most widely used language, has appeared in several languages. Various factors, such as colonization, trade, tourism, and migration have helped the spread of English in many parts of the world, including Africa, Asia, and South America. The spread of English has given birth to different varieties of spoken English such as Nigerian English (NE), Singaporean English (SE), Malaysian English (ME) etc. Resultantly, it is being spoken in diverse accents across the globe [1].

One of the major challenges in speech recognition is to understand speech by non-native speakers. Nigeria had been a British colony and home to multiethnic groups who used distinct English accent dependent on their ethnicity. Therefore, Nigerian accent speech results in phone calls that are not typical to a language and thus makes the speech recognition a very difficult task to achieve. Accent detection or classification can improve the quality of speech recognition. The automatic speech recognition system can first identify the ethnicity of a speaker and then use an automatic speech recognition system that is trained for that particular accent. Beyond accent recognition which provides identification of a speaker's ethnicity, it is crucial in security related applications such as criminal investigations. In real-life applications, it becomes essential to recognize accents from only short snippets of audio recorded from a distance.

Accent classification or accent identification is a topic of interest among the Speech Recognition development community since the accent is one of the limitations, next to gender, that influences Automatic Speech Recognition (ASR) system performances [4]. The training languages in this research work would be Yoruba, Igbo, and Hausa. In this research, we solve this problem by classifying a few seconds accented voice clip as one target language. The approach used in this task of accent classification is feature extraction (extract characters) and Machine Learning classifiers. Our goal is to classify different types of accents, especially Nigerians, in the speaker's native language. When a speaker speaks a script of English words, given a recording of the speaker, we want to predict the speaker's native language by applying a machine learning classifier.

As our country, Nigeria, has been populated by many different ethnic groups, there arise different accents of the pronunciation of English in their speaking, inherited from their own mother tongues phonemes inventory. This complexity presents the most critical issue for automatic speaker-independent speech recognition (SI-ASR) systems [20]. As far as being concerned, there is no special industrial ASR available in the market for Nigerian English (NE) to tackle the deviation due to this ununiformed variation among the population. As speech recognition and intelligent systems

are more prevalent in today's society, we need to account for the variety of accents in spoken language. This project aims to develop a model using some machine learning classifiers to identify accent varieties of three major Nigerian languages (Yoruba, Igbo, and Northerners). The specific objectives are to create a Nigerian dataset for an audio processing project based on the three languages, build a model that can be integrated into existing Automatic Speech Recognition (ASR) system, and successfully classify and identify Nigerian accented speech while predicting the target class for each case in the data.

This paper is divided into five sections and is organized as follows: The first section gives a general introduction to the subject areas and specific introduction to this research concern. A review of the literature of related research and concepts is conducted in the second section while the third section describes the methodology, tools, and techniques deployed in this research work. In the fourth section, we shall discuss our results and conclude the paper in the final, fifth, section

2. LITERATURE REVIEW

Oral Language is the natural source used by humans for communicating the information. When considering speech production, it conveys information of language and spokesperson. Whereas, when we consider speech perception, it also conveys the atmosphere in which the speech was produced and traveled [15]. This reorganization skill inspires the researchers to develop such a system that extracts information from speech through human-machine interaction. Speech recognition is aided by the uniqueness of the voice which comes from the movement of tongue and mouth. [11].

In the 1950s, the first attempts for Speech processing were made after researchers tried to exploit the fundamental ideas of acoustic-phonetics. In Automatic Speech Recognition (ASR) designs, the speech was segmented into several segments and then entities were classified for the regents. The techniques used at the time were Linear Prediction Coefficient (LPC), Spectrum and Fundamental Frequency histogram, Instantaneous Spectral Covariance matrix, and Mel-Frequency Cepstral Coefficients (MFCC). [17] A speech recognition system can operate in many different conditions such as speaker-dependent/independent, isolated/continuous, and small/large vocabulary speech recognition. Speech recognition systems can be separated into several different classes by describing what types of utterances they can recognize.

Cepstrum representation of an audio clip aids in the derivation of Mel-Frequency Cepstrum Coefficients mainly by utilizing Fourier transform of signals. The automatic classification system was for foreign Australian accented English by using the HMM phoneme model [14]. The speech recognition accuracy of 91.89% was reported by [13]. This was achieved by recording continuous speeches of 446 uncommon words by 20 speakers, and MFCC and discrete wavelet packet decomposition (DWPD) features were extracted by MATLAB [13]. Research on the automatic identification of Vietnamese dialects based on Gaussian mixture model (GMM) using MFCC and tonal features was conducted by [10]. To differentiate various South Indian languages, a native identification model based on English accent was developed utilizing GMM [12].

When two distinct classifiers are used, Support Vector Machine (SVM) achieved 48.16% accuracy when trained with MFCC features and 41.18% accuracy rate when trained with PLP features [19]. The study on classification of English accent signifies that the supervised learning algorithm performs better than the unsupervised learning algorithm [6]. The Machine Learning algorithm was used to classify accents from seven foreign-accented English. This showed that the text-independent classifier achieved 41.38% accuracy while that of text-dependent classification rate was 45.12% [2]. Phonetic knowledge was also combined to develop a GMM classifier with PLP features optimized by Heteroscedastic Linear Discriminant Analysis (HLDA). [8]

When the identity of the speaker was recognized by using Urdu Utterances, it was seen that overall performance of MFCC with GMM gave 95.5% and MFCC with Vector Quantization (VQ) with Hidden Markov Model (HMM) gave 96.4% accuracy [18]. As the research goes on for accent classification of Telugu speech, it was deduced that the Deep Belief Network (DBN) based method was more accurate in speech recognition [17]. For accent identification of spoken English, it was seen that Central Neural Network (CNN) developed with two Convolutional/Pooling layers as the best model compared to three-layer CNN and one-layer CNN [7]. To identify the voices of 7 different speakers, MATLAB, the Voicebox tool was utilized which worked on Discrete Cosine Transform (DCT) to retrieve the MFCC coefficient with the CORDIC algorithm [3].

Various Malayalam accents were recognized in MATLAB, by the application of MFCC-GMM method with 89% accuracy [5]. Further, the hybridization of DWT-LPC (Discrete Wavelet Transform and Linear Predictive Coefficient) for Malaysian female speech recognition attained accuracy of 93.25% [20] A continuous speech recognition system for phonetic transcription was developed at AT&T Bell laboratories. Hidden Markov Model in conjunction with an appropriate dynamic programming algorithm was used to do the acoustic to phonetic mapping. The test was performed on DARPA data that has been filtered and down sampled to a 4 kHz bandwidth. In a few informal listing tests, the word intelligibility rate was judged to be approximately 75% [16].

Further research led to the development of a continuous speech recognition system for 100 million British English words using phonologically-constrained morphological analysis at the Department of Phonetic and Linguistic, University College London [9]. Coming towards the Korean Language, a continuous speech recognition system was developed using a phone-based Semi-Continuous Hidden Markov Model (SCHMM) method at the University of Korea. As a result of the speaker-independent experiment, the Discrete HMM (DHMM) method was applied which showed 89.7%word accuracy however, the SCHMM method showed 89.0%.

Reviewing all the above proceedings there are some flaws in the existing systems of speech recognition. One of the common errors in continuous speech recognition is missing out of minuscule gape between words. This happens when the speaker speaks at a high rate. As the accent of various countries is different so there is a fault in ASR and at listening end. Issues of the speed of utterance of words and noise in voice are dominant as the present ASR engine's proficiency is limited.

The existing Automatic Speech Recognition systems (ASR)

are vastly utilized and faced an urgent need for them to cover accented English as they form a large proportion of the population. There have been significant improvements in automatic voice recognition technology. However, existing systems still face difficulties, particularly when used by non-native speakers with accents. The new system will address the problem of identifying the English accented speech of speakers from different backgrounds. Once an accent is identified, the speech recognition software can utilize training set from an appropriate accent and therefore improve the efficiency and accuracy of the speech recognition system. Three different accents were considered and experimented for purpose of this project.

3. METHODOLOGY AND MATERIAL

3.1 Functional and Non-functional Requirements

The functional requirements refer to the services and functionalities which the software renders to the end-users. These are the explicit features of the novel system. For the present system these requirements came out to be; Real-time recording clips for training and testing phase, Modest user interface for effective evaluation of both train and test clip dataset, Domain centered user interface for efficient user familiarity.

Contrary, non-functional requisites are the features of the system which define the quality of the setup. These include; software security, data storage, data integrity, software performance, ease of use, disaster recovery and system accessibility, etc.

3.2 Comprehensive Model for the Accent recognition system

The high-level model for the designed system is given in Figure 1. This has three modules; Preprocessing/Feature Extraction Module, Parametric Features (Training/Testing Module), Model generation Module. This Figure shows the requirements for accent recognition of Nigerian tribal-based language.

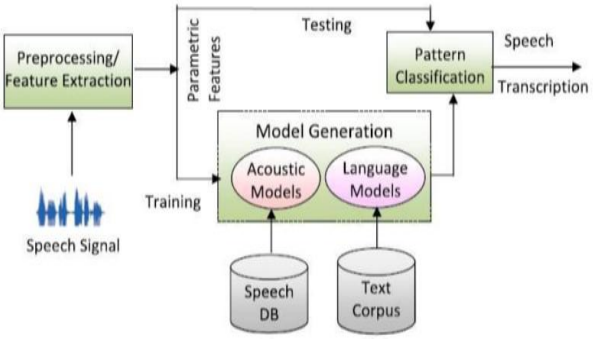

Figure 1: System Depiction of Accent Recognition System

3.3 Methodology

The methodology adopted provides complete measures for the accent classification model which follows the Knowledge Discovery Database (KDD) process. We shall explore machine learning classifier for accent classification, the classifier includes *k*-Nearest Neighbors (KNN), Gaussian Mixture Model (GMM), and Logistic Regression. The approach adopted shown in figure 2 illustrates the analysis of audio signals and extracting important features, these features are considered and respective frames are concatenated to form the input feature fed to the classifier. The results of the classifiers are then validated and compared using different accuracy metrics.

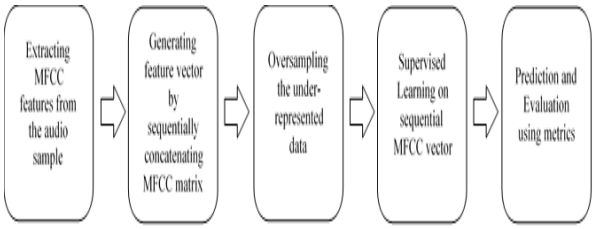

Figure 2: Block Diagram for Methodology description

3.4 Dataset Acquisition

The dataset which serves as research motivation was gathered during system development. It is a collection of audio recordings of Nigerians situated in Lagos state, Kogi state, and Kano state. The speakers, which are generally male, with Yoruba, Igbo, and Northerners accent were recorded uttering the same sentence `Education is Money, Money is Power` (2-5 seconds long). There are 150 audio samples in total which are divided into 80% training and 20% testing. This is shown in Table 1.

Table 1. Dataset Acquisition

Type of Accent	Gender	Number of Speakers	Total Number of Audio Samples
Yoruba	Male	10	50
Igbo	Male	10	50
Hausa	Male	10	50

3.5 Preprocessing and Feature Extraction

The first step to start with is the conversion of an audio clip into a .wav file with the help of M.studio. The next step of pre-processing is the extraction of noise and gap from the sound clip in Audacity as shown in Figure 3.

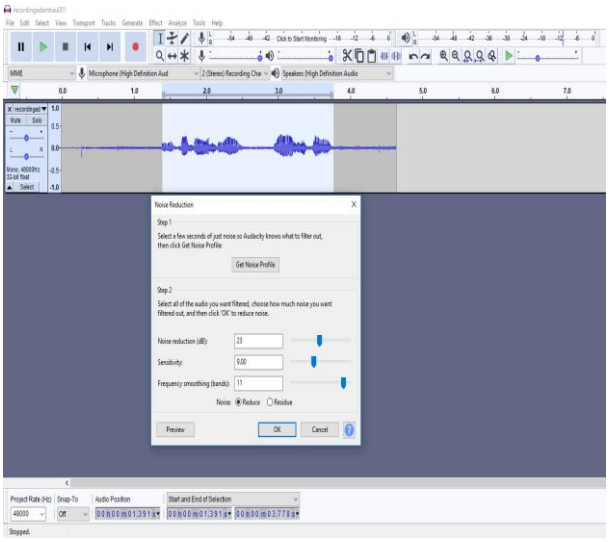

Figure 3: Noise reduction and space removal from an audio clip via Audacity

3.6 Feature Extraction through Python

20 Mel-frequency Frequency Cepstral Coefficient (MFCC) features are calculated for each audio file in the dataset. The MFCC is a tool to disintegrate the accent in various categories Feature extraction is implemented using the Python library 'Librosa'. The following steps are used to calculate the MFCC cepstral features

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

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

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

$$\boldsymbol{MFCC} = \begin{bmatrix} \boldsymbol{Cofo} & \boldsymbol{Cof1} & \boldsymbol{Cof2} \ldots & \boldsymbol{Cofm} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{C19fo} & \boldsymbol{C19f1} & \boldsymbol{C19f2} \ldots & \boldsymbol{C19fm} \end{bmatrix} (2)$$

where ci f0…ci fm is the values of coefficient i for frames 1…m.

the above matrix represents the MFCC coefficients for an audio sample with m frames. This 20 x m matrix needs to be transformed into a format that is recognized by the machine learning model. To contrast the three accents based on their feature sets, MFCC coefficients are sequentially concatenated. This retains enough information required to identify the accent from the feature set. The features are concatenated and flattened into a one-dimensional array of features for each frame. A sample vector of sequential MFCC features from a single audio sample is shown in equation 3.

$$\boldsymbol{MFCC} = [\boldsymbol{Cofo} \ldots \boldsymbol{Cofm} \ldots\ldots\ldots \boldsymbol{C19fo} \ldots \boldsymbol{C19fm}] \quad \textbf{(3)}$$

These features are then used to train the model to distinguish a particular accent from others.

The data mining methodology used in the research project follows the Knowledge Discovery Database (KDD) methodology. It's a structured method for analyzing, designing a system applying the machine learning concepts. It focuses on the collection of data, data preprocessing, data transformation, relevant data selection, and pattern evaluation.

4. RESULT

Place Tables/Figures/Images in text as close to the reference as possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7").

Captions should be Times New Roman 9-point bold. They should be numbered (e.g., "Table 1" or "Figure 2"), please note that the word for Table and Figure are spelled out. Figure's captions should be centered beneath the image or picture, and Table captions should be centered above the table body. System design and system implementation phase which is important in the development of the actual system is discussed in this section and then the effects of the designed system. The objective of developing the new system was the easiness of the users, attractive design interface, quality enhancement, and fulfillment of all the requirements at the user end.

4.1 System Design

The design of the system was built considering the optimization of the training module and architecture of neural networks at the input phase. The system was designed in ordered to match with the existing architecture of the system as shown in figure 4. The first phase of speech recognition is the training phase and then the testing phase. The graphical depiction of the system shows input and output entities of the system with TO and FRO dataflow. Feature extraction is the first phase where speaker aided information is extracted and stored in the speaker database as shown in figure 5.

In the test phase, features are extracted from unknown speaker's speech sample and the compared with the speaker database, this is termed as pattern matching which refers to the algorithm that computes a match score between the models stored in the database and the unknown speaker's features, as shown in figure 6.

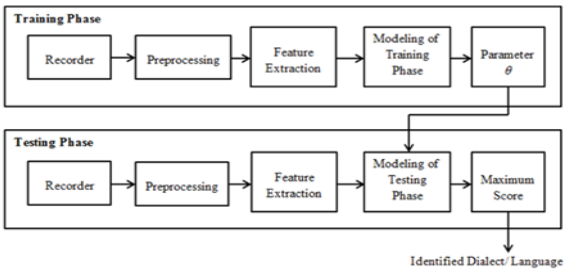

Figure 4: Architecture model of the system

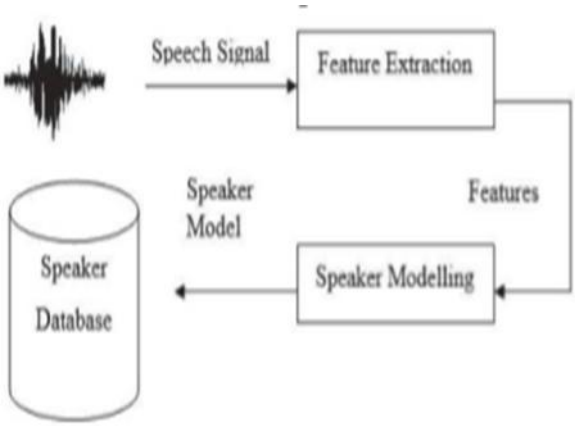

Figure 5: Training phase of the ASR

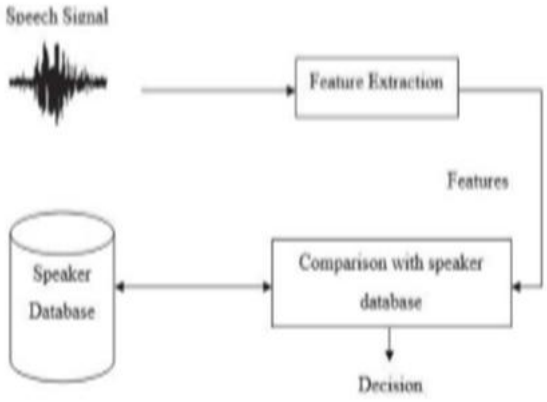

Figure 6: Test Phase of ASR

4.1.1 Input and Output of System

The system input is the audio clip and the extraction of features from the file for the training phase. The interface is shown in figure 7.

Figure 7: Input Data Classifier Interface for ASR

The output of the system is the classified data plotted against the true and false positive rate. These receiver operating characteristics and the confusion matrix of sample KNN are depicted in figure 8.

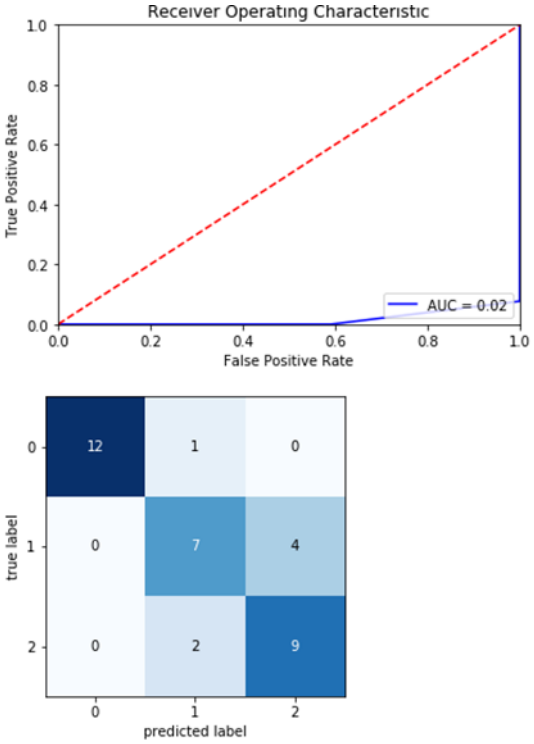

Figure 8: ROC and Confusion Matrix of KNN

4.1.2 Logical Design Model

The block diagram of the MFCC is represented in Figure 9 while the logical model extracted from the speakers of the distinguished accents all uttering the same sentence 'Education is money; money is power', and showing the variation in speech that helps the model predict the class of accents is shown in Figure 10.

4.1.2.1 Training Module

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

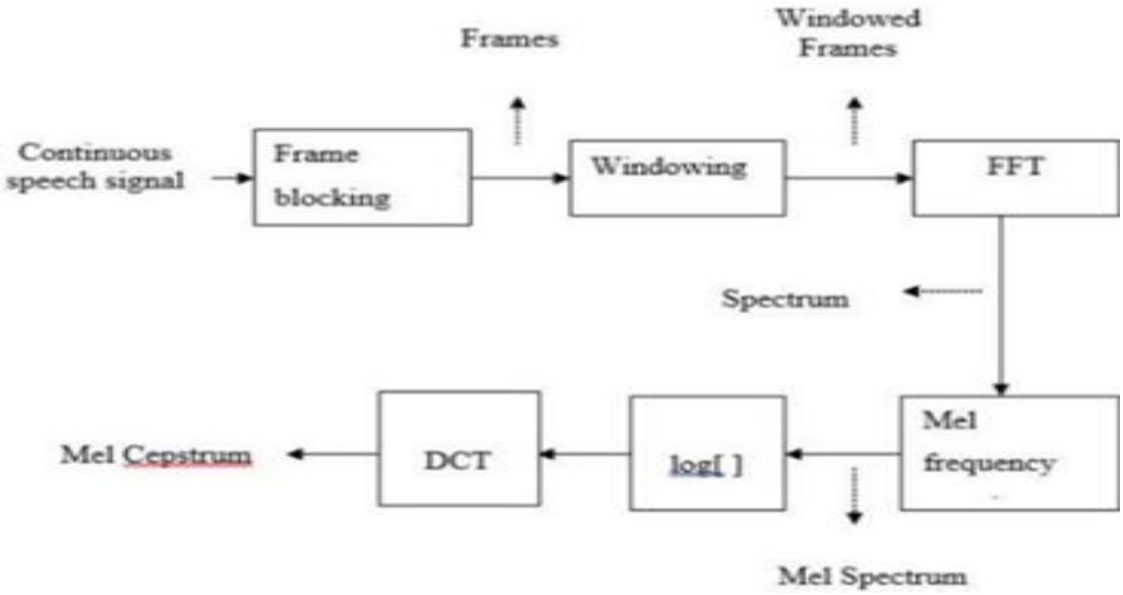

Figure 9: MFCC Block diagram

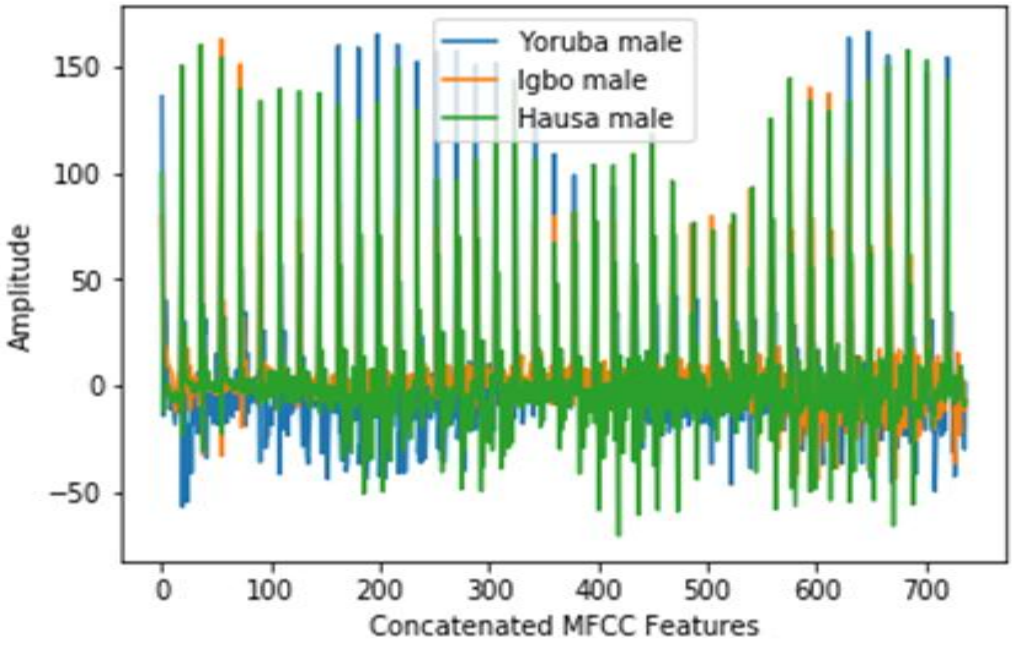

Figure 10: Concatenated Feature of MFC

4.1.2.2 Testing Module

This module validates the test sample by comparing its features with trained features and identify it as the accent that closely matches.

4.1.2.3 Database Design

Database design is a design that produces a simplified data model of the system database. This is given in Table 2.

Table 2. Dataset Distribution for the System

Dataset Class	Train Yoruba Igbo Hausa	Test Yoruba Igbo Hausa
#File	40 40 40	10 10 10
Id	P248 P294 P253	P248 P294 P253
Total duration (hour)	2 2 2	0.5 0.5 0.5

4.1.2.4 Validation Test

K-fold cross-validation is a type of evaluation metric that partition sample into 'k' partitions. These partitions are partitioned at random and are of equal sizes. Among these 'k' partitions, 1 of them is the validation or testing set and the rest serve as training data. The cross-validation process is repeated k times, such that all of the partitions are used as training data once. In general, 'k' remains an unfixed parameter, in our experimentation, k is taken to be as 10. The validation accuracies are tabulated in table 3.

Table 3. Validation for accuracy of audio clip

Model	Mean Validation Score	Standard deviation of Validation Score
KNN	0.61	0.17
Logistic Regression	0.67	0.22
GMM	0.27	0.11

4.1.2.5 Test Case Accuracy

The performance of a model is evaluated using performance metrics like precision, recall, reject rate, and overall accuracy.

The results presented (table 4) apply to the data which is split into training (80% of the data) and testing (20% of the data).

Table 4. Test Case Accuracy recordings

Model	Precision	Recall	f-measure	Reject Rate	Accuracy
KNN	0.71	0.83	0.76	0.86	0.75
Logistic Regression	1.0	0.21	0.35	0.90	0.82
GMM	0.42	1.0	0.58	0.22	0.5

4.1.2.6 Analysis of KNN, Logistic Regression and GMM

GMM classifies the majority of the test case samples of Yoruba '0' as Igbo '1' also Hausa '2' as Igbo '1' which contributes to the high recall but low precision. The overall accuracy, AUC, and K-fold cross-validation results shown in figure 11 (a) and (b) show that GMM does not work well for the task at hand. Coming towards Logistic Regression performance, it acts efficiently in separating the classes and predicting the classes with good accuracy. The problem with Logistic Regression is that it identifies most of the `Igbo` as `Hausa`. KNN Regression technique does not produce as much accuracy as Logistic Regression but it is far more accurate than GMM. This is depicted in figure 12.

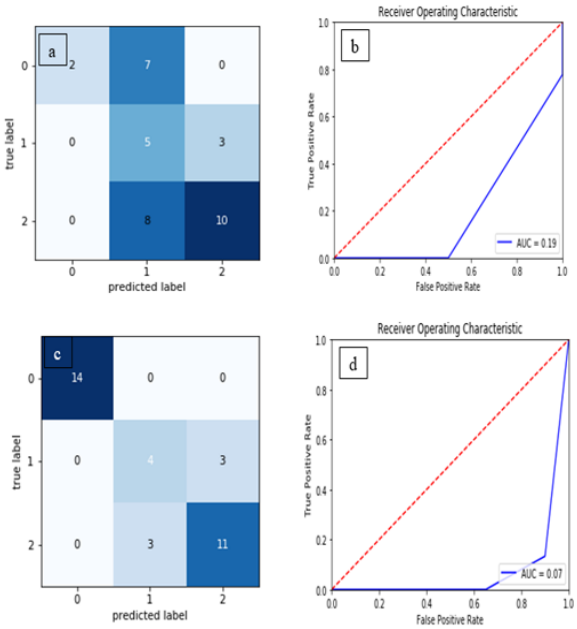

Figure 11: GMM and Logistic Regression Visualization

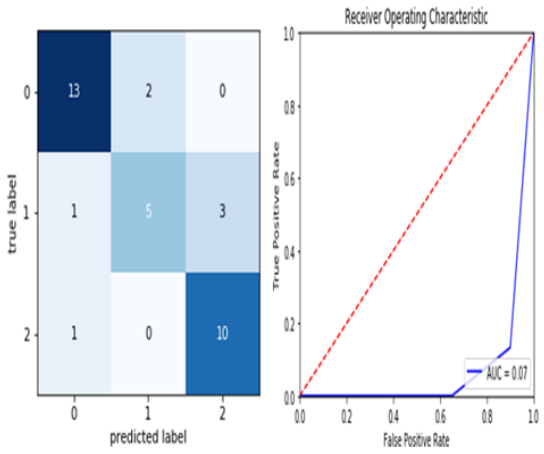

Figure 12: KNN Accuracy Visualization

5. CONCLUSION

Classifying Accents based on their acoustic features using machine learning techniques provides precise results using sequential MFCC features. Accent identification is a preprocessing step to speech recognition. This aids in more proficient speech recognition. In this system, we can solve the

problem of identifying accents 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%). This work can be extended to other accents and accuracy can also be increased in the future.

6. ACKNOWLEDGMENT

Our thanks to the individuals who have contributed towards development of this research

7. REFERENCES

[1] Abdulwahab, A., MohdYusof, S., & Husni, H. (2017). Acoustic comparison of Malaysia and Nigeria English accents. Journal of Telecommunication, Electronic and Computer Engineering, 141-146.

[2] Ahn, E. (2016). A Computational Approach to Foreign Accent Classification.

[3] Andelman, M. (2011). Flow through capacitor basics. Separation and purification technology, 80(2), 262-269.

[4] Angkititrakul, P., & Hansen, J. H. (2006). Advances in phone-based modeling for automatic accent classification. IEEE Transactions on Audio, Speech, and Language Processing, 14(2), 634-646.

[5] Aswathi Sanal, M. N. G. (2017). Accent Recognition for Malayalam Speech Signals. International Journal of Innovative Research in Computer and Communication Engineering.

[6] Bryant, M., Chow, A., & Li, S. (2014). Classification of Accents of English Speakers by Native Language: Stanford University [cited 14 Dec. 2018]. Available from World Wide Web ….

[7] Chionh, K., Song, M., & Yin, Y. (2018). Application of Convolutional Neural Networks in Accent Identification. Project Report, Carnegie Mellon University, Pittsburgh, Pennsylvania.

[8] Ge, Z., Tan, Y., & Ganapathiraju, A. (2015). Accent classification with phonetic vowel representation. Paper presented at the 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR).

[9] Huckvale, M., & Fang, A. C. (2002). Using phonologically-constrained morphological analysis in continuous speech recognition. Computer Speech & Language, 16(2), 165-181.

[10] Hung, P. N., Van Loan, T., & Quang, N. H. (2016). AUTOMATIC IDENTIFICATION OF VIETNAMESE DIALECTS. Journal of Computer Science and Cybernetics, 32(1), 19-30.

[11] Jain, A., Upreti, M., & Jyothi, P. (2018). Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning. Paper presented at the Interspeech.

[12] Krishna, G. R., & Krishnan, R. (2014). Native language identification based on english accent. Paper presented at the International Conference on Natural Language Processing (ICON).

[13] Kumar, A. P., Roy, R., Rawat, S., & Sudhakaran, P. (2017). Continuous Telugu Speech Recognition through Combined Feature Extraction by MFCC and DWPD Using HMM based DNN Techniques. International Journal of Pure and Applied Mathematics, 114(11), 187-197.

[14] Kumpf, K., & King, R. W. (1996). Automatic accent classification of foreign accented Australian English speech. Paper presented at the Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96.

[15] Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., & Shrishrimal, P. P. (2014). Continuous Speech Recognition System: A Review. Asian Journal of Computer Science and Information Technology, 4(6), 62-66.

[16] Levinson, S. E., Ljolje, A., & Miller, L. G. (1990). Continuous speech recognition from a phonetic transcription. Paper presented at the International Conference on Acoustics, Speech, and Signal Processing.

[17] Mannepalli, K., Sastry, P. N., & Suman, M. (2016). MFCC-GMM based accent recognition system for Telugu speech signals. International Journal of Speech Technology, 19(1), 87-93.

[18] Riyaz, S., Bhavani, B. L., & Kumar, S. V. P. Automatic Speaker Recognition System in Urdu using MFCC & HMM. International Journal of Recent Technology and Engineering (IJRTE), 7.

[19] Stanford. (2010). Machine Learrning from csc229.stanford.edu.

[20] Yusnita, M. A., Paulraj, M., Yaacob, S., & Shahriman, A. B. (2012). Classification of speaker accent using hybrid DWT-LPC features and K-nearest neighbors in ethnically diverse Malaysian English. Paper presented at the 2012 International Symposium on Computer Applications and Industrial Electronics (ISCAIE).