A Yorùbá Language Automatic Speech Recognition System using Deep Learning Approach

Ogunjobi Olumide Jospeh
Department of Computer Science
Federal University of Technology, Akure, Nigeria

Agbonifo Oluwatoyin Catherine
ocagbonifo@futa.edu.ng
Federal University of Technology, Akure, Nigeria

Akinwonmi Akintoba E.
Department of Computer Science
Federal University of Technology, Akure, Nigeria

ABSTRACT
In this paper, a Yorùbá Language Automatic Speech Recognition System capable of recognizing words spoken by the users based on preprocessed stored data was designed and implemented. Dataset from Data Science and Computational Research Laboratory, Federal University of Technology, Akure, Nigeria, were made used of. Speech feature sequences were extracted using Mel-frequency cepstral coefficients (MFCC) technique. Kaggle environment was deployed, where python programming language was used to implement the Transformer and LSTM (Long Short-Term Memory) models. The preprocessed data were fed into both models for training, both models were tested and evaluated across the following standard evaluation metrics WER, MER and CER. The results obtained from both models shows a promising approach that could be adopted for Yorùbá Language speech recognition system. The research work describes ASR for standard Yorúbà language. The source language and target language is Yorùbá language, with focus on speaking in Yorùbá language (voice data) and the output in Yorùbá language text form (text data).

General Terms
Artificial Intelligence, Natural Language Processing, Speech Recognition, Deep Learning, Pattern Recognition, Machine Learning.

Keywords
Automatic Speech Recognition (ASR), Mel-Frequency Cepstral Coefficients (MFCC), Speech-to-Text, Low-Resource Language, Deep Learning.

1. INTRODUCTION
The basic means by which humans communicate are via signs/gestures and languages. Communication via languages could be oral or written, which is important for transferring ideas and feelings, interaction within society, to educate, entertainment, among others. An effective means of communication is vital for economy growth and development; it is an important aspect of life because it is only through exchange of ideas and cooperation that a society can grow and develop. However, communication between people of different languages could sometimes serve as a barrier especially when both speakers are unable to comprehend one another or misinterpretation from human translators. Human translators are often required to solve communication barriers between people of different languages, in a situation where perhaps no feasible human translator is available, a machine translator will be most appropriate which is also capable of eliminating misinterpretation. In recent times, people communicate with computers via use of Natural Language Processing (NLP) applications. Computers serve as the translating device between people of different languages to enable effective communication. Once effective communication is established, people communicating will be able to have better understanding of each other and vital information could be easily shared across.

1.1 Natural Language Processing
NLP refers to the branch of computer science and more specifically, the branch of artificial intelligence, concerned with giving computers the ability to understand text and spoken words in much the same way human beings can [1]. It can also be referred to as the automatic manipulation of natural language, like speech and text, by software [2]. NLP combines computational linguistics rule-based modeling of human language with statistical, machine learning, and deep learning models. Together, these technologies enable computers to process human language in the form of text or voice data and to 'understand' its full meaning, complete with the speaker or writer's intent and sentiment [3]. NLP drives computer programs that translate text from one language to another, respond to spoken commands, and summarize large volumes of text rapidly, even in real time. Several NLP tasks break down human text and voice data in ways that help the computer make sense of what it's ingesting. Some of these tasks include the following: Speech recognition, Part of speech tagging, Word sense disambiguation, Named entity recognition, Co-reference resolution, Sentiment analysis and Natural language generation [1].

1.2 Automatic Speech Recognition
Automatic Speech Recognition (ASR) or speech-to-text system identifies spoken words in speech and converts them to written text. ASR refers to a computer aided system that is designed to perform (automate) tasks which are traditionally done by a human. In the process, it converts human speech to text, interprets the recognized text, and then speaks back a response to the user [4]. ASR systems extract acoustic signal characteristics from speech and determine the words in it by pattern matching. Acoustic and language models are used in developing ASR systems. ASR can also be referred to as the task of reliably converting voice data into text data. Speech recognition is required for any application that follows voice commands or answers spoken questions.

1.3 Basic Steps in ASR
The basic sequence of events that make any ASR software, regardless of its sophistication which is to do with pick up and breaking down words for analysis and response goes as follows:

(a). The user speaks to the software via an audio feed

(b). The device spoken to creates wave files of the user's words

(c). The wave file is cleaned by removing background noise and normalizing volume

(d). The resulting filtered wave form is then broken down into what are called phonemes.

(e) Each phoneme appears like a chain link and analyses them in sequence, starting from the first phoneme.

(f). The ASR, having "understood" the words, can now respond in a meaningful way.

1.4 Types of ASR

The two (2) main types of Automatic Speech Recognition software variants are Directed Dialogue Conversations (DDCs) and Natural Language Conversations (NLCs) [5]. Directed Dialogue Conversations are the much simpler version of ASR at work and consist of machine interfaces, which tell verbally to respond with a specific word from a limited list of choices, thus the user's response is narrowly defined. Automated telephone banking and other customer service interfaces commonly use directed dialogue ASR software. However, NLC are the much more sophisticated variants of ASR and instead of heavily limited menus of words, they try to simulate real conversation by allowing users to use an open-ended chat format with them. The Siri interface on the iPhone is a highly advanced example of these systems. NLC is a variant of ASR that comes the closest to allowing real conversation between people and machine intelligence and though it still has a long way to go before reaching an apex of development, for other systems used in business and advanced technology contexts. NLC being the future direction of ASR technology is much more important than directed dialogue in the development of speech recognition systems. NLC is designed to loosely simulate how humans themselves comprehend speech and respond accordingly. Natural language system is designed to react to a much smaller list of selected "tagged" keywords that gives context to longer requests. Thus, using these contextual clues, the system can much more quickly narrow down exactly what the user is saying to it and find out which words are being used so that it can adequately respond.

1.5 Usefulness of ASR

ASR technology could be useful in lots of ways such as;

(a) communication between a well abled body person and a deaf person. People who are having hearing impediments, ASR advancements can improve their ability to complete their daily tasks without obstruction,

(b) it cuts back on misspelled words that some typists may suffer from when typing,

(c) it cuts down on the amount of time editing and fixing spelling corrections,

(d) it is an advantage to people who may suffer from disabilities that affect their writing ability but can use their speech to create text on [6].

In recent times, lots of research work is ongoing in the development of speech recognition systems with better accuracy. Many tools like Rapid Language Adaptation Toolkit (RLAT), Hidden Markov Model Toolkit (HTK), Carnegie Mellon University Sphinx (CMU-Sphinx) and host of others have been developed for the purpose of speech recognition [7]. Currently, many ASR systems are available mainly for very few languages.

Languages addressed are mainly those with available digital resources. However, most of the languages in the world lack such technologies, and research in the area is quite few. Recently, many localization projects were being undertaken for African languages, but not adequate because of the lack of linguistic resources, absence of similar works in the area [8].

1.6 Yorùbá Language Coverage

Yorùbá is considered the second most spoken native language in Nigeria, besides Hausa language. The number of Yoruba speakers is stated as roughly fifty (50) million, plus about two (2) million second-language speakers [9]. It is spoken as well in neighboring countries of the Republic of Benin, Togo, Sierra Leone and Ghana [10].

However, the language has been underrepresented in the digital space due to lack of digital data and poor language tools. Hence, there is need to make available digital data and language tools for Yorùbá language in other for the language to be well represented in the digital space.

2. REVIEW OF RELATED WORKS

In the research of [11] developed an Automatic Speech Recognition for Telephone Voice Dialing in Yorùbá. Their study was motivated due to media interaction is grossly not user friendly and often risky and life threaten in some applications, such as driving and phone dialing. The method involved Audio files collection, pre-processing, feature extraction and recognition. The preprocessing phase includes voice activity detection to remove inter-word silence periods, pre-emphasis files, frame blocking and windowing. The feature extraction, MFCCs feature were extracted, while vector quantization and each clean distance were used for data compression and recognition respectively. They developed a user's friendly human computer interaction based on speech recognition for telephone autodialing in Yorùbá. The system is limited to only recognizing texts with existing templates in the system database.

[7] presented Development of Standard Yorùbá speech-to-text system using HTK. Their study was motivated to provide a system that will be adopted for Standard Yoruba continuous speech recognition system that could also be useable for the foreign speaker. The methods employed involve the use of Praat software which was used to record, analyze and annotate words. Hidden Markov Model Toolkit algorithm was used for Speech to Text conversion, with the help of Cygwin to make the window behave as Linux. Java programming language was also used to build a graphical user interface to make the system accessible and interactive. The system was tested to identify the utterances within the vocabulary, with a moderate accuracy of 76% and 84% respectively for the set bi and tri.

[12] presented Speech Recognition Using Deep Learning Algorithms. Their study was motivated to address high viability of speech signal during the conversion from speech to text. The methods employed involved use of Deep Learning Algorithms. TIMIT acoustic-phonetic continuous speech corpus dataset was used for performance evaluation. They were established as better models than GMM-HMM, DNN-HMM.

Yorùbá Gender Recognition from Speech was developed using Attention-based BiLSTM [13]. Their study was motivated due to availability of gender recognition systems for well-resource languages like English, while for African languages like Yorùbá are not available. The method employed involved two phases, namely, the training and recognition phase. Input speech was preprocessed by removing noise and reducing the dimensionality of the data. Thereafter, acoustic features were extracted from the processed data. These features were fed into BiLSTM for training and the trained model was tested using

unlabeled gender speech to determine its gender. They established the ability to detect Gender from speech via BiLSTM model.

3. METHODOLOGY

The system methodology of the ASR model for Yorùbá language using Speech Transformer Networks and LSTM encoder-decoder model is shown in Figure 1. The conceptual diagram presented in Figure 1 consists of four (4) phases. These include Data Collection, Pre-processing, Model Training and Model Evaluation phase.

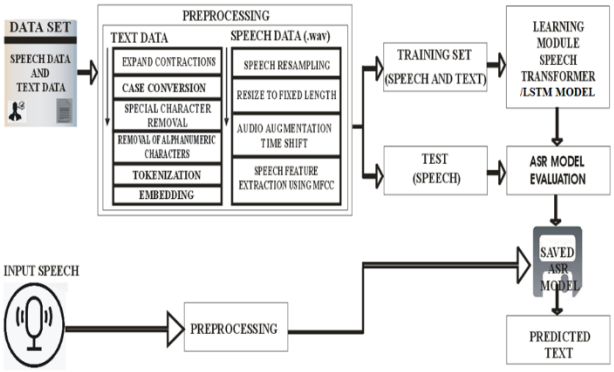

Figure 1: ASR model for Yorùbá language speech

3.1 Data Collection

In the data collection phase, data was gotten from Speech and Language Technology Research Laboratory, Federal University of Technology Akure, Ondo State, Nigeria. The obtained data which is composed of one thousand (1,000) annotated utterances with their corresponding text, were subjected to some preprocessing steps.

3.2 Dataset Pre-processing

In most cases, raw data (speech and text) contains noise, redundant and irrelevant features and if fed into any of the model, it could bring about malfunctioning of the system. Therefore, this research work handled unprocessed text data by means of contraction expansion, case conversion, removal of special and alpha numeric characters, tokenization and embedding, while unprocessed speech data by means of speech resampling, resize to fixed length, audio augmentation time shift and speech feature extraction using MFCC.

3.3 Model Training

In this research work two models were trained namely Transformer model and Long Short-Term Memory (LSTM) encoder-decoder model. In the model training, input data $X = x_1, x_2, x_3 \ldots x_T$, which is a sequence of extracted speech features, was passed into Transformer and LSTM encoder-decoder Models concurrently.

3.3.1 Speech-Transformer Model

The Speech-Transformer is a no-recurrence sequence to sequence model, which transforms speech feature sequence to the corresponding character sequence. Speech-Transformer has two parts, the encoder on the left-hand side and the decoder on the right-hand side, as depicted in Figure 2.

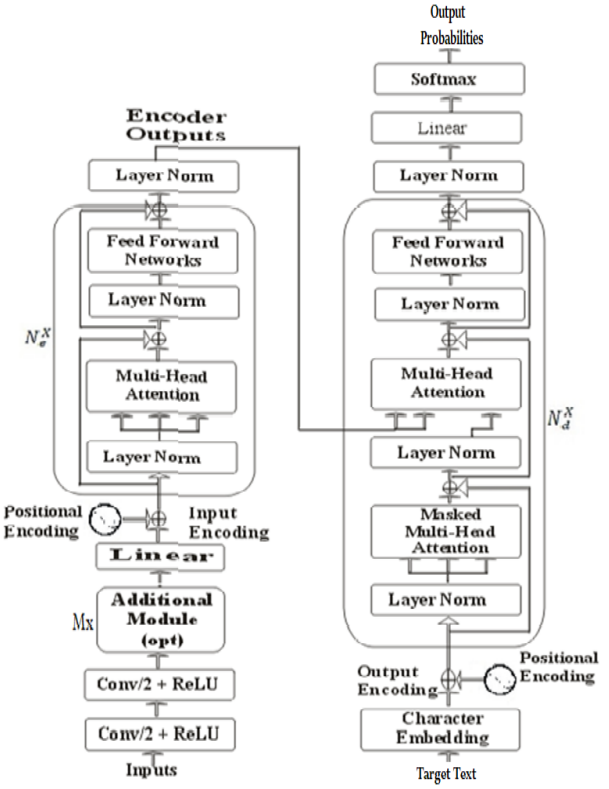

Figure 2: Speech-Transformer model [14]

The speech transformer model is an attention-based encoder-decoder architecture. The encoder and decoder of the transformer model are composed of multihead attention in equation (1) and feed forward neural network (FNN) in equation (3). Multihead is a concatenation of self-attention captured in equation (2) which helps the model to focus on important words in the given input.

$$MHA(Q,K,V) = concat\left(h_1 h_2 \ldots h_h\right) W^0 \quad (1)$$

where $MHA(Q,K,V)$ is a multihead attention; and $Q, K, and\ V$ represent Query, Key and Value respectively, which are used to produce attention scores.

$$h_{i=Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

More so, FNN ensures better contextual representation.

$$FFN(x) = ReLU\left(xW_1 + b_1\right)W_2 + b_2 \quad (3)$$

where x is the input data, $W_1 \in R^{d_{model} \times d_{ff}}$, $W_2 \in R^{d_{ff} \times d_{model}}$ are the weights and $b_1 \in R^{d_{ff}}, b_2 \in R^{d_{model}}$ are the biases.

The output of the encoder is then fed into the decoder block. The decoder block of the transformer gives the sentence representation of the encoder output using the masked multihead and the FFN presented in equation (1) and (3) respectively.

3.3.2 LSTM Encoder-Decoder Model

In the second model, LSTM encoder reads each input $X = x_1, x_2, x_3 \ldots x_T$ sequentially. At each time step, the

hidden state h is updated using the previous hidden state and the current input as presented in equation 4.

$$h_t = f\left(W^{(hh)}\, h_{t-1} + W^{(hx)}\, x_t\right) \quad (4)$$

where h_t is the hidden state, x_t is the current input, $W^{(hh)}$ is the weight at the current neuron, f is function, h_{t-1} is previous hidden state, $W^{(hx)}$ is weight of input neuron and x_t is the current input.

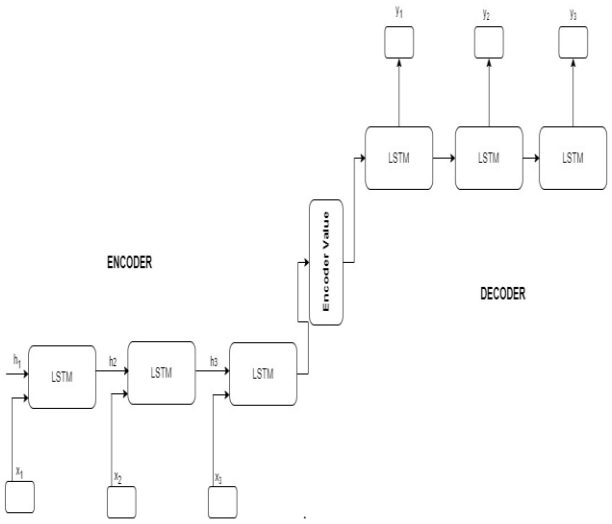

Figure 3: LSTM model

The output of the encoder h_t are vector representations which aims to encapsulate the information for all input elements to help the decoder make accurate predictions.

The decoder converts the hidden vector into the output sequence. The decoder consists of stack of several LSTM units as shown in Figure 3, where each unit predicts an output y_t at a time step t. Each LSTM unit accepts a hidden state from the previous unit h_{t-1}, it also accepts previous layer output y_{t-1} and produces an output as well as its own hidden states. The hidden state of the decoder is computed as presented below.

$$h_t = f\left(W^{(hh)}\, h_{t-1}\right) \quad (5)$$

Then softmax activation function is made use of at the output layer to produce the probability distribution from a vector of values with the target class of high probability. The softmax is calculated as presented in equation (6).

$$y_t = \text{softmax}\,(W^S\, h_t) \quad (6)$$

where h_t is the hidden state at the current time step and W^S is the weight. Finally, the decoder is trained to start generating the output sequence depending on the information encoded by the encoder.

3.4 Model Evaluation

The following performance metrics were made used of to evaluate both models Word Error Rate, Match Error Rate and Character Error Rate

3.4.1 Word Error Rate (WER)

WER was used to evaluate the model in this research work. WER tells how many words were logged incorrectly by the system It is presented as:

$$WER=100\times((S+D+I))/N\% \quad (7)$$

where S represents the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the actual transcription. A lower WER in this research indicates better transcription accuracy.

3.4.2 Match Error Rate (MER)

MER was used to evaluate match word sequences. It measures the errors based on matched words between the actual and the predicted transcription.

$$MER=(S+D+I/(N=H+S+D+I)) = 1-H/N \quad (8)$$

where H is the total number of word hits (Matched words)

3.4.3 Character Error Rate (CER)

Similarly to WER, CER measures the percentage of errors but of characters in the transcribed text compared to the actual text. It is based on Levenshtein distance. It is presented as:

$$CER =100\times((S+D+I))/N \quad (9)$$

where S represents the number of wrong characters, D is the number of missing characters, I is the number of extra characters, and N is the total number of characters in the actual transcription.

4. RESULTS AND DISCUSSION

This section discusses the model's performance based on ASR standard performance metrics, such as WER, MER and CER. The dataset was split into 50% for training and 50% for testing. The model was trained for 2000 epochs.

4.1 Speech Transformer ASR Result

The Figure 4 shows consistent improvement in the Yoruba ASR model as training progresses from 500 to 2000 steps. The WER decreases from approximately 0.19 to 0.165, indicating better word-level recognition. The MER shows the most substantial improvement, dropping from about 0.295 to 0.180, reflecting improved sequence alignment and reduced insertion, deletion, and substitution errors. Similarly, the CER declines steadily from around 0.175 to 0.142, demonstrating enhanced character-level accuracy, which is particularly important for tonal languages like Yoruba.

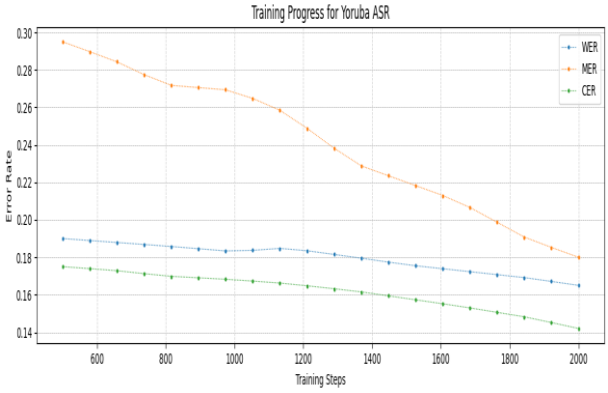

Figure 4: Speech Transformer ASR Training Progress

4.2 LSTM ASR Result

The Figure 5 shows steady but modest performance gains of the Yoruba ASR (LSTM) model as training increases from 500 to 2000 steps. The WER decreases from approximately 0.225 to 0.205, indicating gradual improvement in word-level recognition. The MER shows a clearer reduction, dropping from about 0.340 to 0.275, reflecting improved sequence alignment over time. Similarly, the CER declines from roughly 0.205 to 0.178, demonstrating better character-level accuracy, which is important for handling Yoruba's tonal and orthographic features.

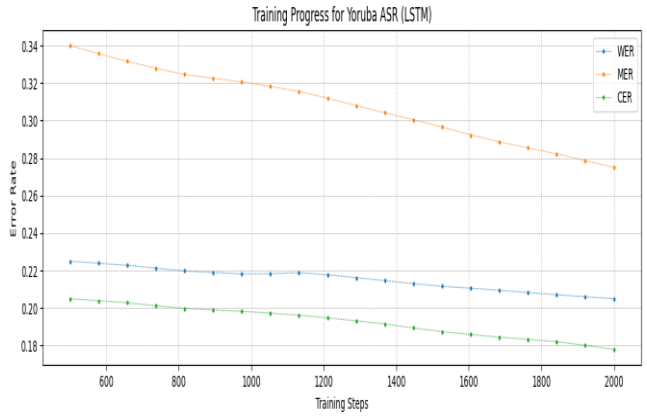

Figure 5: LSTM ASR Training Progress

4.3 Comparison between Speech Transformer Model and LSTM Model

The comparison in Table 1 shows that the Speech Transformer-based ASR consistently outperforms the LSTM-based ASR across all evaluation metrics. The Transformer model achieves lower error rates with an Avg. WER of 0.1796, Avg. MER of 0.2408, and Avg. CER of 0.1616, compared to the LSTM model's Avg. WER of 0.2154, Avg. MER of 0.3082, and Avg. CER of 0.1927. These results indicate that the Transformer-based ASR provides more accurate word, sequence, and character-level recognition, making it a stronger foundation for downstream speech-to-speech translation tasks.

Table 1: Average ASR Performance across Training Steps

Model	Avg. WER	Avg. MER	Avg. CER
Transformer-based ASR	0.1796	0.2408	0.1616
LSTM-based ASR	0.2154	0.3082	0.1927

Table 2 quantifies the performance gains of the Transformer-based ASR over the LSTM-based ASR. The results indicate that the Transformer-based ASR model achieves clear and consistent improvements over the LSTM-based ASR across all evaluation metrics. The WER is reduced by 0.0359, corresponding to a 16.7% improvement, showing better word-level recognition. The largest gain is observed in the MER, which decreases by 0.0674, representing a 21.9% improvement, highlighting significantly improved sequence alignment. Additionally, the CER is reduced by 0.0311, yielding a 16.1% improvement, indicating enhanced character-level accuracy.

Table 2: Relative Performance Gain (Transformer vs LSTM)

Metric	Absolute Reduction	Relative Improvement (%)
WER	0.0359	16.7
MER	0.0674	21.9
CER	0.0311	16.1

The results demonstrate that both models improved with increasing training iterations; however, the Transformer model converged faster and consistently produced lower error rates than the LSTM model. This superiority can be attributed to the self-attention mechanism of the Transformer, which effectively captures long-range dependencies in speech sequences without suffering from the vanishing gradient problems common in recurrent architectures. The reduction in CER is particularly significant for Yorùbá because tonal marks and diacritics play important roles in semantic interpretation. Lower CER values therefore indicate better preservation of linguistic information.

Table 3: Final Performance of the Transformer vs LSTM

Model	Final WER	Final MER	Final CER
Transformer	0.165	0.180	0.142
LSTM	0.205	0.275	0.178

5. CONCLUSION

This study presented the development of an Automatic Speech Recognition system for the Yorùbá language using deep learning approaches based on Speech Transformer and LSTM encoder-decoder architectures. Speech features were extracted using Mel-Frequency Cepstral Coefficients (MFCC), and both models were trained and evaluated using Word Error Rate (WER), Match Error Rate (MER), and Character Error Rate (CER).

Experimental results demonstrated that the Speech Transformer model consistently outperformed the LSTM model, achieving lower error rates across all evaluation metrics. Specifically, the Transformer model attained average WER, MER, and CER values of 0.1796, 0.2408, and 0.1616 respectively, indicating superior capability in capturing long-range contextual dependencies within Yorùbá speech sequences.

The findings suggest that attention-based architectures provide a promising direction for low-resource African language speech recognition systems. The developed model can potentially be integrated into applications such as voice assistants, educational technologies, assistive systems for individuals with disabilities, and speech-driven translation platforms.

Future work will focus on evaluating the model using larger and more diverse Yorùbá speech corpora, incorporating multiple dialects and noisy environments, exploring pre-trained self-supervised speech representations such as wav2vec 2.0 and Whisper, and extending the framework toward end-to-end Yorùbá speech-to-speech translation systems.

6. ACKNOWLEDGMENTS

The authors sincerely appreciate the Federal University of Technology, Akure, Nigeria. The authors also acknowledge the valuable contributions of colleagues and researchers whose

insights and suggestions helped improve the quality of this work.

7. REFERENCES

[1] IBM Cloud Education (2020). Natural Language Processing (NLP). Retrieved from https://www.ibm.com/cloud/learn/natural-language-processing

[2] Brownlee, J. (2019). Deep learning for Natural Language Processing (NLP). Retrieved from https://machinelearningmastery.com/natural-langage-processing/

[3] IBM Cloud Education (2020). What is Speech Recognition?. Retrieved from https://www.ibm.com/topics/speech-recognition

[4] IGI Global (2021). What is ASR. Retrieved from https://www.igiglo.com/dictionary/asr/1566

[5] Zajechowski, M. (2021). Automatic Speech Recognition (ASR) Software: An Introduction. Retrieved from. https://usabilitygeek.com/automatic-speech-recognition-asr-software-an-introduction/

[6] Raut and Deoghare (2016). Automatic Speech Recognition and its Applications. IRJET, 03(05), 2368-2371

[7] Adetunmbi, O. A., Obe, O. O., and Iyanda, J. N. (2016). Development of Standard Yorùbá speech-to-text system using HTK. International Journal of Speech Technology, 19(4), 92-94.

[8] Abiola, O.B., Adeyemo O.A., Saka-Balogun O.Y., and Okesola F. (2020). A Web-Based Yorùbá to English Bilingual Lexicon for Building Technicians. International Journal of Advanced Trends in Computer Science and Engineering, 9(1), 793- 800

[9] Eberhard, D. M., Gary F. S., and Fennig, C. D. (2021). Ethnologue: Languages of the World. Twenty-fourth. Retrieved from. https://www.ethnologue.com/language/yor

[10] Okanlawon, J. (2016). An Analysis of the Yorùbá Language with English Phonetics, Phonology, Morphology and Syntax. Retrieved from. https://cos.northeastern.edu/wp-content/uploads/2018/09/Jolaade-Okanlawon-An-Analysis-of-Yorùbá-with-English.pdf

[11] Ibiyemi, T.S., and Akintola, A.G. (2012). Automatic Speech Recognition for Telephone Voice Dialing in Yorùbá. International Journal of Engineering Research & Technology (IJERT), 1(4), 1-6

[12] Zhang, Y., and Yzhang5, S. I. (2013). Speech Recognition Using Deep Learning Algorithms, pp. 1-5

[13] Modupe, I. A, Sefara, T. J, and Ojo, S.O. (2019). Yorùbá Gender Recognition from Speech using Attention-based BiLSTM. In Proceedings of the First International Workshop on NLP Solutions for Under Resourced Languages (NSURL) co-located with ICNLSP. Short papers, pp. 16-22

[14] Dong, L., Xu, S. and Xu, B. (2018). Speech-Transformer: A No-Recurrence Sequence-To-Sequence Model for Speech Recognition. International Conference on Acoustics, Speech and signal processing, pp. 5884-5888