Acoustic Nudging-Based Model for Vocabulary Reformulation in Continuous Yorùbá Speech Recognition

Lydia Kehinde Ajayi[1](✉), Ambrose Azeta[2], Isaac Odun-Ayo[3], and Enem Theophilus Aniemeka[4]

[1] ENSI Department, National Identity Management Commission, Lagos, Nigeria
lydia4reel@gmail.com
[2] Department of Computer Science, Namibia University of Science and Technology, Windhoek, Namibia
aazeta@nust.na
[3] Department of Computer Science, Chrisland University, Abeokuta, Nigeria
isaac.odun-ayo@chrislanduniversity.edu.ng
[4] Department of Cybersecurity, Faculty of Computing, Air Force Base, Kaduna, Nigeria
ta.enem@afit.edu.ng

Abstract. Speech recognition is a technology that aid processing of speech signals through communicating with computer applications**.** Previous studies exhibits speech recognition errors arising from users' acoustic irrational behavior. This research paper provides acoustic nudging-based model for reformulating the persistence of automatic speech recognition errors that involve the user's acoustic irrational behavior and distortion of speech recognition accuracy. Gaussian mixture model (GMM) helped in addressing the low-resourced attribute of Yorùbá language to achieve better accuracy and system performance. From the implemented results, it was observed that the proposed acoustic nudging-based model improves accuracy and system performance based on Word Error Rate (WER), validation, testing and training accuracy. The evaluation results for the mean WER was 4.723% when compared to existing models. This approach thereby reduces error rate when compared with previous models by GMM (1.1%), GMM-HMM (0.5%), CNN (0.8%), and DNN (1.4%). Therefore, this work was able to discover a foundation for advancing the current understanding of under-resourced languages and development of an accurate and precise model for speech recognition.

Keywords: Acoustic nudging model · Gaussian mixture model · Automatic speech recognition · Communication and nudging

1 Introduction

Speech is one of the most convenient means of communication between people and humans [1–4]. Human-computer interaction is based on the syntactic arrangement of lexical names drawn from huge vocabularies [3]. Yorùbá language is an indigenous

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022
O. Gervasi et al. (Eds.): ICCSA 2022, LNCS 13375, pp. 494–508, 2022.
https://doi.org/10.1007/978-3-031-10522-7_34

language exiting in the second largest ethnic group in Nigeria spoken by more than 50 million people with 30 million speaking Standard Yorùbá [5] [6].

There are three different spoken languages in Nigeria which are Yorùbá, Igbo, and Hausa with their accents, however, the focus of this study is on Standard Yorùbá. 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 these languages are be located [7, 8].

Yorùbá also borrow 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 language such as Yoruba is a current field of research that has experienced significant progress. Under-resourced language is defined as a language that lacks unique writing system and has 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 system to realize voice to speech synthesis [2]. There are different classifications of automatic speech recognition which are based on utterances that are broken down into isolated, continuous, spontaneous and connected. This paper is organized as follows: Sect. 2 highlights related works, Sect. 3 described the acoustic nudging-based models. The results and discussions are contained in Sect. 4, while Sect. 5 provides conclusion and further works of the study.

2 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 text data [10, 11]. [12] designed HMM-based Tamil Speech Recognition 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 the speech signal by MFCC for multiple isolated words to train audio files to get spoken words recognized and gave an accuracy of 98%, but the limitation is that it uses isolated words. Automatic speech recognition for Tunisian dialect, which is also an under-resourced language was developed in [17]. The HMM-GMM model using MFCC and HMM-GMM with LDA were compared in the study. These approaches gave a WER of 48.8% and 48.7%. [17] proposed an enhanced ASR system for Arabic using Kaldi Toolkit to build the system around, after which the acoustic models were trained using the HMM-GMM model, and the data collected was based on Standard Arabic news broadcasts. The language model was trained with two Corpora, including GigaWord3 Arabic corpus (1000 words) and the acoustic training data transcription (315000 words). The result gave a WER of 14.42%.

[3] provided a Mobile Tourist Assistance for Yorùbá Language developed for tourists and implemented on Android application. The researchers developed a speech-to-text system for easy communication between locals and tourists. The text data were gathered from on-site interaction with the native speakers in four different domains: Market, Media, 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 and a male voice. The accuracy of the system was established to be 85% for clarity and 88% for naturalness. The limitation of this system is that it is a one-way interactive system, based on isolated words. [18] also developed a Home Automation speech-to-text system that lets a user control computer function and dictates text by voice using HMM as acoustic modeling, MFCC for feature extraction and for feature training of the dataset. [19] developed a standard Yorùbá Isolated speech-to-text system using HTK, which can recognize isolated words spoken by users using previous data. The system adopted a 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%, and no inclusion of tone diacritization.

All the existing studies have no consideration for automatic correction of ASR errors involving user's acoustic irrational behavior. Moreso, existing studies didn't consider tone diacritization as it is needed especially in Yoruba language to avoid ambiguity. To improve speech recognition accuracy and system performance, acoustic nudging-based model is proposed in this paper. The research process reduces word error rate and sentence error rate at a very significant level when compared to other existing models.

3 The Acoustic Nudging-Based Model

Acoustic nudging is a concept adapted from behavioural science for aligning speech interactions that appears distorted. This study focused on large vocabulary, speaker independent, and tone-diacritized continuous standard Yorùbá speech recognition, which is the recognition of continuously spoken words with a comparatively high number of different words. The continuous word recognition system for standard Yorùbá based on the AN model is developed in three stages as follows: (i) System training with standard Yorùbá speech samples, (ii) System validation with standard Yorùbá speech samples, and (iii) 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 [6]. The training and recognition of parameters are based on an AN 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 (see Fig. 1).

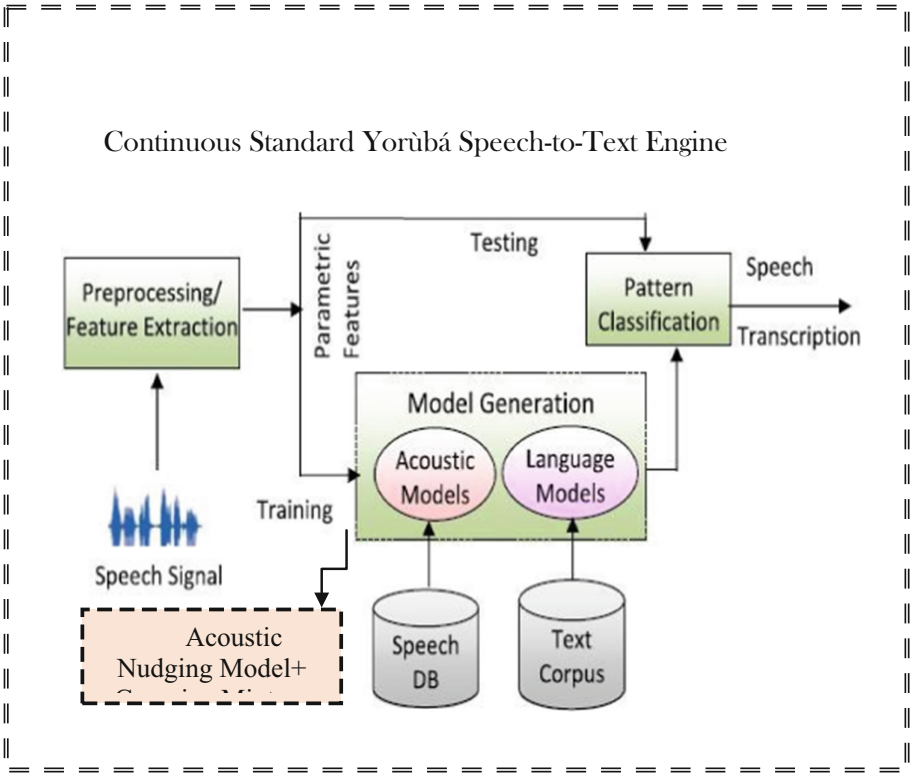

Fig. 1. The proposed acoustic nudging-based model

Acoustic Nudging Model: The entire process that makes up the acoustic nudging model is adapted from this study [20]. The acoustic nudging model is needed for automatic reformulation of the user's acoustic irrational behavior which involves tracking/monitoring, detecting, and reformulation of user's acoustic behavior in real-time.

Gaussian Mixture Model: The Gaussian Mixture Models (GMM) was used to address the low-resourced attribute of Yoruba Language [21]. Research has recorded GMM to be a tremendous success in the speech domain due to their high word detection accuracy ratio for language with low training data. GMM has been applied in many domains like artificial intelligence, image recognition and phoneme classification. 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 assumes that an M component mixture model with component weights $P(w_m)$ and its parameters θ_m represents the shape of each spectral. A univariate mixture model is represented in [21]. The developed ANGM modeling technique 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 purposes. The ANGM model is a universal approach that can deal with both discrete and continuous data.

The following process were engaged in the development of CSYSTT:

Data Collection Requirement. The development of the automatic speaker-independent large 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 a small noise in the background area. Also, the requirement for the text corpus development involves the speech data file collection as textual transcription.

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 text can have different pronunciations (such as 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, which can either be diacritized text or non-diacrtized text. The Standard Yorùbá diacritization process includes marking the Yorùbá letters using the orthographic symbols called diacritics or tone marks (See Fig. 2). Using non-diacritized texts poses a challenge to Yorùbá automatic speech recognition as missing tone/short vowels leads

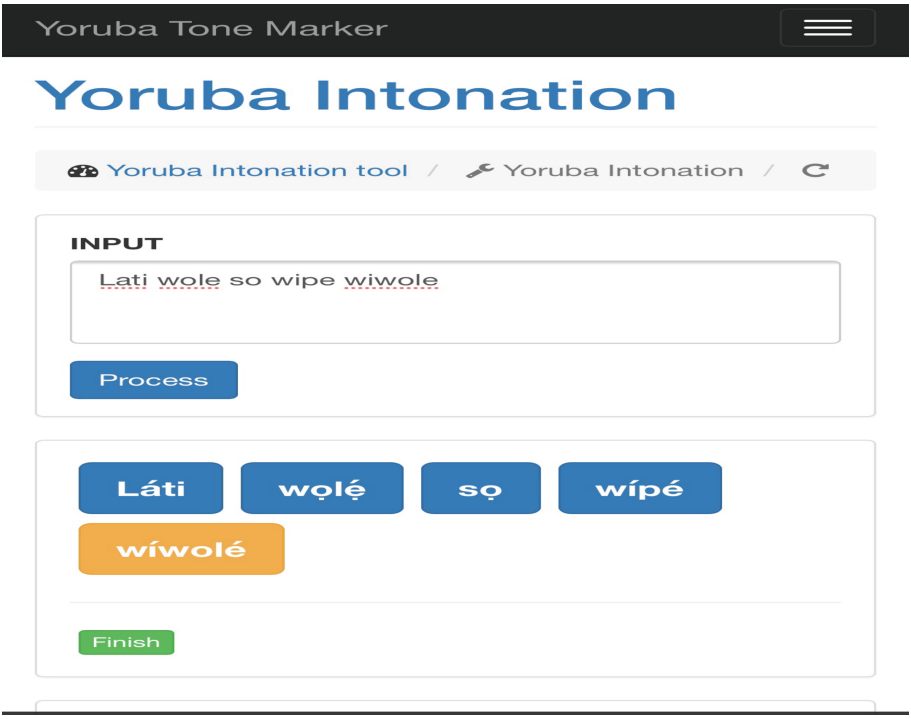

Fig. 2. A Semi-Supervised Standard Yorùbá Tone Diacritization Process

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.

In *Yorùbá* language, for diacritization process, having deep knowledge of syllables is essential and the allowed syllables in *Yorùbá* language are V, CV, and N", where V represents vowel sound, CV represents Consonant-Vowel sound and N represents Syllabic Nasal. Furthermore, *Yorùbá* language contains twenty-six '26' letters (consonants and vowels), seven '7' oral vowel sounds, and eighteen '18' consonant sounds. One of the consonant sounds is a digraph denoted as "D" and the rest as "C" with three '3' upper tonal signs (high, mid and low: á, a, and à) and under vowel sign (o, ọ) to distinguish words with the same spellings but different meaning and pronunciation. They are known as the consonants – sounds and vowel-sounds given below;

Consonant-sounds

b, d, f, g, gb, j, k, l, m, n, p, r, s, ṣ, t, u, w, y

Vowel-sounds

Oral-vowel:	a,	e,	ẹ,	i,	o	ọ	u
Nasal vowel:	an	ẹn	in	ọn	u		

Recorded Speech Signal (Speech Database Development). The speech corpus was created through data collection by training and validating and testing Yorùbá speech samples. The collection of data to develop the speech corpus is based on very large vocabulary for continuous speech recognition. A very large vocabulary generally implies 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 (such as single word utterances, multiple word utterances, long sentences, etc) one after the other in different record notes. The data was further divided into training, validation and testing dataset (See Table 1). Moreso, for the acoustic nudging model process, 16 speech data were collected based on a normal, angry, panicked, sore throat, and stressed acoustic behavior from different age brackets for males and females. The speakers are from different states in Nigeria, consisting of female and male speakers of a large variety of ages, i.e., both children and adults. The dataset was gathered from different social apps such as Facebook, WhatsApp, Messenger, Twitter, as the speakers recorded their voices in short mp3 files.

Table 1. 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

The data gathered from different locations; each location consisted of varying noises. Some voice samples embedded low/noise room recordings, while some were affected by outdoor noises (such as 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 noise-free voice samples were given as a batch of different voice notes by each speaker. (See Table 2).

Table 2. System parameters

Speaking mode	Continuous words
Sampling Rate	44 kHz
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
Language	Standard Yorùbá

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 a semi-supervised diacritizer. The full vocalization of the Yorùbá script was 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 steps include the following: (i) Entering Standard Yorùbá text, one paragraph at a time, (ii) 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. Table 3 contains standard Yoruba continuous speech sample Dataset and its English version.

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	Káàb`ọ j`ọw´ọ fí ímeèlì àti `ọr`ọ ìgbaniwólé rẹ
2	I want to buy men's watches	Mo fẹ rà aago àwọn okúnrin
3	I want to buy women's shoes	Mo fẹ rà bàtà àwọn obìnrin
4	Diving Watch	Ìluw`ẹ´ẹ 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. All the data collected must be in proper order. After all the speech samples were processed, it was necessary 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. Thereafter, the speaker's name was added to each speech sample and each speech sample was then divided to put each word is in one separate file.

Pre-processing/Feature Extraction. As the data were gathered from different locations, it followed with 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 to minimize this problem, it is essential to create a high-end, accurate, and robust voice system where noise and silence removal is required. 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 the 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 were converted into a wav format to execute the feature extraction algorithm. This study makes use of the Mel-frequency cepstral coefficient (MFCC). MFCC in speech research has been known as one feature extraction algorithm that works well with continuous speech recognition tasks for finding features compared to other feature extraction algorithms [19]. 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].

Acoustic Reformulation and Modeling. In this study, the statistical representation adopted is Acoustic Nudging (AN) model and the Gaussian Mixture Model (GMM). The ANGM (GMM + AN) Model was used in statistically representing the relationships between the speech signals and the language phonemes. These models were combined 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, to minimize the error rate embedded in speech data and at the same time, address the issue of low-resourced attributes of Yoruba Language.

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 was created after a deep study of Standard Yorùbá phonetics and also, different rules were used in pronouncing the words. Multiple entries are entered for words that are diacritized due to their homonym's 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 acts as an intermediary between the acoustic model and the language model to achieve a good recognition result.

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	
àmì	a mi	meanings	
àwọ	a woo		
àwọn	a wo n	.	
àsàyàn	a sa yan	.	
baagì	ba gi	.	
báyìí	ba yi	.	
bẹ́ẹ̀ bẹ́ẹ̀	bee be		
bèèrè	beee re		
èka	e ka	Yan – ya a un	

4 Result and Discussion

To facilitate comparison with previous results, the ANGM model for the CSYSTT system was evaluated based on Word Error Rate (WER). The WER was calculated using Eq. 1:

$$\text{WER} = 100 \times \frac{\text{Sw} + \text{Ds} + \text{ls}}{\text{N}} \tag{1}$$

$$\text{Where N} = \text{D} + \text{S} + \text{C}$$

Also, the Word Recognition Rate (WRR) metric is a complement of the Word Error Rate (WER) given by Eqs. 1 and 2:

$$WRR = WER^{1} \tag{2}$$

$$\text{Accuracy}(WRR) = (1 - \text{WER}) * 100\% \tag{3}$$

The ANGM model achieved a word error rate (WER) average score of 4.723% and mean accuracy (WRR) of 95.277. The best performance in terms of WER using 15 speakers was achieved using 128 MFCCs with 2.453% from speaker 10 (female adult) (see Table 5).

Table 5. 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

(*continued*)

Table 5. (*continued*)

S/N	Speakers	No of speech files	Mean accuracy (WRR)%	WER %
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**

Comparison with other Existing Models. To evaluate the efficacy of the ANGM model for the CSYSTT system, this study compares the performance of the ANGM model with existing models that have been popularly used in literature with the application as either single models or hybrid models such as GMM, GMM-HMM, CNN, GMM-CNN and 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. In the results presented in Table 6 and Fig. 4, the WER of each model was obtained by taking the average of each continuous Yorùbá speech sentence from the 2300 test dataset using different speakers with different age brackets conducted under a natural environment where there is noise interference.

Table 6. 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 k`ẹk´ẹ ìgb´ẹru	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 wẹwẹ kanfasi	3.623	5.256	4.701	4.813	4.276	5.600
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

(*continued*)

Table 6. (*continued*)

S/N	Sentences used	ANGM	GMM	GMM-HMM	CNN	GMM-CNN	DNN
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**

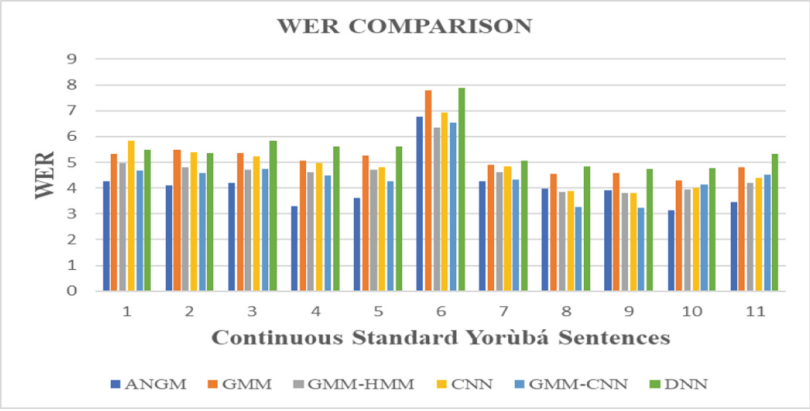

Fig. 3. WER comparison chart

The WER for ANGM, GMM, GMM-HMM, CNN, GMM-CNN and DNN models were estimated to be 4.091%, 5.219%, 4.596%, 4.915%, 4.436% and 5.503% respectively. The WER decrease was obtained between two systems and analyzed using Eq. 4.

$$\%\text{Decrease} = \left| \frac{\text{WER}_{p2} - \text{WER}_{p1}}{\text{WER}_{p1}} \right| \times 100 \tag{4}$$

The average CSYSTT performance which is based on the ANGM model demonstrates a WER decrease of GMM (1.1%), GMM-HMM (0.5%), CNN (0.8%), and DNN (1.4%). when compared with previous models. The results of the experiment obtained in Tables 4 and 5 are compared with a column chart in Fig. 4. In this study, WER comparison was also made based on existing literature studies for under-resourced languages [22–24] with this present study having different WER ranges of 11.27%, 10.07%, 5.48%, 5.45%, 4.92%, etc. The 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, the ANGM model has a significantly lower average WER of 4.091% using 2300 test sentences.

5 Conclusion and Future Works

After the development of the indigenous vocabulary Continuous Standard Yorùbá Speech-to-Text reformulation engine for seamless communication, the system was evaluated using WER. The mean Word Error Rate (WER) is 4.723%. 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 an application as either single models or hybrid models using 11 CSY test datasets. 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 the ANGM model demonstrates a WER decrease when compared with previous models- GMM (1.1%), GMM-HMM (0.5%), CNN (0.8%), and DNN (1.4%).

Existing African language speech recognition also achieved a WER range of 9.48%, 14.83%, and 18%, and when compared with this current study, the ANGM model has a significantly lower average WER of 4.091%. Our study caters for the observed limitations of the existing speech domain through the development of a large vocabulary Continuous Standard Yorùbá Speech-to-Text Engine as opposed to the existing Isolated Yorùbá Speech-to-Text Engine with consideration of tone diacritization and an Acoustic Nudging-based Gaussian Mixture (ANGM) Model. The essence is 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 a platform to host the applications in Media, [25–27]. This study provides future research direction including 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

References

1. Gaikwad, S.K., Gawali, B.W., Yannawar, P.: A review on speech recognition technique. Int. J. Comput. Appl. **10**(3), 16–24 (2010). https://doi.org/10.5120/1462-1976
2. Nwakanma, P.C., Ibe, R.C.: Globalization and economic growth. An econometric dimension drawing evidence from Nigeria. Int. Rev. Manage. Bus. Res. **3**(2), 771 (2014)
3. Deborah, N.O., Rhoda, I.A., Williams, O.A.: Development of a Mobile Tourist Assistance for a Local Language. **6**(1), 5–9 (2017). https://doi.org/10.5923/j.tourism.20170601.02
4. Prachi, U., Bhope, G.: Voice Based Collaborative Banking (2015)
5. Akintola, A., Ibiyemi, T.: Machine to Man Communication in Yorùbá Language. Annal. Comput. Sci. Ser. **15**(2) (2017)
6. Sebastian, B.W., 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, V.B., Besacier, L.: Automatic speech recognition for under-resourced languages: application to Vietnamese language. IEEE Trans. Audio Speech Lang. Process. **17**(8), 1471–1482 (2009). https://doi.org/10.1109/TASL.2009.2021723
11. Saksamudre, K.S., Shrishrimal, P.P., Deshmukh, R.R.: A review on different approaches for speech recognition system. Int. J. Comput. Appl. **115**(22), 23–28 (2015). 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, N.T., Kraus, F., Schultz, T.: Rapid building of an ASR system for under-resourced languages based on multilingual unsupervised training. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 3145–3148, August 2017
14. Vyas, J.V.: Study of Speech Recognition Technology and its Significance in Human-Machine Interface. **3**(10), 416–422 (2017)
15. Mohanty, S., Swain, B.K.: Markov Model Based Oriya Isolated Speech Recognizer-An Emerging Solution for Visually Impaired Students in School and Public Examination. Special Issue IJCCT **2**(2,3,4), 107–111 (2010). https://www.researchgate.net/publication/266268918
16. Ltaief, A., Ben, E.Y., Graja, M., Belguith, L.H.: Automatic speech recognition for Tunisian dialect, pp. 1–8 (2016)
17. Das, P., Acharjee, K., Das, P., Prasad, V.: Voice Recognition System : Speech-To-Text, July 2016
18. Adetunmbi, O.A., Obe, O.O., Iyanda, J.N.: Development of Standard Yorùbá speech-to-text system using HTK. Int. J. Speech Technol. **19**(4), 929–944 (2016). https://doi.org/10.1007/s10772-016-9380-2
19. Laleye, F.A., Besacier, L., Ezin, E.C., 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), pp. 477–482. IEEE, September 2017
20. Galvan, R.F., Barranco, V., Galvan, J.C., Batlle, S., Fajardo, F., García, S.: We are IntechOpen , the world' s leading publisher of Open Access books Built by scientists, for scientists TOP 1%. Intech, i(tourism) **13** (2016). http://dx.doi.org/https://doi.org/10.5772/57353

21. 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. J. Comput. Sci. **16**. 17ss31–1741. https://doi.org/10.3844/jcssp.2020.1731.1741
22. Beserra, A.A.V., Silva, W.L.S., Serra, G.L.D.O.: A GMM/CPSO speech recognition system. In: IEEE International Symposium on Industrial Electronics, 2015-September (December 2018), 26–31. https://doi.org/10.1109/ISIE.2015.7281438
23. Stuttle, M.N.: A Gaussian Mixture Model Spectral Representation for Speech Recognition, p. 163, July 2003
24. 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., pp. 185–188 (2016)
25. Azeta, A., Da-Omiete, A.I., Azeta, I.V., Emmanuel, O.I.,. Fatinikun, D.O., Ekpunobi, E.: Implementing a medical record system with biometrics authentication in E-Health**.** In: IEEE AFRICON*:* Science, Technology and Innovation for Africa, AFRICON2017, pp. 18–20. September 2017; The Avenue V&A Waterfront Cape Town South Africa
26. Ajayi, L.K., et al.: Enhancing the low adoption rate of M-commerce in Nigeria through Yorùbá voice technology. In: Abraham, A., Hanne, T., Castillo, O., Gandhi, N., Nogueira Rios, T., Hong, T.-P. (eds.) HIS 2020. AISC, vol. 1375, pp. 516–524. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-73050-5_52
27. Azeta, A.A., Ayo, C.K., Atayero, A.A., 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., pp. 92–108 (2009)