Indigenuous Vocabulary Reformulation for Continuousyorùbá Speech Recognition In M-Commerce Using Acoustic Nudging-Based Gaussian Mixture Model

Kehinde Lydia Ajayi (lydia4reel@gmail.com)
Covenant University https://orcid.org/0000-0003-1544-2939
Victor Azeta
National Productivity center, Kaduna
Isaac Odun-Ayo
Covenant University
Ambrose Azeta
Covenant University
Ajayi Peter Taiwo
Joesph Ayo Babalola University
Felix Chidozie
Covenant University

Research

Keywords: Acoustic Nudging Model, Gaussian Mixture Model, Automatic Speech Recognition, Word Error Rate, Nudging, ASR Error, Acoustic Irrational behavior, Behavioral Economies.

Posted Date: February 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-211622/v1

Indigenuous Vocabulary Reformulation For Continuousyorùbá Speech Recognition In M-Commerce Using Acoustic Nudging-Based Gaussian Mixture Model

Lydia Kehinde Ajayi[1], Ambrose Azeta[2], Isaac Odun-Ayo[3], Victor Azeta[4], Ajayi Peter Taiwo[5], Felix Chidozie[6]

Corresponding Author

Correspondence to lydia4reel@gmail.com

Abstract

One of the current research areas is speech recognition by aiding in the recognition of speech signals through computer applications**.** In this research paper, Acoustic Nudging, (AN) Model is used in re-formulating the persistence automatic speech recognition (ASR) errors that involves user's acoustic irrational behavior which alters speech recognition accuracy. GMM helped in addressing low-resourced attribute of Yorùbá language to achieve better accuracy and system performance. From the simulated results given, it is observed that proposed Acoustic Nudging-based Gaussian Mixture Model (ANGM) improves accuracy and system performance which is evaluated based on Word Recognition Rate (WRR) and Word Error Rate (WER)given by validation accuracy, testing accuracy, and training accuracy. The evaluation results for the mean WRR accuracy achieved for the ANGM model is 95.277% and the mean Word Error Rate (WER) is 4.723%when compared to existing models. This approach thereby reduce error rate by 1.1%, 0.5%, 0.8%, 0.3%, and 1.4% when compared with other models. Therefore this work was able to discover a foundation for advancing current understanding of under-resourced languages and at the same time, development of accurate and precise model for speech recognition.

Keywords: *Acoustic Nudging Model*, *Gaussian Mixture Model*, *Automatic Speech Recognition*, *Word Error Rate*, *Nudging*, ASR Error, *Acoustic Irrational behavior, Behavioral Economies.*

Introduction

Speech is one of the most convenient means of communication betweenpeople and the primary key need by human[1] [2][3] [4]. Human-computer interface is the communication used for human-computer interaction [5]. Human computer interaction is based on the syntactic arrangement of lexical names drawn from huge vocabularies [3]. There are three different spoken languages in Nigeria which are Yorùbá, Igbo, and Hausa with their accents, but the focus of this study is on Standard Yorùbá. Yorùbá language is the second largest ethnic group in Nigeria spoken by more than 50 million people with 30 million speaking Standard Yorùbá [5] [6]. Yoruba is one of the 12 languages of the Edekiri branch from the family of West Benue-Congo. Yorùbá is a native language spoken by Nigeria, Togo, Ghana, Sudan, Cote D'Ivoire, Sierra-Leone, and Benin. Standard Yorùbá is also spoken beyond Africa in countries like Cuba, Brazil, Trinidad and Tobago where a large number of this language can be located[7] [8].Yorùbá also loans different words from Arabic, Hausa, Igbo, and English [9].

Yorùbá language is a language that falls under the low-resourced languages as it doesn't have a massive amount of speech and text information for speech recognition. Speech processing for under-resourced languages is a current field of research that has experience a significant progress, but the focus of this work is on Yorùbá, which is under-resourced. Under-resourced language is defined as a language that lacks a unique writing system and a limited

presence on the Web. They are also called low-density, low-data, low-resourced, or resource-poor languages [10].Several efforts have been made over the years to develop vocally interactive computers to realize voice to speech synthesis which has been of great benefits [2]. This speech processing are either Automatic Speech Recognition also called speech-to-text (ASR or STT), text-to-speech Synthesis (TTS), and speech Coding. There are different classification of automatic speech recognition which are based on utterances which are broken down into isolated, continuous, spontaneous and connected but the focus of this study is on Continuous speech recognition.

Background and Related Work

ASR. also known as speech-to-text converts a speech signal into textual information i.e., a sequence of spoken words using an algorithm that is implemented by a hardware or software module into a text data [10] [11]. [12] designed an HMM-based Tamil Speech Recognition is based on limited Tamil words exhibiting low recognition accuracy. [13] developed a GMM based isolated speech recognition using MATLAB, where he designed a speaker-dependent speech recognition system based only on isolated digits of 0-9 and gives an accuracy ratio greater than 70%. [14] developed a speech-to-text converter using GMM where the paper focused on the extraction of features of speech signal by MFCC for multiple isolated words to train audio files in order to get spoken words recognized and gave an accuracy of 98%, but the limitation of the system is based on isolated words.

[15] developed a speaker-independent continuous Amazigh language using CMUSphinx tools. The accuracy recognition percentage achieved is 90.5%. [16] also proposed a Markov Model-based Oriya isolated speech recognizer for visually impaired students in school and public examination. The study focused on problems with visually impaired learners of Orissa in schools and pubic examinations as their assessment procedure is not suitable for students. This was done using 1800 isolated answers Oriya words, collected from 30 different speakers in training stage and also, the testing stage is carried out by five ’5’ speakers. The word accuracy yielded 76.23% for seen data and 58.86% on unseen data.

Automatic speech recognition for Tunisian dialect, which is also an under-resourced language was developed in[17], where HMM-GMM model using MFCC, and HMM-GMM with LDA were compared. These approaches gave a WER of 48.8%, 48.7%. [17] proposed an enhanced ASR system for Arabic (MSA) using Kaldi Toolkit to build the system around, after which the acoustic models was trained using HMM-GMM model and the data collected was based on Standard Arabic news broadcasts. The language model was trained with two Corpora, which are GigaWord3 Arabic corpus (1000 words), and the acoustic training data transcription (315000 words). It gave a WER OF 14.42%.

[3] provided a Mobile Tourist Assistance for Yorùbá Language developed for tourists and implemented on Android application. He developed speech-to-text system for easy communication between locals and tourist. The text data were gathered from on-site interaction with the native speakers in four different domains: Market, Hospital, Motor-Park, and Restaurant. The recording of the Yorùbá phone-set was done using Praat software through a male voice. The recording of the Yorùbá phone-set was done using Praat software through a male voice. The accuracy of the system is established to be 85% for clarity and 88% for naturalness. The limitation of this system is it is a one-way interactive system, which is based on isolated words. [18] also developed a Home Automation speech-to-text system

that lets a user control computer functions and dictates text by voice using HMM as acoustic modeling, MFCC for feature extraction, VQ for feature training of the dataset.

[19] developed a standard Yorùbá Isolated speech-to-text system using HTK, which has the ability to recognize isolated words spoken by users using previous data. The system adopted syllable-based approach using six '6' native speakers speaking 25 bi-syllabic and 25 tri-syllabic words under an acoustically-controlled room based on HMM and MFCC. The overall accuracy recognition ratio gave 76% and 84% for bi-syllabic and tri-syllabic words.[21] proposed a continuous Fongbe ASR system, an African language spoken in Benin, Togo, and a minimal part of Nigeria, but the system exhibited low accuracy of 71.07%, no inclusion of tone diacritization. All the existing studies also have no consideration for automatic correction of ASR errors involving user's acoustic irrational behavior and also, existing studies doesn't consider tone diacritization as it is needed especially in Yoruba to avoid ambiguity. In order to improve speech recognition accuracy and system performance, Acoustic Nudging based Gaussian Mixture Model (ANGMM) is proposed in this paper. This research process method reduces word error rate and sentence error rate at a very significant level when compared to other existing models.

The paper is organized as follows: The proposed ANGM model for large vocabulary-continuous standard Yoruba speech recognition (CSYSTT) is described in section 3. Section 4 illustrates the performance evaluation of the existing models and the proposed model. Section 5 concludes this paper.

Methodology

This work focused on large vocabulary speaker independent tone-diacritized continuous standard Yorùbá speech recognition, which is the recognition of continuous spoken words with a comparatively high number of different words. The continuous word recognition system for standard Yorùbá based on ANGM model is developed and designed in three stages;

- System training with standard Yorùbá speech samples.
- System validation with standard Yorùbá speech samples.
- System testing with standard Yorùbá test samples.

The training phase of the CSYSTT system consists of building and learning the acoustic model, which is regarded as the major component of any automatic speech recognition "ASR" or speech-to-text "STT" engines, language model together with the pronunciation dictionary [6]. The training and recognition of parameters is based on an ANGM speaker-independent tone-diacritized continuous standard Yorùbá speech recognition system capable of handling large vocabularies. The approach for modeling standard Yorùbá sounds consists of generated/trained acoustic models and language models for standard Yorùbá speech data. The following steps were engaged for the development of CSYSTT (see Figure 1).

Data Collection Requirement

The development of the automatic speaker-independent large vocabulary continuous Standard Yorùbá speech recognition (CSYSTT) is made from a large amount of data that contains both the speech signals (for acoustic modeling) and text data (for language modeling). This stage describes the methodology of how texts and audio

signals of standard Yorùbá language are collected for designing the CSYSTT system. The requirement for data collection for speech database development is done through several standard Yorùbá speakers, and these speakers are required to record their voices by uttering and pronouncing the standard Yorùbá words and sentences in a pure voice with small noise in the background area.Also, the requirement for the text corpus development is done after the speech data file collection which is the textual transcription (see Figure 1).

- **Tone Diacritization**

Before developing any speech or text corpus, Tone diacritization is needed in any Yorùbá words to achieve high accuracy in speech recognition and reduce ambiguity. One un-diacritized texts can have different pronunciation (ile and ile - house or land). Therefore, having a Yorùbá text without diacritics leads to different pronunciation forms. In speech recognition, there are two different forms of textual training in speech recognition, which can either be diacritized text or non-diacrtized text. This stage, the Standard Yorùbá diacritization process includes marking the Yorùbá letters using the orthographic symbols called diacritics or tone marks. Using non-diacritized texts poses a challenge to this Yorùbá automatic speech recognition as missing tone/short vowels leads to some confusion in the learning and training process. Identification of syllables with the appropriate tone helps to determine the stress and intonation of that particular word (See Figure 1).

Using non-diacritized texts poses a challenge to this Yorùbá automatic speech recognition as missing tone/short vowels leads to some confusion in the learning and training process. Identification of syllables with the appropriate tone helps to determine the stress and intonation of that particular word (See Figure 1).

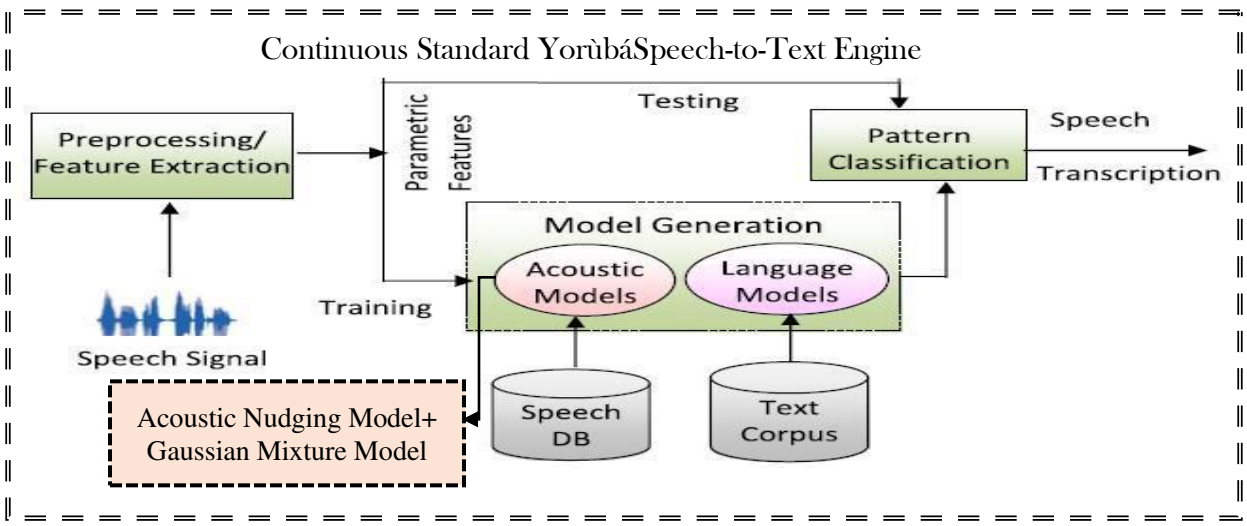

Figure 1: The Proposed Acoustic Nudging-Based Gaussian Mixture Model

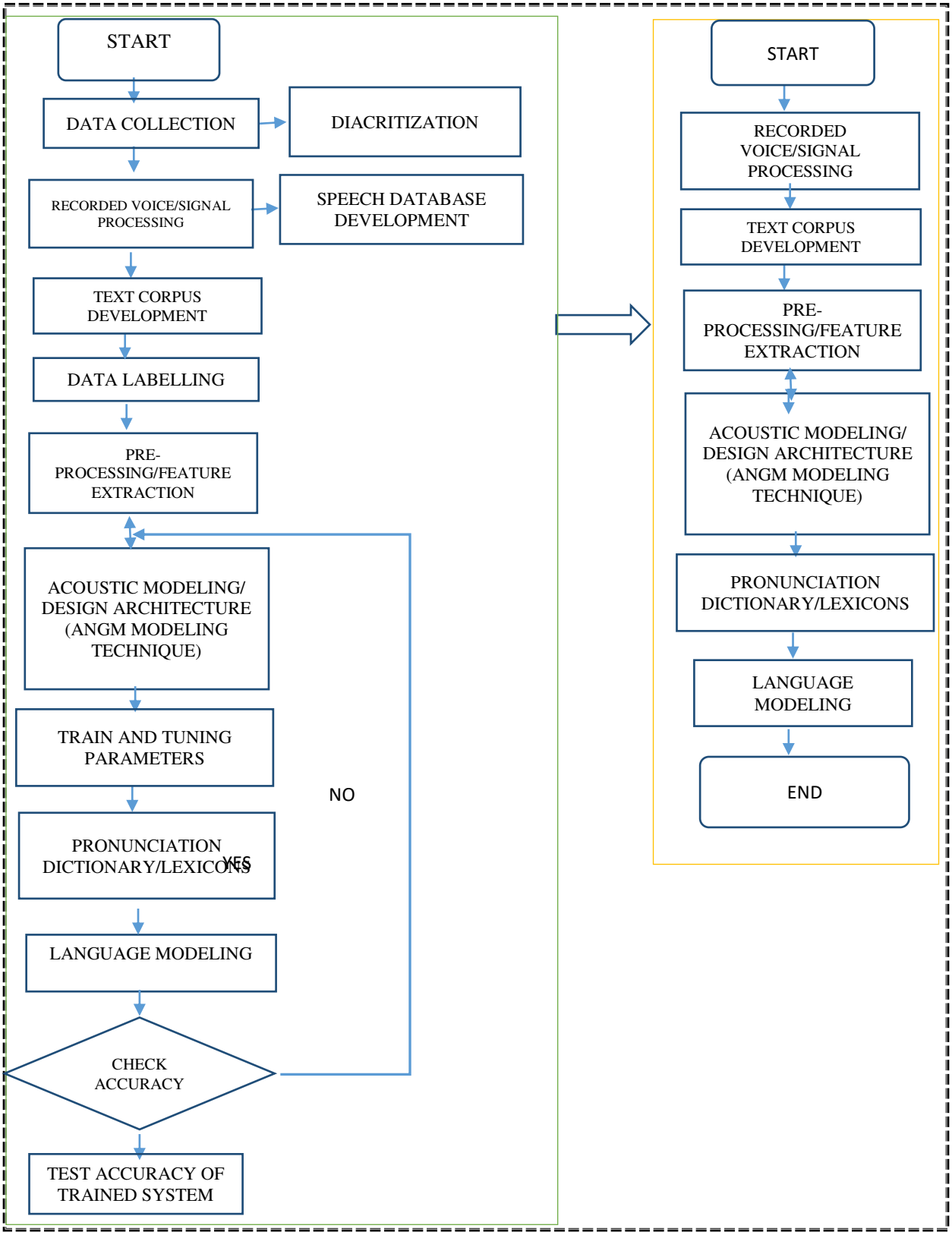
START
DATA COLLECTION
DIACRITIZATION
RECORDED VOICE/SIGNAL PROCESSING
SPEECH DATABASE DEVELOPMENT
TEXT CORPUS DEVELOPMENT
DATA LABELLING
PRE-PROCESSING/FEATURE EXTRACTION
ACOUSTIC MODELING/ DESIGN ARCHITECTURE (ANGM MODELING TECHNIQUE)
TRAIN AND TUNING PARAMETERS
NO
PRONUNCIATION DICTIONARY/LEXICONS
YES
LANGUAGE MODELING
CHECK ACCURACY
TEST ACCURACY OF TRAINED SYSTEM
START
RECORDED VOICE/SIGNAL PROCESSING
TEXT CORPUS DEVELOPMENT
PRE-PROCESSING/FEATURE EXTRACTION
ACOUSTIC MODELING/ DESIGN ARCHITECTURE (ANGM MODELING TECHNIQUE)
PRONUNCIATION DICTIONARY/LEXICONS
LANGUAGE MODELING
END

Figure 2: The Methodology Process Flowchart of the Continuous Speech-to-Text (CSYSTT) Model

Recorded Voice/Signal Processing (Speech Database Development)

The speech corpus was created through data collection by training, validating, and testing of Yorùbá speech samples. The collection of data to develop the speech corpus is based on verylarge vocabulary continuous speech recognition. Very large vocabulary generally means that the system will have a vocabulary of roughly 20,000 to 60,000 words.

The dataset was collected using 47 different Yorùbá speakers making a total of 22,312 data points, and sums up to 2880.8 hours as each speaker utters different standard Yorùbá words and sentences (single word utterances, multiple word utterances, long sentences, etc.) one after the other in different record notes. The data is further divided into training, validation, and testing dataset (See Table 1). Also, for the acoustic nudging model process, 16 speech data was collected based on normal, angry, panicked, sorethroat and stressed acoustic behavior from different age brackets for males and females. 8 normal/neutral speech samples and 8 acoustic irrational speech samples. The speakers are from different states in Nigeria, consisting of female and male speakers with a large variety of ages, i.e. both children and adults. The dataset was gatheredfrom different social apps like Facebook,WhatsApp, Messenger, Twitter, as the speakers recorded their voices in short mp3 files (See Figure 1).

Table1: Speech Corpus Statistics

	No of speakers	**No of Files**	**No of Sentences**	**No of words**
Train	22	17,850	6392	11458
Validation	10	2,162	851	1311
Test	15	2,300	918	1382
Total	**47**	**22312**	**8161**	**14151**

Table2: System Parameters

Speaking Mode	Continuous words
Sampling Rate	44kHz
Training/Enrolment	Speaker-Independent
Vocabulary Size	Large
Equipment	Smart Voice Recorder Application (QuickRec), Woefzela,Social Media Application
Number of Channels	1, Mono
Audio Data File Format	.wav
Speech Acquisition	2-Level (Word Level and Sentence Level)
Speech Corpus	22,312 words and sentences
Number of Speakers	47

As the data was gathered through different location, each locations consisted of varying noises. Some voice samples embedded low/minimal noise (room recordings), while some were affected by outdoor noises (by air-conditioner, children playing, car horning, dog barking, siren, street music, low music playing, people talking, etc.). Therefore to create a high-end accurate and robust voice system, noise and silence removal was required as there are varying voice samples based on these outdoor noises. Therefore, it was important to perform pre-processing/feature extraction on each signal separately. The voice samples were given as a batch of different voice notes by each speaker, which is ultimately necessary for each sample to be noise-free (See Figure 1).

Text Corpus Development

In addition to the speech database, there is the resultant text corpus, which involves developing the text corpus. The collection of proper and suitable text is significant for the development of a text corpus. For the standard Yorùbá text corpus, the text collection was mainly done through the utilization of newspaper articles, Yorùbá online books, newspapers, texts embedded on the m-commerce platform and online Yorùbá text dictionary. The next process was diacritizing each text using the Yorùbá tonal marks with the aid of semi-supervised diacritizer. The full vocalization of Yorùbá script is provided by the insertion of diacritics in the text. The approach employed in this process is a semi-automatic diacritization approach using Online Yorùbá Tone Marker (OYTM). The step process includes;

- Entering Standard Yorùbá text, one paragraph at a time.
- Tone mark each paragraph one syllable at a time by choosing syllable option using the correct tone mark either do, re, or mi with the tone under-marking for short vowel.

The whole process is given in Figure 3.

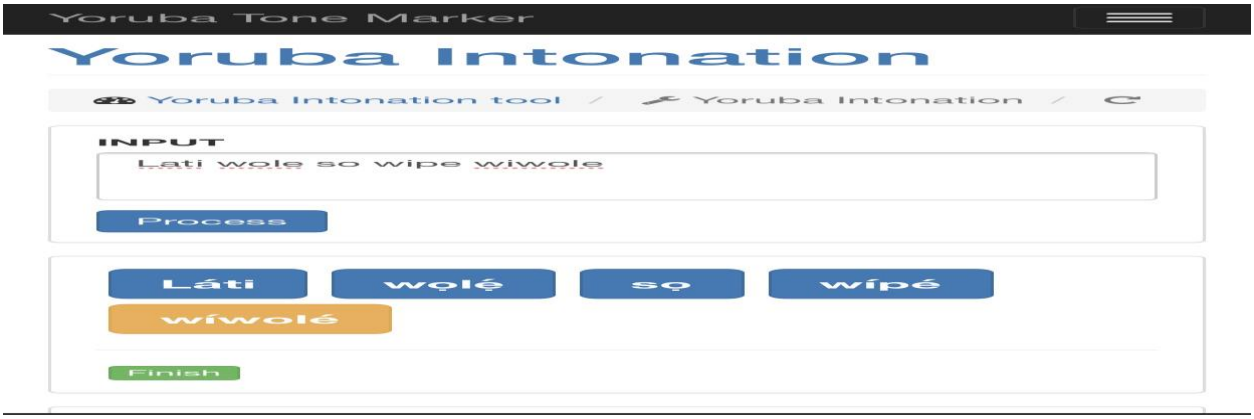

Figure 3: A Semi-Supervised Standard Yorùbá Tone Diacritization Process

Table 3: A Standard Yorùbá Continuous Speech Sample Dataset and its English Version

S/N	English Version	Standard Yorùbá Version
1	Welcome, please put in your email and password	Káàbọ̀ jọ̀wọ́ fí ímeèlì àti ọ̀rọ̀ ìgbaniwólé rẹ
2	I want to buy men's watches	Mo fẹ́ rà aago àwon okúnrin

3	I want to buy women's shoes	Mo fẹ rà bàtàawọn obìnrin
4	Diving Watch	Ìluwẹ̀ẹ́ aago
5	Pay now by saying pay now	Sánwó ní báyìí nípa sísọ sánwó ní báyìí
6	men's bag	baagì àwọn okúnrin
7	women watch	aago àwọn obìnrin

Data Labeling

It is highly essential that all the data collected are in proper order. After all, the speech samples were processed, it was required for the speech samples to be divided and labeled, which is done by labeling and exporting multiple speech samples from one sample using the corresponding text transcription and then, the speaker's name was added to each speech sample. Each speech sample was then divided so that each word is in one separate file only (See Figure 1).

Pre-Processing/Feature Extraction

As the data were gathered through different locations, each location consisted of varying noises. Some voice samples embedded low/minimal noise (room recordings), while some were affected by outdoor noises (by air-conditioner, children playing, car horning, dog barking, siren, street music, low music playing, people talking, etc.). This is referred to as environmental variations, and in order to minimize this problem, it is essential to create a high-end, accurate, and robust voice system where noise and silence removal are required as there are varying voice samples based on these outdoor noises. Therefore, it was important to perform pre-processing/feature extraction on each signal separately. The voice samples were given as a batch of different voice notes by each speaker, which is ultimately necessary for each sample to be noise-free. It involves compressing all the speech signal samples to a vector that makes meaningful information.

Before developing and training the model, feature extraction of audio files is required to eliminate unwanted signals such as noise and silence. This stage includes noise removal and silence removal. The silence removal is to eliminate unvoiced and silent art of the speech signal. This stage explains how the file format of all the separated words/sentences with 22,312 speech files is converted into a wav format to execute the feature extraction algorithm. This study makes use of Mel-frequency cepstral coefficient (MFCC).MFCC in speech research has been known as one feature extraction algorithms that works well with continuous speech recognition tasks for finding features compared to other feature extraction algorithms (Galvan et al., 2016). It reduces vulnerability to noise disturbance. They are created by taking the spectrum of a spectrogram ('a cepstrum') and discarding some of the higher frequencies that are less significant to the human ear [11].

Design Architecture (Acoustic Reformulation and Modeling)

In this study, the statistical representation adopted is Acoustic Nudging (AN) model and Gaussian Mixture model (GMM).The ANGM (GMM + AN) Model statistically represents the relationships between the speech signals and the language phonemes.

There are different methods and techniques to train acoustic models but for this study which is based on continuous automatic speech recognition for standard Yorùbá, Gaussian Mixture Model (GMM), and Acoustic Nudging (AN)

Model will be utilized. These models are combined together to develop the acoustic representation model of this study. This proposed model (ANGM) helps in automatically reformulating user's acoustic irrational behavior, by correcting ASR error, so as to minimize the error rate embedded in speech data and at the same time, address the issue of low-resourced attributes of Yoruba Language.

Acoustic Nudging Model

The entire process that makes up the acoustic nudging model is adapted from this study [23]. The acoustic nudging model is needed for automatic reformulation of user's acoustic irrational behavior which involves tracking/monitoring, detecting and reformulation of user's acoustic behavior in real time.

Gaussian Mixture Models (GMM)

Research has recorded GMM to be a tremendous success in speech domain due to their high word detection accuracy ratio for language with low training data. GMM has been applied in many domain like artificial intelligence, image recognition, phoneme classification, etc. one of the powerful attributes of GMM is its ability to form smooth approximations to arbitrarily shaped densities [24]. The Gaussian mixture density model for speech is assumes that an M component mixture model with component weights $P(w_m)$ and its parameters $\boldsymbol{\theta}_m$ represents the shape of each spectral.A univariate mixture model is represented in [24].The developed ANGM modeling technique is to be trained more than a trial to optimize the final training parameters. The accuracy of the model and loss is checked when tuning the parameters for optimization purpose. The ANGM model is a universal approach that can deal with both discrete and continuous data (See Figure 1).

Pronunciation Dictionary

The Pronunciation dictionary (PD) is also known as the lexicon. It contains all the 37,536 words with the sentences when broken down into a single word followed by their pronunciation called the phonetic transcription based on Standard Yorùbá language. The pronunciation dictionary is created after a deep study of Standard Yorùbá phonetics and also, different rules are used in pronouncing the words. Multiple entries are entered for a word which are diacritized due to their homonyms characteristics e.g. words with the same spelling but different pronunciation. Table 4 presents the phonetic dictionary list of some words used in training the system. This pronunciation dictionary act as an intermediary between the acoustic model and the language model to achieve a good recognition result (See Figure 1).

Table 4: The Phonetic Dictionary used in the Training

Aago	a a go	Ìgbaniwólé	e gba ni wo le
Àbájáde	a ba ja de		
àdírẹ́sì	a di re sii		
àkọ́ọ́lẹ̀	a koo lee	look for the same spellings and different meanings	
àmì	a mi		
àwọ	a woo	.	
àwọn	a wo n	.	
àsàyàn	a sa yan	.	
baagì	ba gi	.	
báyìí	ba yi		
bẹ́ẹ̀ bẹ́ẹ̀	bee be		
bèèrè	beee re	Yan – ya a un	

èka	e ka

System Testing Phase

After the CSYSTT training, it is required to perform output analysis which is done by verification of the output known as validation. The trained CSYSTT system is then applied to test Standard Yorùbá Continuous speeches to estimate the accuracy of the system. The testing of the trained CSYSTT is in two ways. The test data is with a total of 2300 words/sentences using 15 Standard Yorùbá speakers. The acoustic nudging modeling technique is applied to the test data and voices are unlabeled to calculate the accuracy of the CSYSTT model then calculate the accuracy of different continuous speech accumulated in the 22,312 words/sentences. Having a well-laid structure of the CSYSTT primarily affects the output. If the system is not well-structured and organized, it then leads to an inaccurate system. The accuracy of the testing/decoding phase is the percentage of the word error rate and the sentence error rate (See Figure 1).

PSEUDOCODE FOR THE ANGMMODEL

Algorithm Development

The Acoustic Nudging-Based Gaussian Mixture (ANGM) Model serves as the model for designing the continuous Standard Yorùbá speech recognition (CSYSTT) in m-commerce context. The algorithm for the ANGM model gives the sequence of activities used in designing the m-commerce system. The Acoustic Nudging algorithm is a reformulation algorithm for the user's acoustic irrational behavior to detect and correct ASR error which was modified and adapted from improved digital nudging. These are all presented in Figure 4.

1: ***Begin***
2: ***Generate*** the user's corrected acoustic rational behavior$-(Xm - Xp(5))$
3: ***Input*:** Five Heuristics and biases P_1, P_2, P_3, P_4 and P_5 randomly each having different Q variation values.
4: ***for***e ≥ 1 **do**
5: ***for****i*=1 to number of experiment
6: e*l* **do**
7: Evaluate the desired value ($P_{(5)}$) of experiment *i*
8: ***end for***
9: ***for****P* = 1 to number of variation sliderpoints Q **do**
10: Q_1= P, Q_2= L, Q_3= At, Q_4= Dt, and Q_5= T_bW
11: Evaluate effects of the heuristics and biases P_1, P_2, P_3, P_4 and P_5
12: ***for*** P Values= 1248Hz ≤ 1355Hz ***do***
13: ***for*** L= Gain of -50dB ≤ 48dB ***do***
14: ***for*** At = 0.12s≤ -0.06s ***do***
15: ***for*** Dt = 0.11s ≤ -0.05s ***do***
16: ***for*** T_bW= Gain of 0.12s ≤ 0.10s ***do***
17: ***for*** At = 0.12s ≤ -0.06s ***do***
18: ***end for***
19: ***Formulate***: draw $(-(Xm - Xp(5)))$ independently for every *i = 1…… N*
20: ***Acoustic Nudging***: Choose a set of model field (heuristics and biases) $P_{(5)}$ C [N], then compute $(-(Xm - Xp(5)) = (\bar{X}m + X'm) - (C + X'p(5)))$ for every $I \in P_{(5)}$, where $(-(Xm - Xp(5)) = (X'm - X'p(5)) - (\bar{X}m - \bar{X}p(5)))$ for every i $\in [N] \setminus P_{(5)}$
21: ***Re-formulate***: draw $(-(X'm - X'p(5)) = -(Xm - X\acute{p}(5)))$ for every
22: $(X\acute{p}(5) = Xp(5) - \bar{X}p(5) + \bar{X}m)$ independently for *i = 1…… N*
23**:** ***end for***

24: Generation of the Acoustic Nudging (AN) parameters $(-(Xm - X\acute{p}(5)))$

25: ***Initialize:*** Acoustic Nudging (AN) parameters $(-(Xm - X\acute{p}(5)))$ into $p(x)|w_m, \boldsymbol{\theta}_m = \frac{1}{\sqrt{2\pi}\sigma^2{}_m} exp\left[-\frac{(x-\mu_m)^2}{2\sigma^2{}_m}\right]$and set μ = 1 and σ = 1

26: While Converged ***do***

27: Set the step-Size Schedule μ_1 and σ_1, appropriately

28: ***Repeat***

29: Sample Yorùbá speech data-point λ_1 uniformly from the Yorùbá speech corpus and the resultant English Corpus

30: ***Compute:*** The ANGM Parameter μ_2 and σ_2

31: Update: The Current extimate of the ANGM parameters μ_2 and σ_2

32: ***end while***

33: Generation of the ANGM Parameter (μ_2 and σ_2)

34***: end for***

35: ***end***

Figure 4: The ANGM Algorithm

Variables used in the ANGM Algorithm

X: User's acoustic rational behavior
m: Acoustic Nudging Model predicted values
P (5): Acoustic Nudging Model prescribed values for the heuristics and biases
$\acute{p}(5)$: Replaced Heuristics and Biases.
S: User's speech signal
N: Number of speech signals
X^1: User's acoustic irrational behavior
i: Number of experiment
P: Pitch
L: Loudness or Sound Pressure
At: Timbre Ascend Time
Dt: Timbre Descend Time
T_bW: Time between each words
μ_1:The first GMM parameter "mean"
σ_1: The first GMM parameter "standard deviation"
μ_2:The ANGM parameter "mean"
σ_2: The ANGM parameter "standard deviation"
Q_1 - Q_5: Variation Slider Endpoints

Line 1-24 presents the Acoustic Nudging algorithm for reformulating user's acoustic irrational behavior to detect and correct ASR error whose details are contained in Section VI-A of 3.5.1. Line 25-35 presents the Gaussian Mixture Model (GMM) to address low resourced attribute of Yorùbá language whose details are contained in Section VI-B of 3.5.1.

Results

In achieving the CSYSTT service mode, there are different components that are put in place starting from data collection, tone diacritization, speech database development, text corpus development, data labeling, pre-processing/feature extraction, acoustic modeling, language modeling, pronunciation dictionary development, etc.

The Standard Yorùbá text data for developing the text corpus for the CSYSTT service mode was collected and analyzed using Antconc software given in Figure 5, 6 and 7.

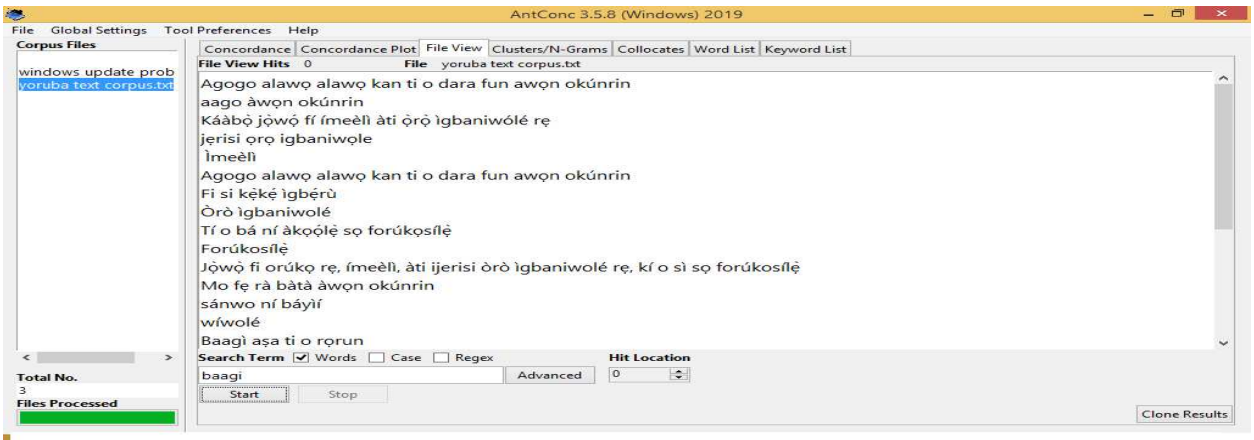

Figure 5: The Standard Yorùbá Text Data Collection using AntConc

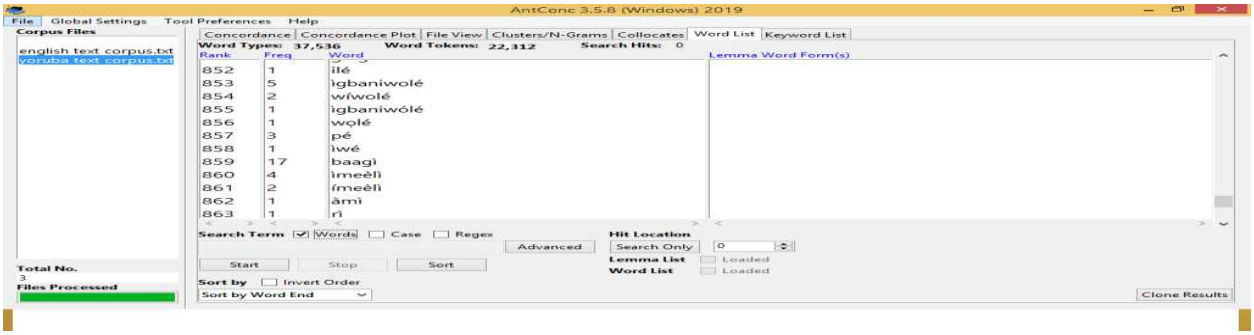

Figure 6: The Standard Yorùbá Text Data having 22,312 Data-Points with 37,536 Word Types using AntConc.

The pre-processing/feature extraction goal is to get the speech signal of each word or sentence spoken which is done by taking the speech sample, segmenting the speech file, detecting silence regions and noise filtering. After the raw speech signal is digitized and segmented, it is then windowed and encoded into a set of initial parameters which is defined by MFCC. In this context, silence and noise region was conducted using MFCC for all the audio files. The MFCC steps were implemented using TensorFlow. For instance, Whenever a user speaks into the system, his/her voice is recorded for up to 60 seconds using a 44 KHz sampling rate, then the speech sample was saved as a wav format to utilize along the process. After the speech sample has been taking, the next step is to perform the necessary feature extraction from the user voice using MFCC.

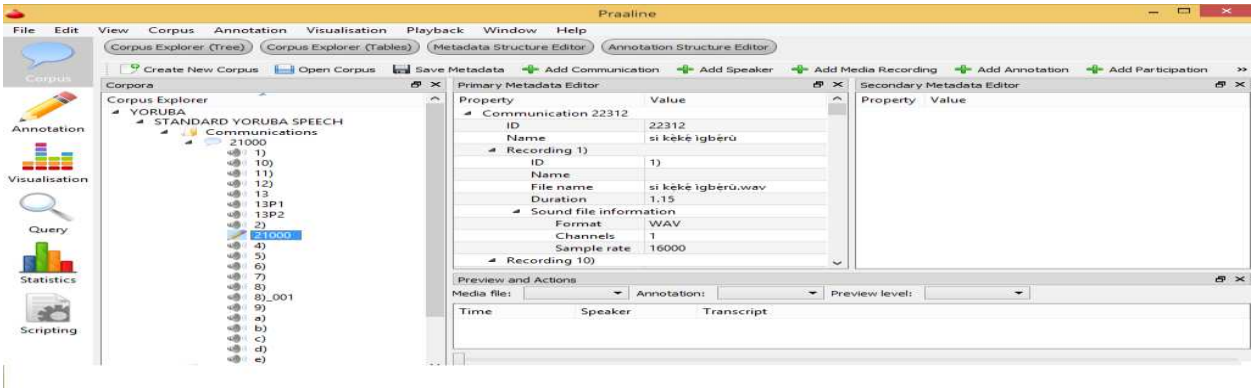

Figure 7: The Standard Yorùbá Speech Data Collection using Praaline

Each of the speech sample were extracted using 128 MFCCs. The raw continuous speech signal are transformed into Mel frequency cepstral coefficients (MFCCs) so as to mimic the way the human ears perceive sounds and at the same time remove outliers (noise and silence). After the feature extraction process is the acoustic modeling process. The acoustic modeling stage has the responsibility of constructing and training the models needed for the CSYSTT service mode. The goal of this stage is to get a Gaussian distributed Acoustic Nudging MFCC clean speech. It loads the MFCC features, and it then iterates through them while performing parameters updation. After getting the Acoustic Nudging MFCC speech then Gaussian Mixture distributed model is utilized, which serves as the final algorithm.

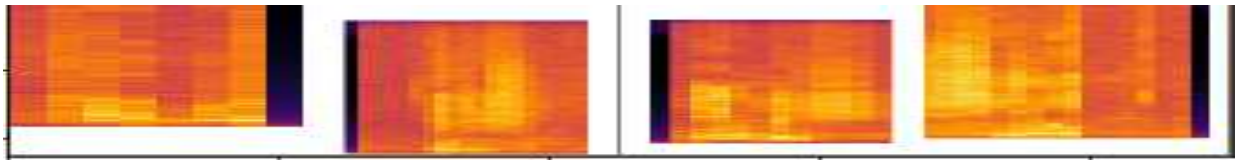

Figure 8: An Acoustic Irrational and Re-formulated Training Speech Signals.

Discussion

Having validated that ANGM model can be used to design a robust CSYSTT, The ANGM model is evaluated with respect to performance. The evaluation of this system is built on Quantitative (objective) tests which is based on accuracy. The main aim of experiment-1 is to determine the recognition accuracy level of the system in real time. In order to achieve this step, a simulation was undertaken to determine the accuracy which is evaluated based on Loss-(training, validation and testing), Word Recognition Rate (WRR), and Word Error Rate (WER). Also, the approach used in this experiment is comparative. It is of utmost importance that the experiment is planned so that data is collected to enable comparison between the ANGM model and other speech recognition methods. The data generated from applying the aforementioned evaluation metrics was then used to determine which method is the optimum best.

H_0: The performance in terms of accuracy level of the ANGM model for a user spoken words or sentences is directly proportional to the reference words or sentences.

Experimental Dataset

Training Dataset: 17,850 (22 speakers), 1200.5 hours (3-12 seconds each)

Validation Dataset: 2,162 (10 speakers), 860.1 hours (3-12 seconds each)

Testing Dataset: 2,300 (15 speakers), 880.3 hours (3-20 seconds each)

Training dataset is used in performing the ANGM model training by initializing and building the weights of the network.The validation dataset is used after the network has been trained which is used in tuning the network's hyperparameters (leaning rate, batch size, etc.), compare them and foresee how the tuning affects the predictive accuracy of the ANGM model. The test dataset is used in testing the predictive accuracy of the trained ANGM model on previously unseen data after training and validation processing has taken place.The experimental dataset

for the performance evaluation process is the test dataset which comprises of 2300 speech files that belongs to 15 different speakers of gender (male and female) comprising of both children and adults with a total of 880.3 hours. Moreover, each speech files has its corresponding transcription. Each test speaker recorded a data of the range 80-280 speech files as given in Table 5.

Table 5: The Continuous Standard Yorùbá Test Dataset

S/N	Speakers	Hours	No of Speech Files
1	Speaker 1(Male Adult)	85.6	170
2	Speaker 2 (Female Adult)	89.8	190
3	Speaker 3 (Female Adult)	85.4	160
4	Speaker 4 (Female Child)	63	80
5	Speaker 5 (Male Adult)	83.2	148
6	Speaker 6 (Male Child)	64.1	90
7	Speaker 7 (Female Adult)	88.4	185
8	Speaker 8 (Male Adult)	86.2	160
9	Speaker 9 (Male Child)	69.1	85
10	Speaker 10 (Female Adult)	90.2	190
11	Speaker 11 (Male Adult)	80.2	160
12	Speaker 12 (Male Child)	66.3	82
13	Speaker 13 (Female Adult)	175.9	280
14	Speaker 14 (Male Adult)	86.3	170
15	Speaker 15 (Male Adult)	80.2	150
	413.6	**Total = 880.3**	**Total = 2300**

The goal of the training and validation objective is to make the cross entropy loss function as small as possible. If validation loss > training loss, then there is overfitting for the ANGM model. If the validation loss << training loss then there is underfitting of the ANGM model but if the training loss == validation loss, then there is perfectfitting. The aim is to make sure that the validation_loss and train_loss as low as possible and at the same time, making sure that the train_accuracy and validation_accuracy as high as possible to achieve a robust speech recognition. Table 6 presents the comparison between validation_loss values and train_loss values which shows that their values are almost the same when rounded up to one-significant figure. Table 7 also presents a sample comparison between validation_accuracy and train_accuracy with validation_accuracy slightly higher than the training accuracy but almost the same when rounded up to one significant figure which automatically means that the ANGM model build is learning correctly.

Table 6: A Sample Comparison between Validation_Loss and Train_Loss Values

Epoch	Time Taken (s)	Validation_Loss Values	Train_Loss Values
1/24	04.19	0.3823	0.2651
2/24	04.16	0.2674	0.1648
3/24	04.13	0.2367	0.0954
4/24	04.17	0.2267	0.0875
5/24	04.18	0.1945	0.0787
6/24	04.19	0.1674	0.0077
7/24	04.20	0.0863	0.0576
8/24	04.17	0.0563	0.0456
9/24	04.18	0.0421	0.0334
24/24	04.19	0.0376	0.0328

Table7: A Sample Comparison between Validation_Accuracy and Train_Accuracy Values

Epoch	Time Taken (s)	Validation_Accuracy Values	Train_Accuracy Values
1/24	04.19	0.9215	0.9024
2/24	04.16	0.9418	0.9115
3/24	04.13	0.9434	0.9385
4/24	04.17	0.9685	0.9523
5/24	04.18	0.9558	0.9551
6/24	04.19	0.9581	0.9523
7/24	04.20	0.9595	0.9543
8/24	04.17	0.9554	0.9523
9/24	04.18	0.9587	0.9576
24/24	04.19	0.9789	0.9728

In this study, test_accuracy and test_loss metrics is also adapted in this study. Table 9 presents a sample comparison between the train_accuracy and test_accuracy which shows that train_accuracy is slightly higher than test_accuracy but almost the same when rounded up to one significant figure which gives the ANGM model a perfectfitting.

Table 8: A Sample Comparison between Train_loss and Test_Loss Values

Epoch	Time Taken (s)	Train_Loss Values	Test_Loss Values
1/24	06.09	0.3672	0.4223
2/24	06.07	0.2126	0.3271
3/24	06.10	0.1867	0.1785
4/24	06.11	0.1019	0.2217
5/24	06.12	0.0787	0.0975
6/24	06.09	0.0698	0.0718
7/24	06.11	0.0698	0.0540
8/24	06.12	0.0623	0.0433
9/24	06.13	0.0545	0.0359
24/24	06.14	0.0485	0.0257

Table 9: A Sample Comparison between Train_Accuracy and Test_Accuracy Values

Epoch	Time Taken (s)	Train_Accuracy Values	Test_Accuracy Values
1/24	06.09	0.9234	0.9180
2/24	06.07	0.9367	0.9256
3/24	06.10	0.9440	0.9467
4/24	06.11	0.9596	0.9485
5/24	06.12	0.9554	0.9594
6/24	06.09	0.9567	0.9561
7/24	06.11	0.9578	0.9576
8/24	06.12	0.9592	0.9585
9/24	06.13	0.9695	0.9592
24/24	06.14	0.9799	0.9696

a) Word Error Rate (WER)

The WER is calculated given by

$$\text{WER} = 100 * \frac{\text{Sw+Ds+1s}}{\text{N}} \quad \textbf{1.1}$$

Where N = D + S+ C

Also, the Word Recognition Rate (WRR) metric is a complement of the Word Error Rate (WER) given by

$$WRR = WER^{1} \quad \textbf{1.2}$$

Accuracy $(WRR) = (1 - WER) * 100\%$ …….…….....**1.3**

Table 10: WER Results for 15 different Test Speakers using ANGM Model

S/N	Speakers	No of Speech Files	Mean Accuracy (WRR)%	WER %
1	Speaker 1(Male Adult)	170	94.576	5.424
2	Speaker 2 (Female Adult)	190	95.692	4.308
3	Speaker 3 (Female Adult)	160	96.785	3.215
4	Speaker 4 (Female Child)	80	92.452	7.548
5	Speaker 5 (Male Adult)	148	96.683	3.317
6	Speaker 6 (Male Child)	90	92.643	7.357
7	Speaker 7 (Female Adult)	185	95.856	4.144
8	Speaker 8 (Male Adult)	160	96.932	3.068
9	Speaker 9 (Male Child)	85	93.593	6.407
10	Speaker 10 (Female Adult)	190	97.547	2.453
11	Speaker 11 (Male Adult)	160	95.327	4.673
12	Speaker 12 (Male Child)	82	94.735	5.265
13	Speaker 13 (Female Adult)	280	96.855	3.145
14	Speaker 14 (Male Adult)	170	94.954	5.046
15	Speaker 15 (Male Adult)	150	94.532	5.468
		Total = 2300	**Average = 95.277**	**Average = 4.723**

The ANGM model achieved a word error rate (WER) average score of 4.723% and a mean accuracy (WRR) across 2,330, 880.3 hours test dataset. The best performance in terms of WER using 15 speakers was achieved using 128 MFCCs with 2.453% from speaker 10 (female adult).

5.2.4 Comparison of the ANGM model with Popular Speech Recognition Models

To evaluate the efficacy of the ANGM model, this study compare the performance of the ANGM model with existing models that have been popularly used in literature with application as either a single models or hybrid models s which are as GMM, GMM-HMM, CNN, GMM-CNN, DNN. These models have been widely studied in different application scenarios where the training dataset was large and small with a dataset of 500 and above. The main purpose of this comparative study is to show how the ANGM model performs compared with popular speech recognition models. The test dataset for this study is utilized for this comparative study. The results of this comparative study are presented in Table 11.

Table 11: WER Results for 5 Existing Speech Recognition Models with ANGM Model

S/N	Sentences Used	ANGM	GMM	GMM-HMM	CNN	GMM-CNN	DNN
1	Mo fẹ rà bàtà àwọn okúnrin	4.248	5.324	4.957	5.823	4.667	5.483
2	Fi si kẹ̀kẹ́ ìgbẹ́ru	4.108	5.478	4.802	5.392	4.598	5.356
3	Àjọsọpọ ti nmí kekere bàtà tàwọn obìnrin	4.203	5.362	4.723	5.237	4.747	5.823
4	Sánwó ní báyìí	3.315	5.056	4.623	4.965	4.492	5.623
5	Mo fẹ rà ọmọ	3.623	5.256	4.701	4.813	4.276	5.600

	wẹwẹ kanfasi						
6	Mo fẹ rà baagì àwọn obìnrin	6.757	7.789	6.334	6.934	6.534	7.893
7	Mo fẹ rà bàtà awọ ilẹ	4.255	4.911	4.612	4.833	4.335	5.064
8	mo fẹ rà bàtà àwọn obìnrin	3.967	4.567	3.845	3.867	3.267	4.834
9	Mo fẹ rà bàtà àwọn okúnrin	3.921	4.578	3.822	3.805	3.235	4.745
10	Àwọn bata alawọ goolu	3.145	4.283	3.941	4.002	4.122	4.782
11	Mo fẹ rà aago àwọn okúnrin	145	4.803	4.197	4.397	4.523	5.324
Average WER		**4.091**	**5.219**	**4.596**	**4.915**	**4.436**	**5.503**

In Table 11, the WER of each model was obtained by taking the average of each continuous Yorùbá speech sentences from the test dataset using different speakers with different age brackets conducted under natural environment where there is noise interference. The ANGM model WER was estimated at 4.091%, the GMM WER model was estimated at 5.219%, the GMM-HMM model WER was estimated at 4.596%, the CNN WER model was estimated at 4.915%, the GMM-CNN model WER was estimated at 4.436%, the DNN model WER was estimated at 5.503%. The WER decrease is obtained between two systems and analyzed using Equation 1.2. The average CSYSTT performance which is based on ANGM model demonstrates a WER decrease of 1.1%, 0.5%, 0.8%, 0.3%, and 1.4% when compared with other models. The results of the experiment obtained in Table 11 and 12 are compared with a column chart in Figure 9.

$$\%Decrease = \left|\frac{WER_{P2}-WER_{P1}}{WER_{P1}}\right| * 100 \quad \text{.....} \quad \textbf{5.4}$$

Table 12: Performance of Six Speech Recognition Models

S/N	Speech Recognition Models	WER% (Rounded up)
1	ANGM	4.1
2	GMM	5.2
3	GMM-HMM	4.6
4	CNN	4.9
5	GMM-CNN	4.4
6	DNN	5.5

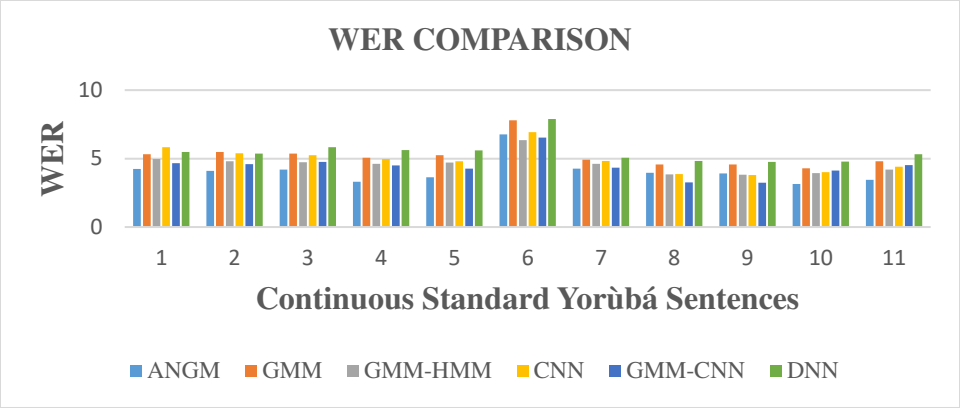

Figure 9: WER Comparison Chart

In this study, WER comparison was also made based on existing literature studies for under-resourced languages [24] [25] [26] with this present study having different WER range of 11.27%, 10.07%, 5.48%, 5.45%, 4.92%, etc. Existing African language speech recognition system [8] [20] also achieved a WER range of 9.48%, 14.83%, 18% and when compared with this current study, ANGM model has a significant lower average WER of 4.091% using 2300 test sentences.

Conclusion

After applying the acoustic nudging on both the training, validation and test dataset, the speech signals were re-formulated in real time, after which the Gaussian mixture model was applied. The comparison between validation_loss values and train_loss values shows that their values are roughly the same. Also, and athe comparison between the validation_accuracy values and train_accuracy values shows that validation_accuracy is slightly higher than the training accuracy but roughly the same when rounded up to one significant figure. Also, the comparison between the train_accuracy and the test-accuracy shows that the train_accuracy is slightly higher when compared with the test_accuracy, which all evidently expatiates that ANGM model is a perfectfitting.

The mean accuracy (WRR) achieved for the ANGM model is 95.277% and the mean Word Error Rate (WER) is 4.723% across 2,330, 880.3 hours test dataset, and the best performance in terms of WER using 15 speakers was achieved using 128 MFCCs with 2.453% from a female adult speaker. To evaluate the efficacy of the ANGM model, this study compared the performance of the ANGM model with existing models that have been popularly used in literature with application as either a single models or hybrid models using 11 CSY test dataset. These models have been widely studied in different application scenarios where the training dataset was large and small with a dataset of 500 and above. The main purpose of this comparative study is to show how the ANGM model performs compared with popular speech recognition models. The ANGM model achieved the least mean WER % of 4.091 compared to the afore-mentioned existing speech recognition models with mean WER% of 5.219, 4.596, 4.915, 4.436, and 5.503. The average CSYSTT performance which is based on ANGM model demonstrates a WER decrease of 1.1%, 0.5%, 0.8%, 0.3%, and 1.4% when compared with other models.

Existing African language speech recognition also achieved a WER range of 9.48%, 14.83%, and 18%, and when compared with this current study, ANGM model has a significant lower average WER of 4.091% using 2300 test sentences. Our study caters for the observed limitations of existing speech domain through development of a large vocabulary Continuous Standard Yorùbá Speech-to-Text Engine with consideration of Tone diacritization and an Acoustic Nudging-based Gaussian Mixture (ANGM) Model to allow automatic correction of ASR errors, involving user's acoustic irrational behavior in speech with low-resourced attributes for any language to achieve better accuracy and system performance. The Yoruba speech engine derived from this study may also be used as platform to host the applications in [27] [28]. This study provides future research which includes the design of a Voice Engine System for other Yorùbá Accents, design of a Voice Engine System for other African Languages, and application of the ANGM model in other languages for enhanced accuracy and system performance.

Abbreviations

ANGM**:** Acoustic Nudging-based Gaussian Mixture Model; WRR: Word Recognition Rate; WER: Word Error Rate; ASR: Automatic Speech Recognition; STT: Speech-to-Text; TTS: Text-to-Speech; CSYSTT: Continuous Standard Yoruba Speech-to-Text; VQ: Vector Quantization; HMM: Hidden Markov Model; CNN: Convolution Neural Network; DNN: Deep Neural Network; CSY: Continuous Standard Yorùbá

Ethics Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
Not Available.

Competing Interests
The authors in this research do not have any competing interest.

Funding
No funding for this project.

Authors' contribution

LKA researched and conducted the experiment as well as writing the manuscript, ***AA***provided publication recommendations, guidance during the project experimentation phase, reviewed the manuscript as well as supporting the publication of the manuscript, ***IO-A***provided guidance during project experimentation phase and contributed on reviewing the manuscript,***VA***did the final language-editorial work, corrected spelling errors and ensured grammar compliance, ***APT***Revised the manuscript and ensured grammar compliance.

Acknowledgements

I wish to acknowledge my profound appreciation to Covenant University for their support towards this research.

Author Information
[1]*Department of Computer and Information Sciences, Covenant University, Ogun, Nigeria;*
[2]*Department of Computer and Information Sciences, Covenant University, Ogun, Nigeria;*
[3]*Department of Computer and Information Sciences, Covenant University, Ogun, Nigeria.*
[4]*National Productivity Center, Kaduna*
[5]*Department of Mass Communication, Joseph Ayo Babalola University, Osun, Nigeria.*
[6]*Department of Computer and Information Sciences, Covenant University, Ogun, Nigeria.*

References

[1]. Gaikwad, SK., Gawali, BW, Yannawar, P. A Review on Speech Recognition Technique. *International Journal of Computer Applications*, 2010; *10*(3), 16–24. https://doi.org/10.5120/1462-1976
[2]. Nwakanma, PC, Ibe, RC. Globalization and economic growth. An econometric dimension drawing evidence from Nigeria. *International Review of Management and Business Research*, 2014; *3*(2), 771.
[3]. Deborah, NO, Rhoda, IA, Williams, OA. *Development of a Mobile Tourist Assistance for a Local Language*. 2017; *6*(1), 5–9. https://doi.org/10.5923/j.tourism.20170601.02
[4]. Prachi, U., & Bhope, G. (2015). *Voice Based Collaborative Banking*.

[5]. Akintola, A., Ibiyemi, T. Machine to Man Communication in Yorùbá Language. Annals. Computer Science Series, 2017; 15(2).
[6]. Sebastian BW, Sebastian B, Pinar T, Leon D. Software Development: Advanced Computing. IT University of Copenhagen.October, 2018.
[7]. Levis, J, Suvorov, R.. Automatic speech recognition. *The encyclopedia of applied linguistics*.2012; https://doi.org/10.1002/9781405198431.wbeal0066
[8]. Wahab A., Atanda, F, Azmi, S, Yusof, M, & Hariharan, M. . *Yorùbá Automatic Speech Recognition : A Review Yorùbá Automatic Speech Recognition : A Review*. (June 2013).
[9]. Oluseye A,. Yorùbá: A Grammar Sketch: Version 1.0; Journal of National Technical Support, 2014.
[10]. Le, VB, Besacier, L. . Automatic speech recognition for under-resourced languages: Application to Vietnamese language. *IEEE Transactions on Audio, Speech and Language Processing*, 2009;*17*(8), 1471–1482. https://doi.org/10.1109/TASL.2009.2021723
[11]. Saksamudre, KS., Shrishrimal, PP, Deshmukh, RR. A Review on Different Approaches for Speech Recognition System. *International Journal of Computer Applications*, 2015; *115*(22), 23–28. https://doi.org/10.5120/20284-2839
[12]. Thangarajan, R. Speech Recognition for Agglutinative Languages. *Modern Speech Recognition Approaches with Case Studies*. 2012;https://doi.org/10.5772/50140
[13]. Vu, NT, Kraus, F, Schultz, T. Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, (August 2017); 3145–3148.
[14]. Vyas, JV. *Study of Speech Recognition Technology and its Significance in Human-Machine Interface*. 2017; *3*(10), 416–422.
[15]. Chauhan, V, Dwivedi, S, Karale, P, Potdar, PSM. *Speech to Text Converter Using Gaussian Mixture Model (GMM) of Electronics and Telecommunication Engineering*. 2016; 125–129.
[16]. Satori, H., & Elhaoussi, F. Investigation Amazigh speech recognition using CMU tools. *International Journal of Speech Technology*, 2014; *17*(3), 235–243. https://doi.org/10.1007/s10772-014-9223-y
[17]. Mohanty, S, Swain, BK. Markov Model Based Oriya Isolated Speech Recognizer-An Emerging Solution for Visually Impaired Students in School and Public Examination. 2010; *Special Issue of IJCCT*, *2*(2,3,4), 107–111. Retrieved from https://www.researchgate.net/publication/266268918.
[18]. Ltaief, A. Ben, EY, Graja, M, & Belguith, L. H. *Automatic speech recognition for Tunisian dialect*. 2016; 1–8.
[19]. Das, P, Acharjee, K, Das, P, Prasad, V. *Voice Recognition System : Speech-To-Text*. (July 2016).
[20]. Adetunmbi, O., Obe, O, & Iyanda, J. Development of Standard Yorùbá speech-to-text system using HTK. *International Journal of Speech Technology*. 2016; https://doi.org/10.1007/s10772-016-9380-2
[21]. Laleye, FA, Besacier, L, Ezin, EC., Motamed, C. First automatic fongbe continuous speech recognition system: Development of acoustic models and language models. In *2017 Federated Conference on Computer Science and Information Systems (FedCSIS),* 2017, September; (pp. 477-482). IEEE.
[22]. Galvan, RF., Barranco, V, Galvan, JC., Batlle, Sebastian FeliuFajardo, S, & García. We are IntechOpen , the world ' s leading publisher of Open Access books Built by scientists ,for scientists TOP 1 %. *Intech*, *i*(tourism), 13. 2016; https://doi.org/http://dx.doi.org/10.5772/57353
[23]. Ajayi L.K., Azeta, A.A., Odun-Ayo I., Chidozie, F, Ajayi, PT. *Automatic Re-Formulation of user's Irrational Behavior in Speech Recognition using Acoustic Nudging Model; Journal of Computer Science. 16. 17ss31-1741. 10.3844/jcssp.2020.1731.1741.*
[24]. Beserra, AA. V., Silva, WLS, Serra, GLD. O. A GMM/CPSO speech recognition system. *IEEE International Symposium on Industrial Electronics*, *2015-September* (December 2018), 26–31. https://doi.org/10.1109/ISIE.2015.7281438
[25]. Stuttle, MN.*A Gaussian Mixture Model Spectral Representation for Speech Recognition*. July 2003; 163.
[26]. Huggins-daines, D., Kumar, M, Chan, A, Black, AW, Ravishankar, M, Rudnicky, AI, Avenue, F,*Pocketsphinx : A Free , Real-Time Continuous Speech Recognition System For Hand-Held Devices Language Technologies Institute (dhuggins , mohitkum , archan , awb , rkm , air)@ cs . cmu . edu*. 2016; 185–188.
[27]. Azeta,A.,Da-Omiete A.I.,, Azeta, I.V.,Emmanuel. O.I.,. Fatinikun, D. O., EkpunobiE. "Implementing a Medical Record System with Biometrics Authentication in E-Health"**.** *IEEE AFRICON:* Science, Technology and Innovation for Africa, AFRICON2017. 18-20. September 2017; The Avenue V&A Waterfront Cape Town South Africa.
[28]. Azeta A. A., Ayo C. K., Atayero A. A. and Ikhu-Omoregbe N. A. Application of VoiceXML in e-Learning Systems", Cases on Successful E-Learning Practices in the Developed and Developing World: Methods for the Global Information Economy. Chapter 7, Published in the United States of America by Information Science Reference (an imprint of IGIGlobal). Edited by Bolanle A. Olaniran., 2009; PP. 92-108.

Figures

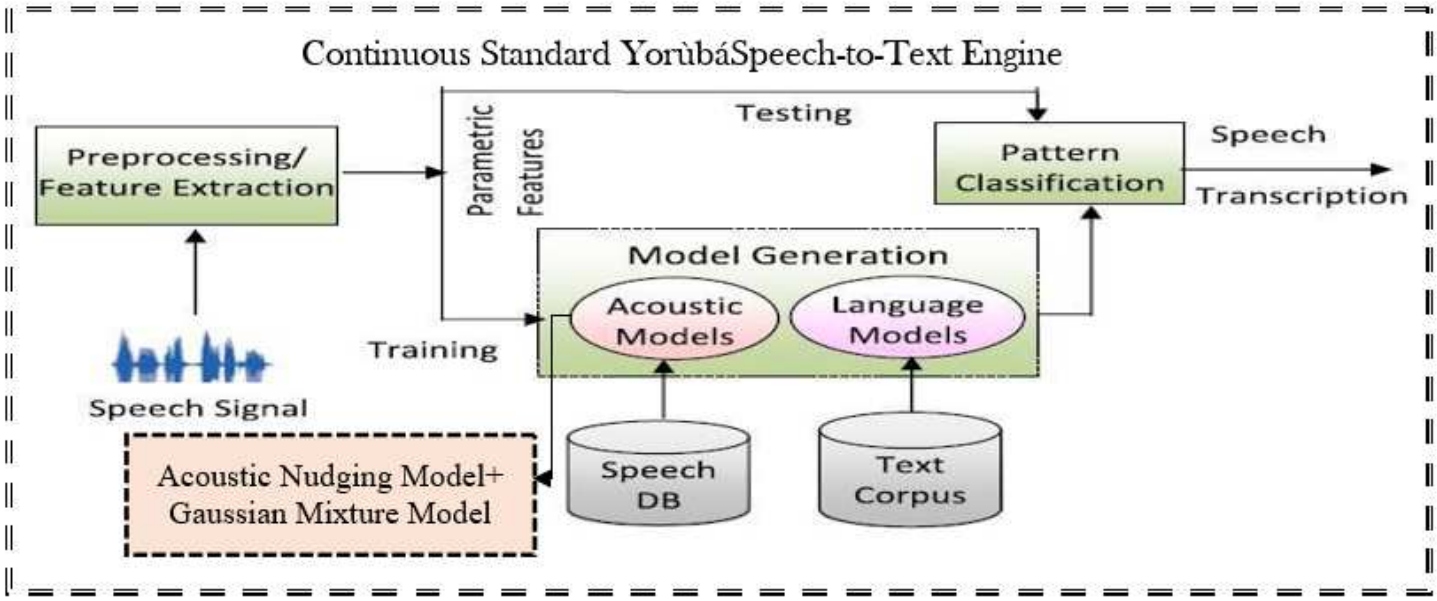

Figure 1

The Proposed Acoustic Nudging-Based Gaussian Mixture Model

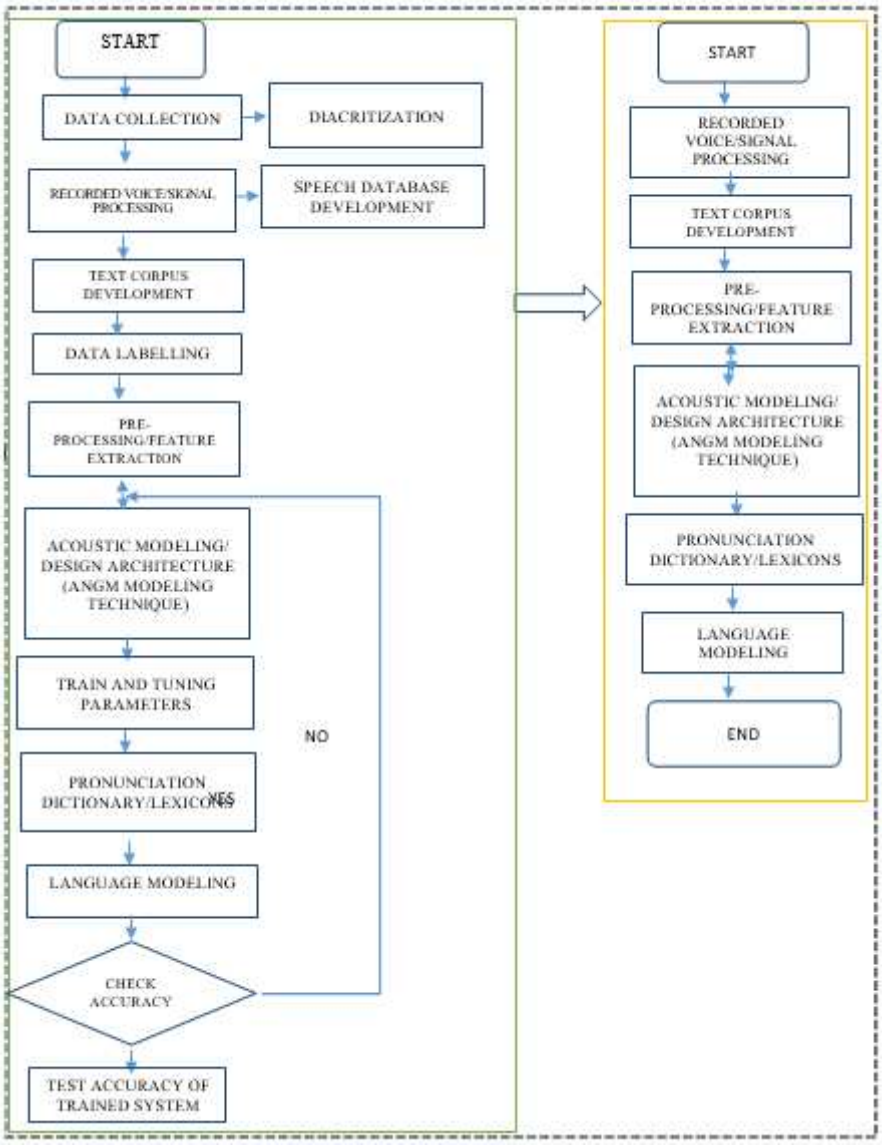

Figure 2

The Methodology Process Flowchart of the Continuous Speech-to-Text (CSYSTT) Model

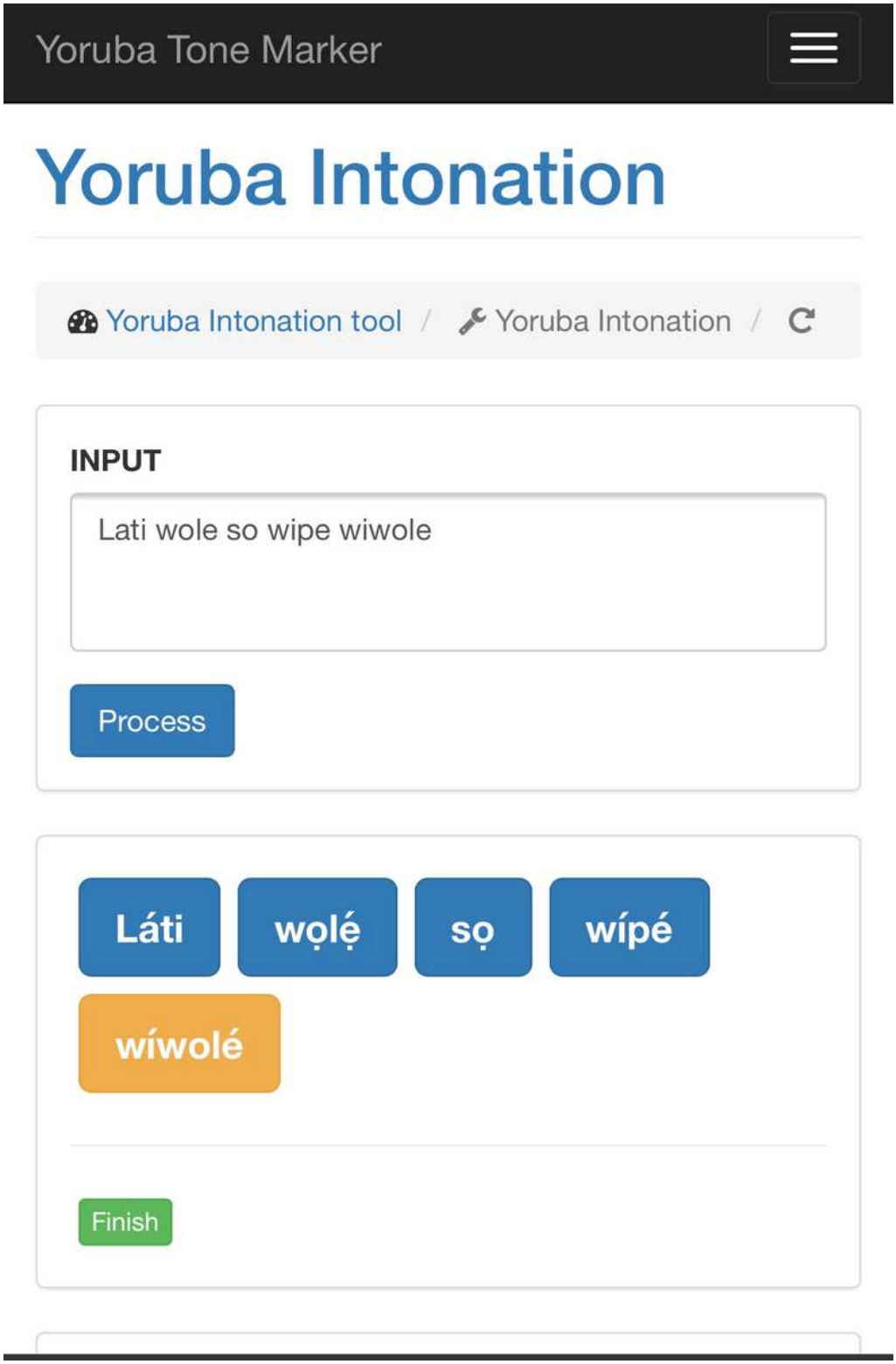

Figure 3

A Semi-Supervised Standard Yorùbá Tone Diacritization Process

Variables used in the ANGM Algorithm

X: User's acoustic rational behavior
m: Acoustic Nudging Model predicted values
P (5): Acoustic Nudging Model prescribed values for the heuristics and biases
ṕ(5): Replaced Heuristics and Biases.
S: User's speech signal
N: Number of speech signals
X^1: User's acoustic irrational behavior
i: Number of experiment
P: Pitch
L: Loudness or Sound Pressure
At: Timbre Ascend Time
Dt: Timbre Descend Time
T_bW: Time between each words
μ_1: The first GMM parameter "mean"
σ_1: The first GMM parameter "standard deviation"
μ_2: The ANGM parameter "mean"
σ_2: The ANGM parameter "standard deviation"
Q_1 - Q_5: Variation Slider Endpoints

Figure 4

The ANGM Algorithm

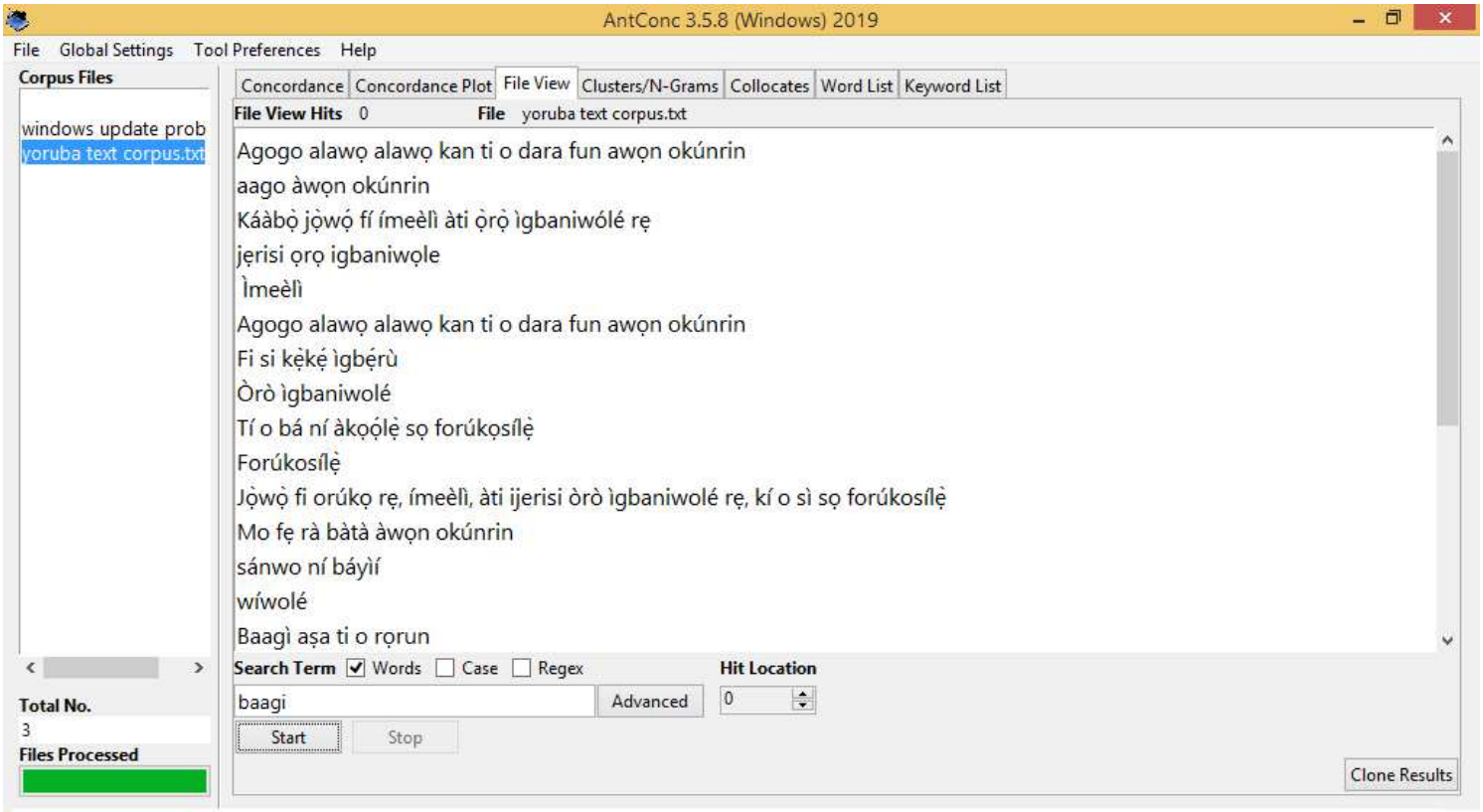

Figure 5

The Standard Yorùbá Text Data Collection using AntConc

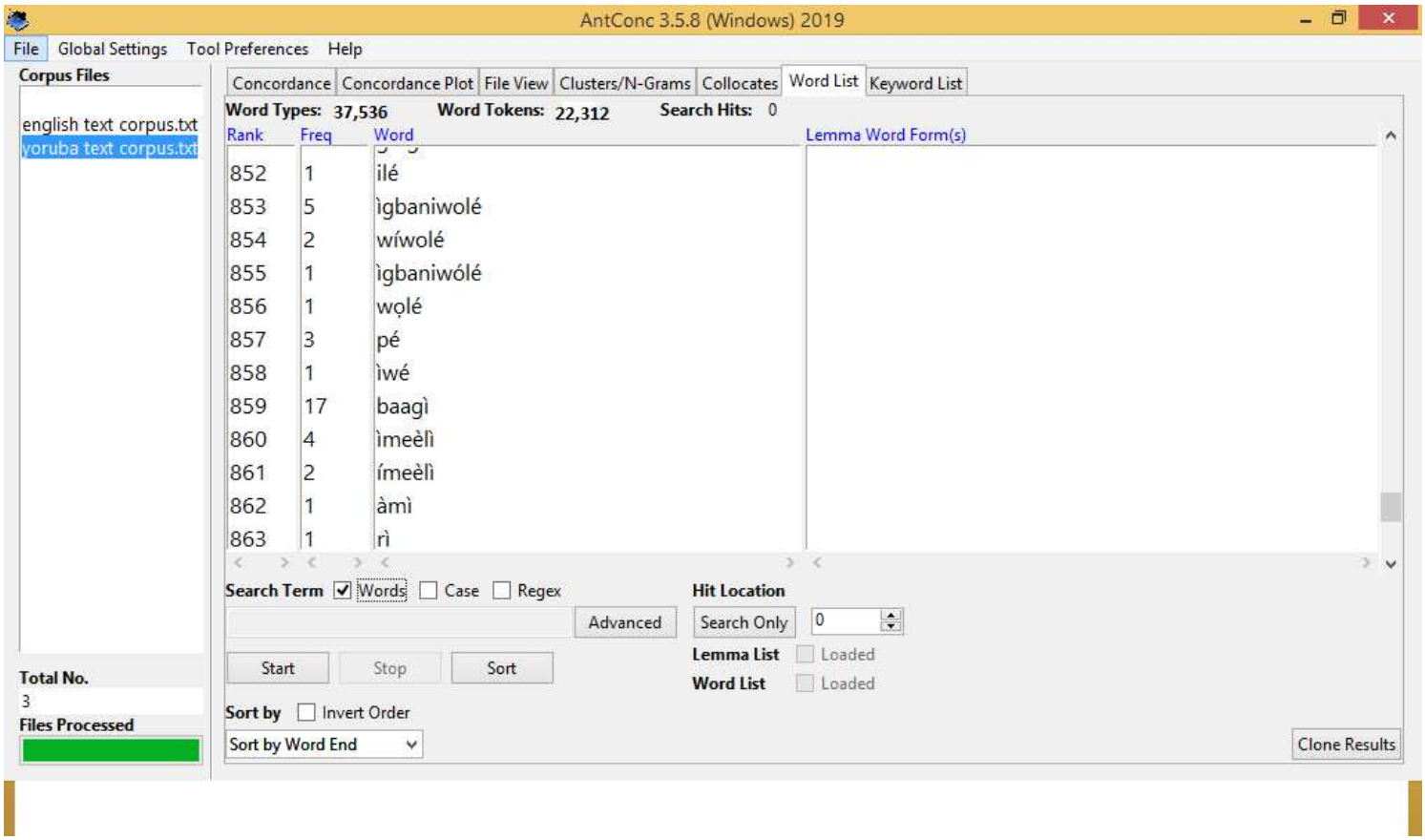

Figure 6

The Standard Yorùbá Text Data having 22,312 Data-Points with 37,536 Word Types using AntConc.

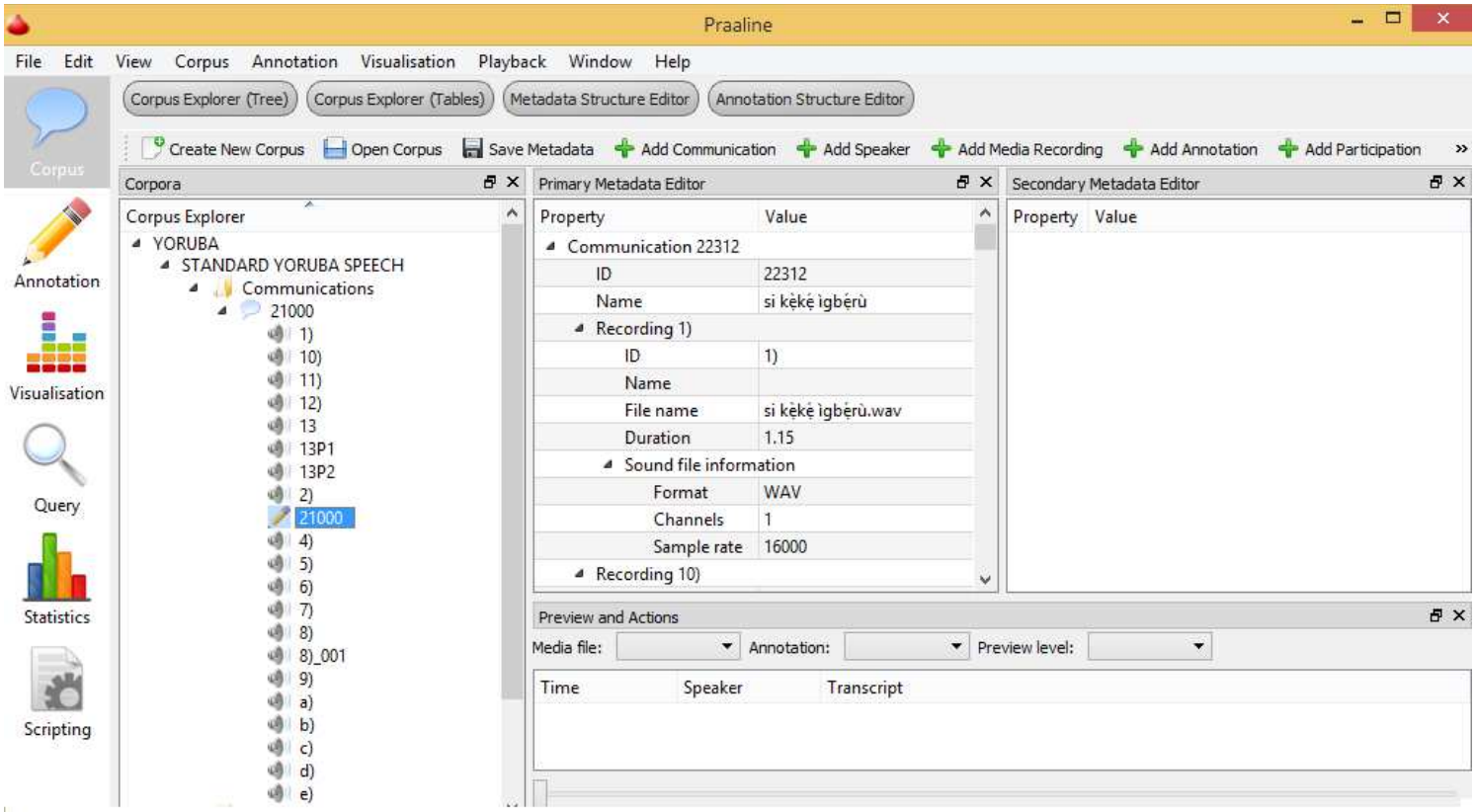

Figure 7

The Standard Yorùbá Speech Data Collection using Praaline

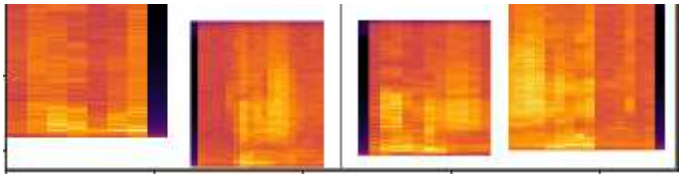

Figure 8

An Acoustic Irrational and Re-formulated Training Speech Signals.

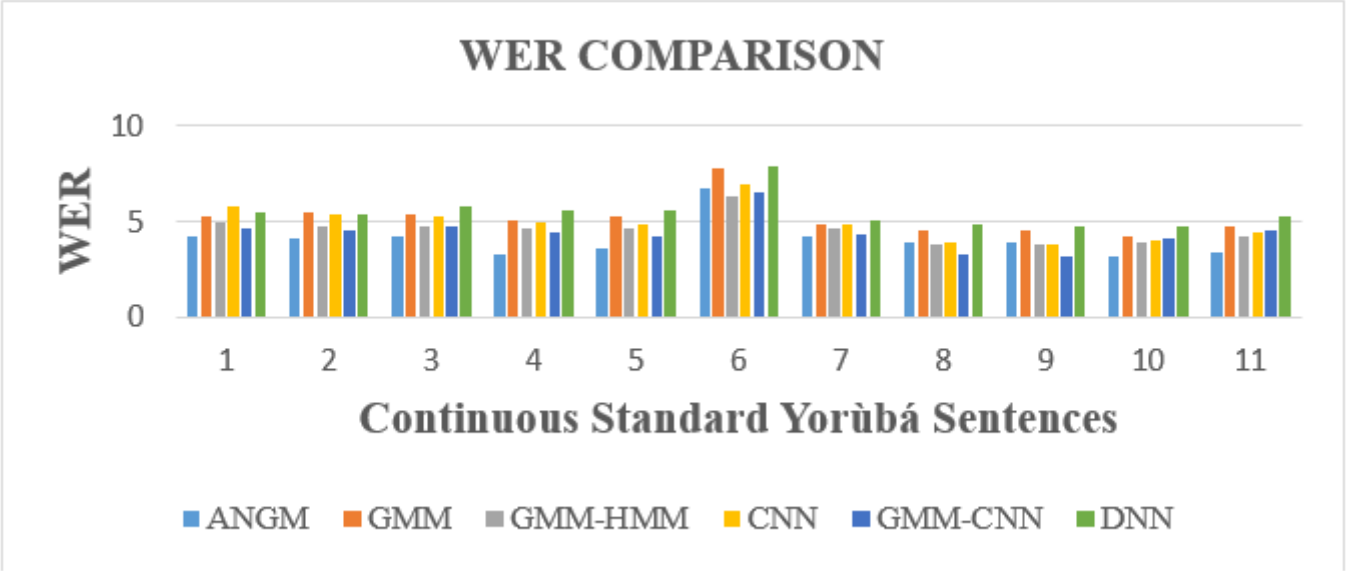

Figure 9

WER Comparison Chart