FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 8 No. 6, December, 2024, pp 329 - 338
DOI: https://doi.org/10.33003/fjs-2024-0806-2890

SPEECH-TO-TEXT: A SECURED REAL-TIME LANGUAGE TRANSLATION PLATFORM FOR STUDENTS

***Anazia Eluemunor Kizito, Eti Erife Friday, Ovili Peter Henry and Ogbimi O. Francis**

Department of Information Systems and Technology, Delta State University of Science and Technology, Ozoro.

*Corresponding authors' email: kaymax07@yahoo.com

ABSTRACT

In order to establish effective communication and understanding among students regardless of language background, there is need to develop a common platform that will support this motive. This necessity has led to the emergence of Speech-to-Text (S-to-T) translation framework that enabling students with diverse languages to communicate directly without relying on intermediaries. English has become the foremost lingua franca in Nigeria, spoken widely across various ethnic groups. However, this continuous use of English language has affected the subsistence of indigenous Nigerian languages, leading to many children growing up unable to speak their native language. In modern, civilized societies, speech remains a primary and essential means of communication, allowing individuals to express a range of ideas from their minds through organized words, phrases, and sentences that follow grammatical rules. This work, Speech-to-Text: A secure real-time language translation platform for students, translates speech to English and Yoruba languages during chat. It was developed using ASP.Net with C# as the base technology. The model was developed with CSS, Bootstrap, JQuery, and JavaScript, ensuring responsiveness, while a secure SQL Server database repository supports data storage. The software is structured using the Object-Oriented Methodology (OOM). This platform presents a user-friendly and intuitive web interface that allows students of both English and Yoruba speakers to easily access and interact with other in real-time thereby bridging the communication gap between them.

Keywords: Boostrap, Google Cloud Translation, JavaScript, Hypertext Markup Language, Speech-to-Text Translation

INTRODUCTION

Speech is a core and widely used form of communication in contemporary societies, serving as the main method through which people naturally and effectively share information (Prachi and Bhope, 2015). The transformation of written text into spoken language is termed Text-to-Speech synthesis, as described by (Adekunle, Agbonifo and Olaniyan, 2020). Rather than hindering interaction, speech facilitates communication, allowing individuals to convey their feelings, thoughts, emotions, and beliefs. The field focused on developing communication between machines and humans is known as natural language processing. A key component within language processing method is speech recognition techniques, which translates natural spoken languages (voice) into text, fostering fluid communication and interaction between people and machines (Sanchit, Aniket and Tanvi, 2022).

Recently, urban residents have found it increasingly difficult to communicate in local dialects. Nevertheless, preserving and using these indigenous languages remains essential. In Nigeria, there are about 500 to 550 local languages and dialects, with Igbo, Hausa, and Yoruba as the three major languages (Benjamin and Eludiora, 2020). Nigeria's large population, around 234 million, and its robust economy enhance its linguistic diversity. Currently, approximately 80% of the population resides in urban areas, where English is predominantly spoken, while those in rural areas mainly use local languages and dialects and often experience communication barriers among peers (Akinwaleet al., 2015). Creating a website for translating documents, text, and real-time speech from English to Yoruba fulfills a crucial need for language accessibility and communication across communities. Translation tools support language learners in comprehending English and Yoruba texts, facilitate real-time communication between students and teachers from different language backgrounds and enhance multilingual education, cultural discussion and learning opportunities (Esan et al., 2020).

In the carried out by Ajibade and Eludiora (2021), they pointed out that most Yoruba speakers who lack English proficiency are unable to access the vast range of valuable information available online. Although existing translation tools have made major progress, a specialized platform developed to the needs of Yoruba speakers translating to and from English is still required. Developing a comprehensive web-based solution capable of translating typed text, uploaded documents, and real-time speech between English and Yoruba and vice versa would address this need (Oloruntoyin, 2014 and Adigun, et al, 2024).

This work examines the comprehensive development of speech-to-text systems, tracing their historical evolution, core methodologies, and modern applications. Speech-to-text technologies have changed from basic systems with limited vocabulary to sophisticated models capable of handling various accents and languages, marking a rapid and transformative journey Totare et al., 2023). According to Vasilakes, Zhou and Zhang (2020), speech recognition or speech-to-text refers to a model or program's ability to detect spoken language and convert it into texts that are readable to human. While basic speech recognition software is limited to recognizing a set vocabulary and requires clear speech, more advanced systems can handle natural speech, diverse accents, and multiple languages (Oise and Konyeha, 2024). The development of speech recognition technology draws upon extensive research across computer science, linguistics, and computer engineering. Today, many devices and text-centric applications feature speech recognition, allowing for more convenient, hands-free operation (Padmane, 2022). Speech recognition stands as an interdisciplinary subfield, bridging computer science and computational linguistics, with a focus on designing systems that accurately recognize and transcribe spoken language.

This computer field is referred to also as computer speech recognition which draws its principles from computer engineering, information and communication technology and linguistics (Adebara, Abdul-Mageed and Silfverberg, 2022). Speech recognition significantly contributes to the advancement of human-computer interaction, working in tandem with the computer reverse process theory, speech synthesis (Chauhan et al., 2016). Gaussian Mixture Models are generally used in acoustic modeling within speech recognition and recent studies have focused on improving GMMs through techniques like speaker adaptation and the integration of deep learning techniques to enhance the modeling of acoustic features (Ren et al., 2019). . Convolutional neural networks, initially widely used in image processing, have been successfully adapted in the development of speech recognition models. Recent studies demonstrate Convolutional neural network's ability to capture hierarchical features from spectrograms and other acoustic representations, improving overall recognition accuracy (Graves et al., 2013).

In the carried out by Xiao and Zhu (2023), it was suggested that the multi-head attention mechanism which is very important, allows for efficient parallelization during training. This approach is particularly advantageous for handling large datasets, a common challenge in speech processing (Siddique et al., 2023). According to Madahana (2022), they highlighted the significance of speech as a fundamental form of interaction and introduce human-computer interaction through a human-computer interface while focusing on a system developed using Raspberry Pi technologies.

As proposed by Akintola and Ibiyemi (2017), using transformers have gained significant attention in the natural language processing and speech processing fields due to their remarkable performance on so many applications and platform. The rapid emergence and widespread adoption of models that uses transformer in speech to text processing have prompted extensive research into the distinctive features that contribute to their outstanding performance (Chauhan et al., 2016). It was opined by Ajao, Yusuff and Ajao (2022), that significant lack of evidence regarding the use and optimization of machine learning and artificial intelligence algorithms as potential solutions for people with hearing impairments. The data collected primarily focuses on the home environment, featuring familiar words commonly used in domestic settings. The selected words for translation reflect those typically encountered in everyday home life (Sawai, Paik and Kuwana, 2021).

Additionally, interactions with native speakers were crucial to the process, mostly because of the scarcity of readily available resources or libraries for Yoruba, which is regarded as one of the many under-resourced African languages (Adewole, et al., 2017). indicate that speech classification based on utterances categorizes speech into isolated words, connected text, ongoing speech, and spontaneous speech. Furthermore, the classification of natural language processing techniques according to presenter mode distinguishes between two approaches related to speaker models: speaker-dependent and speaker-independent methods (Eludiora and Odejobi, 2016).

The system's primary objectives include establishing a login system for user access, implementing a translator to fulfill users' language translation needs, and delivering translation output in both voice and text formats (Akintola and Ibiyemi, 2017), In the study by target speech representations are generated and acquired through unsupervised learning. This adaptable approach can be used in any language, regardless of whether it has a written form. It was observed by Amin (2022) that the area of speech recognition has significantly enhanced the translation of speeches into written text by leveraging advanced technologies such as machine learning and artificial intelligence. Notable examples of speech to text technologies include Google Speech Recognition, Microsoft Azure Speech Recognition, IBM Watson Speech to Text, CMU Sphinx (Pocketsphinx), and Dragon NaturallySpeaking, among others (Sneha et al., 2023).

MATERIALS AND METHODS

One of the major importances of this platform is its capability to bridge the communication barriers, enhancing communication efficiency and inclusivity. Through meticulous analysis and thoughtful design, a robust and user-friendly platform that helps users to meet the need of communicating diverse languages was developed. The following tools will be employed for the design and development of the proposed system:

Active Server Pages (ASP)

Active Server Pages is a server-side scripting technology developed by Microsoft that facilitates the creation of dynamic web pages and applications. It allows for the integration of server-side code, typically written in languages such as C#, directly within HTML pages.

C# (C-Sharp)

This is a modern, object-oriented programming language created by Microsoft. It is widely utilized for developing Windows applications, web applications (using ASP.NET), and various software solutions. C# is appreciated for its simplicity, type safety, and scalability.

Bootstrap 5

Bootstrap 5 is the latest version of the open-source front-end framework aimed at streamlining the development of responsive and mobile-first web applications. With its enhanced grid system, utility classes, and customizable features, Bootstrap simplifies the design process through pre-built components and a responsive layout grid.

HTML 5 (Hypertext Markup Language)

This is the standard markup language for web pages,. This latest version introduces new elements and APIs that enhance the structure, multimedia capabilities, and semantics of web pages. HTML 5 supports multimedia elements, offline web application APIs, and improved form controls.

Google Cloud Translation API

The Google Cloud Translation API is a service that enables developers to seamlessly integrate automatic language translation into their applications, allowing for programmatic translation of text across different languages.

JavaScript

JavaScript is a versatile scripting language for web development that facilitates the creation of content that are dynamic and improves user experiences on web pages. Supported by web browsers, JavaScript is crucial for client-side scripting, providing interactivity and dynamic functionalities. The diagrams below in figure 1 and 2 show the hierarchy chart of the entire system and model of the proposed system.

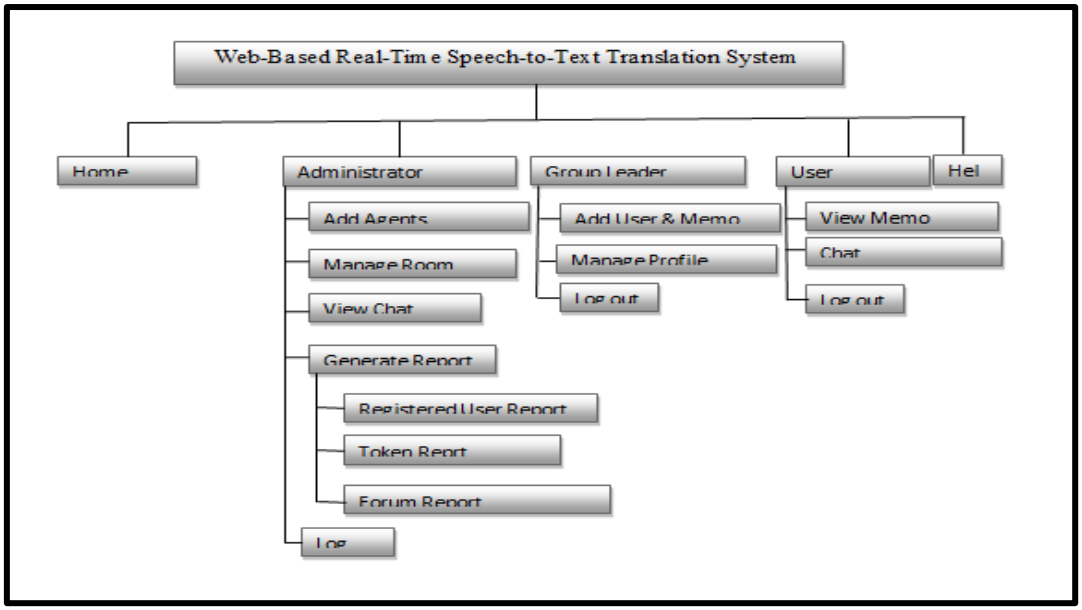

Figure 1: Hierarchy Chart of the Entire System

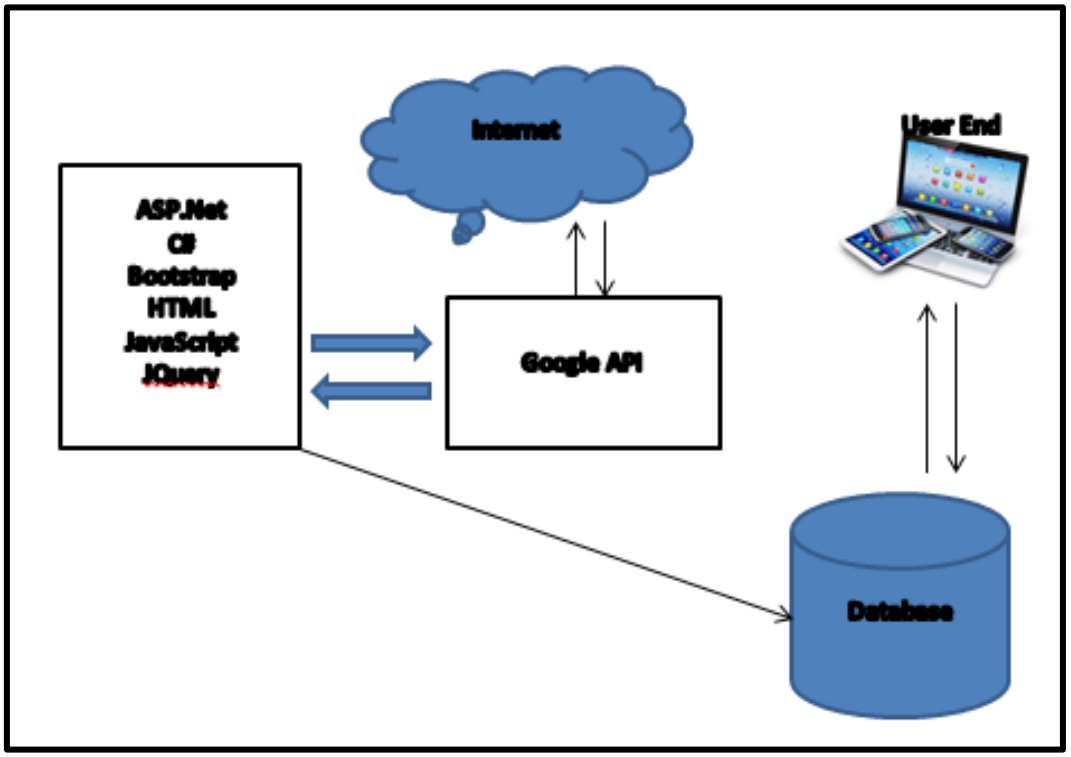

Figure 2: Model of the Proposed System

The Model's Components

***User Interaction*:** The central component of the system is the user interface, accessible through web browsers. Users usually have access to the platform in real-time by speaking into their device's microphone or uploading audio files. They can effortlessly switch between English and Yoruba languages for speech input.

Interactive Gateway

Serving as the interactive gateway, our web application connects users to the speech-to-text translation system. Whether accessed through a computer or a mobile device, users can easily input speech in either English or Yoruba, initiating real-time translation.

Data Repository

At the heart of our system is the database, which functions as a comprehensive digital repository. Speeches in English and Yoruba languages are securely stored alongside with their corresponding translated text. This centralized hub ensures efficient storage and retrieval of translation data, facilitating a unbroken user experience.

Google Language Translation API

The Google Language Translation API is a service provided by Google Cloud that empowers developers to incorporate language translation modules in the applications or websites. This API enables the programmatic translation of speech to text across designated languages, leveraging Google's

advanced machine learning models and language processing algorithms.

The Process of Speech Translation

Input Text

Users submit the text they wish to translate to the API.
Language Identification: If the language is not specified, the API automatically detects the intended language of the input text.

Translation

Utilizing advanced neural machine translation models, the API translates the input text into the desired target language.

Output

The API returns the converted text to the user's application, making it ready for display or further use.

Creating Google Speech-to-Text API

To create an API key, follow these steps:

i. Open the Navigation menu.
ii. Select APIs & Services > Credentials.
iii. Click Create Credentials and choose API Key.
iv. Copy and securely store the generated keys.
v. Click Close.

In the shell (SSH), run the following command, replacing <your_api_key> with your actual API key. Build your request to the API in a request.json file. To create the request.json file and open the file using your preferred command-line editor (such as nano, vim, or emacs) or using the gcloud command, then add the following content to your request.json file, using the URI value of the sample raw audio file: The system's algorithms and E-R diagram of the model are showed in figure 3 and 4 below.

Figure 3: Algorithms for Creating Google Speech to Text API

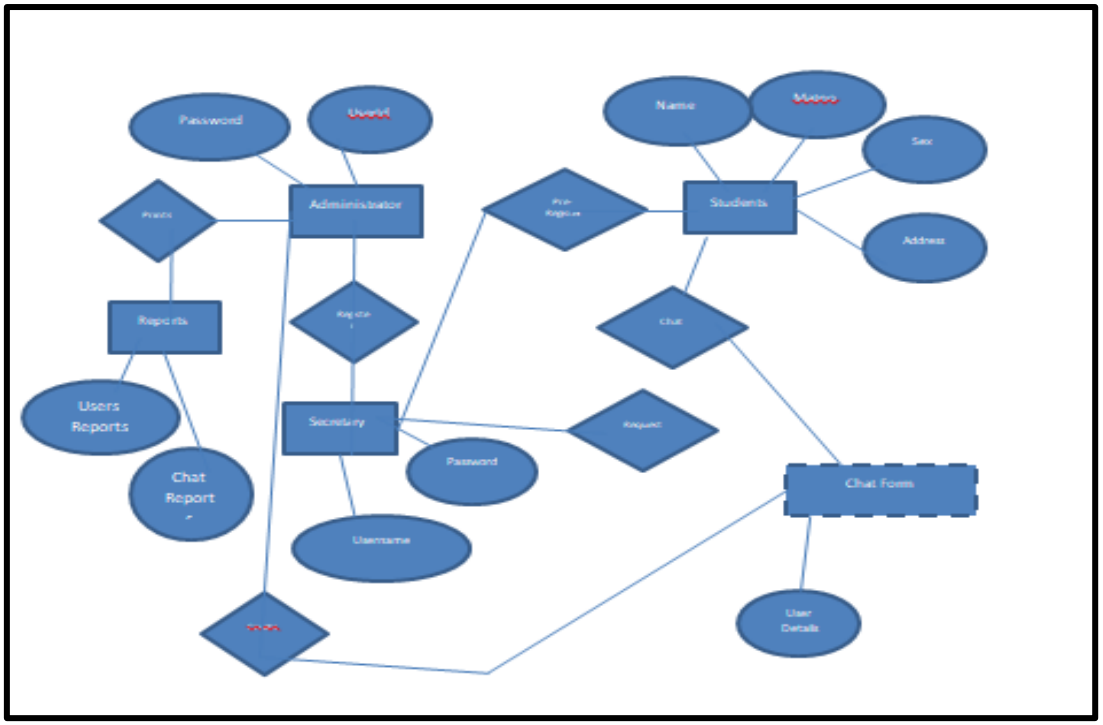

Figure 4: E-R Diagram of the model

The diagram in figure 5 shows the flowchart of the model.

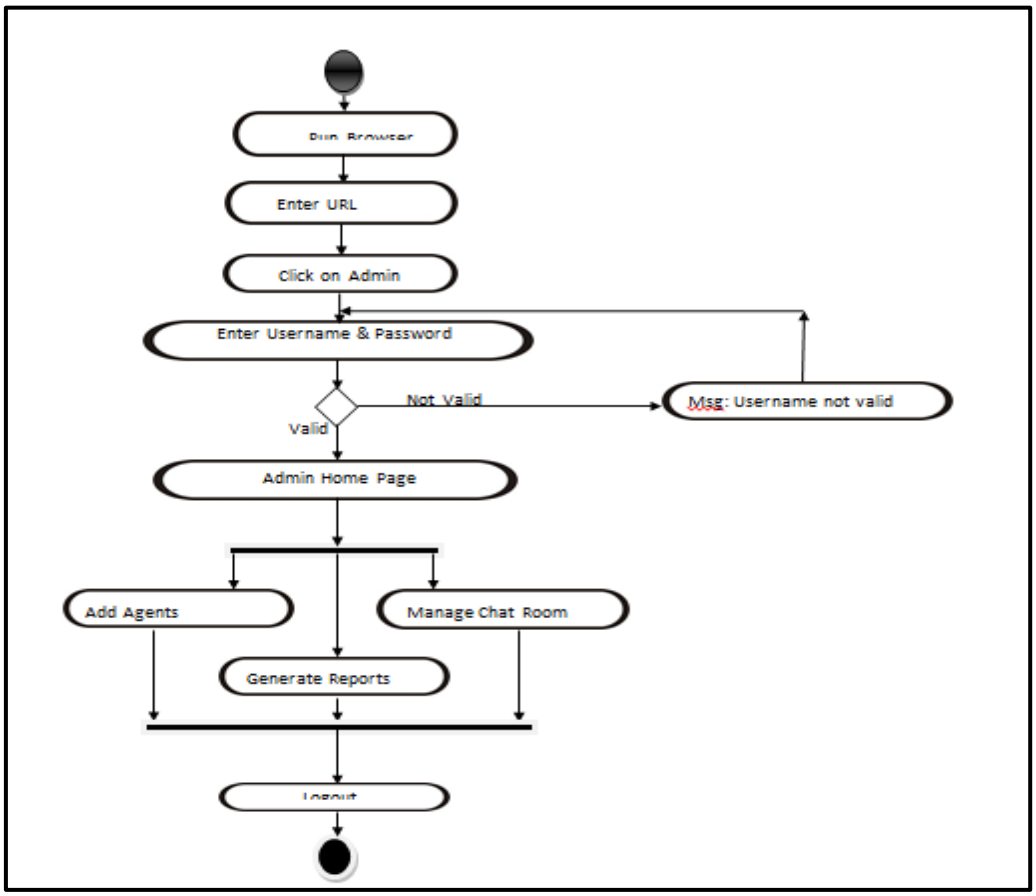

Figure 5: The flowchart of the Model

RESULTS AND DISCUSSION

The main objective of this model is to create a robust and accurate model tailored specifically for a web-based real-time speech-to-text translation system to English and Yoruba languages, meeting the increasing demand for dependable and efficient transcription tools. This model aims to encompass a

variety of functionalities that will improve speech to text translation, speech recognition, including accents, dialects, speaking speeds, ambient noise, and context within language use. By the integration of these diverse data sources, the goal is to build an advanced predictive framework capable of analyzing complex and varied speech data through the use of sophisticated linguistic models, acoustic processing methods, and intricate language modeling. This holistic approach is designed to manage uncertainties typical of spoken language data, thereby improving the accuracy, reliability, and real-world relevance of speech-to-text systems. Ultimately, this effort aims to make important contributions to accessibility, improve communication among diverse groups, and enrich user experiences in real-time speech-to-text translation services.

Algorithm For Implementing the Language Translation Model

Step 1: User Input

Speech Input (Optional): If it detects speech recognition mechanism, the system captures the audio when it detects English language.

Text Input: The user types an English sentence straight into the web interface.

Step 2: Speech-to-Text Conversion (Optional)

Audio Processing: The system enhances the audio signal to improve clarity and reduce background noise.

Speech Recognition: The speech recognition model converts the spoken English sentence into text.

Step 3: Preprocessing the English Text

Text Cleaning: Unnecessary characters, such as special characters and extra spaces, are removed.

Tokenization: The English sentence detected is broken down into small parts..

Step 4: Translation Process

Encoding: The tokenized English sentence is processed by the encoder of a sequence-to-sequence model, which generates a context vector—a numerical representation of the input.

Attention Mechanism: If implemented, the attention mechanism assists the model in focusing on relevant parts of the input during translation.

Decoding: Using the context vector, the decoder generates the corresponding Yoruba translation word by word.

Step 5: Postprocessing the Yoruba Text

Detokenization: The generated Yoruba tokens are combined into a coherent sentence.

Text Normalization: Adjustments are applied in order to make sure that the output is grammatically correct and natural in Yoruba.

Step 6: Output Display

Text Output: The translated Yoruba sentence appears on the web interface.

Speech Output (Optional): If text-to-speech functionality is enabled, the translated sentence is also converted into audio and played for the user. The diagram below in figure 6 shows the flow diagram of speech to text conversion

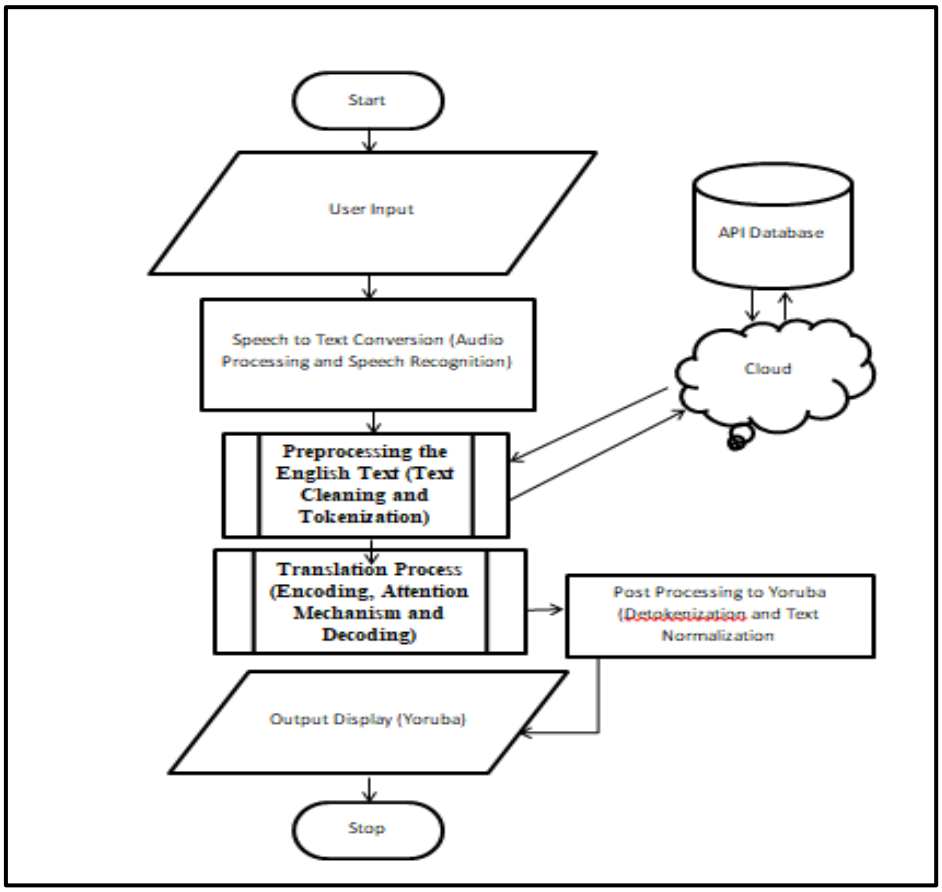

Figure 6: Flow diagram of Speech to Text Conversion

Result Evaluation Performance

Evaluating the performance of a web-based real-time English speech-to-text translation system involves measuring accuracy, speed, robustness, user experience, and reliability. Key accuracy metrics include Word Error Rate (WER) and Character Error Rate (CER), with lower values reflecting

fewer transcription errors. Potential and quantity are needed for real-time performance, ensuring fast responses and efficient handling of multiple requests. Robustness is assessed by testing the system's ability to handle various accents, dialects, and background noise without major accuracy losses. User experience is evaluated through feedback on satisfaction and ease of correcting errors, ensuring the system is intuitive and meets user needs. Reliability is measured through metrics like uptime and consistent response times to confirm the system remains operational and responsive in diverse conditions. A thorough assessment of these factors provides an accurate view of the system's effectiveness in real-world applications.

System Requirements

Outlined below are the requirements for the development of the model that is made up of the hardware and software requirements; Processor: Core i5 Celeron, 1 Terabyte HDD, 4 Gigabytes of RAM, Windows 10 or any compatible platform, Microsoft Visual Studio (versions 2017 to 2022) for front-end development and Microsoft SQL Server for back-end support.

How to Chat with the System

A main aim of this model is to develop a secured platform where students can communicate using a language translator to overcome language barriers, regardless of their location.

Step 1

Power on a computer system or android phone that is internet enabled and connected.

Step 2

Open a web browser and go to the software's URL.

Step 3

On the homepage, click the login tab in the menu bar and sign in as a registered user.

Step 4

Enter username and password to login the user and redirect the user to the Chat Page. Figures 7 and 8 shown below are the login and chat page interface of the platform.

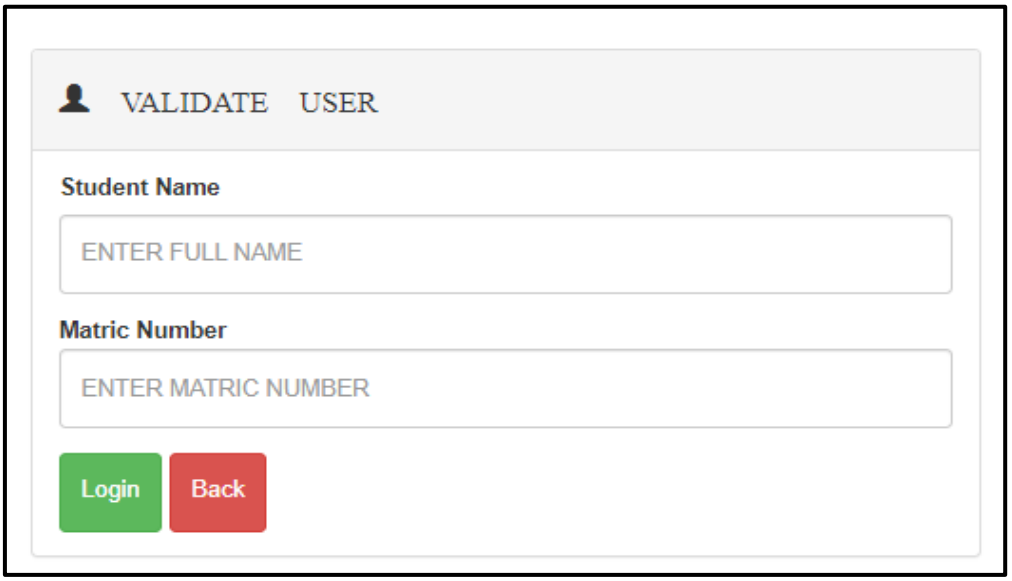

Figure 7: Login Interface

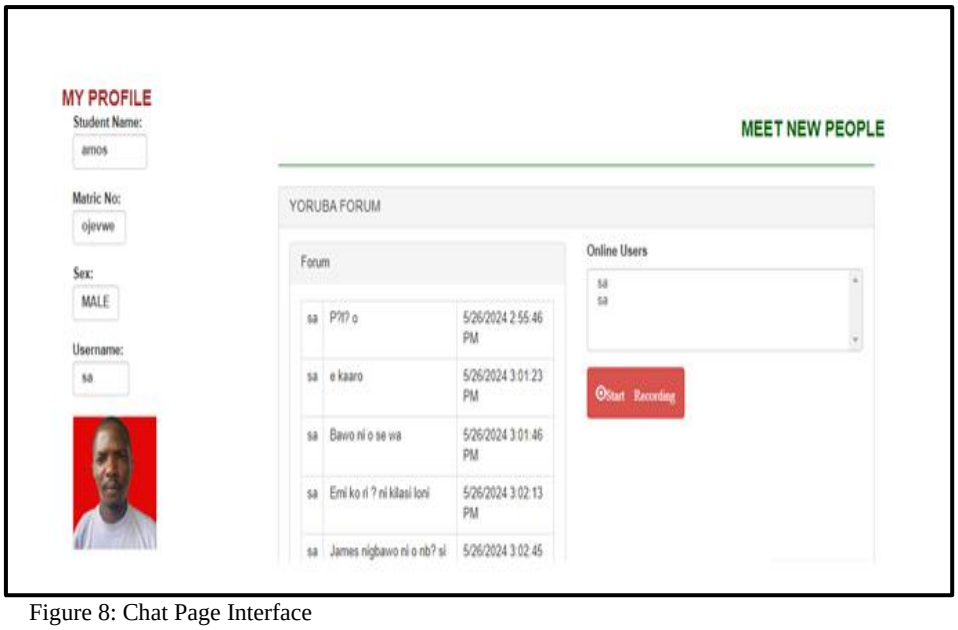

Figure 8: Chat Page Interface

The table below in Table 1 shows the speech translation sample of the platform

Table 1: Speech Translation Sample

S/N	Speech	Yoruba
1	Hello	P?IO (No Yoruba translation)
2	Good morning	E kaaro
3	Good evening	Ka a ale
4	Hello	P?IO (No Yoruba translation)
5	Good morning	E kaaro
6	We are going to school today	a nlo si ile-iwo loni
7	My name is John	John ni oruko mi
8	My name is John	John ni oruko mi

The diagrams in figure 9 and figure 10 indicate speech translation to text in English language and text in English language to Yoruba language.

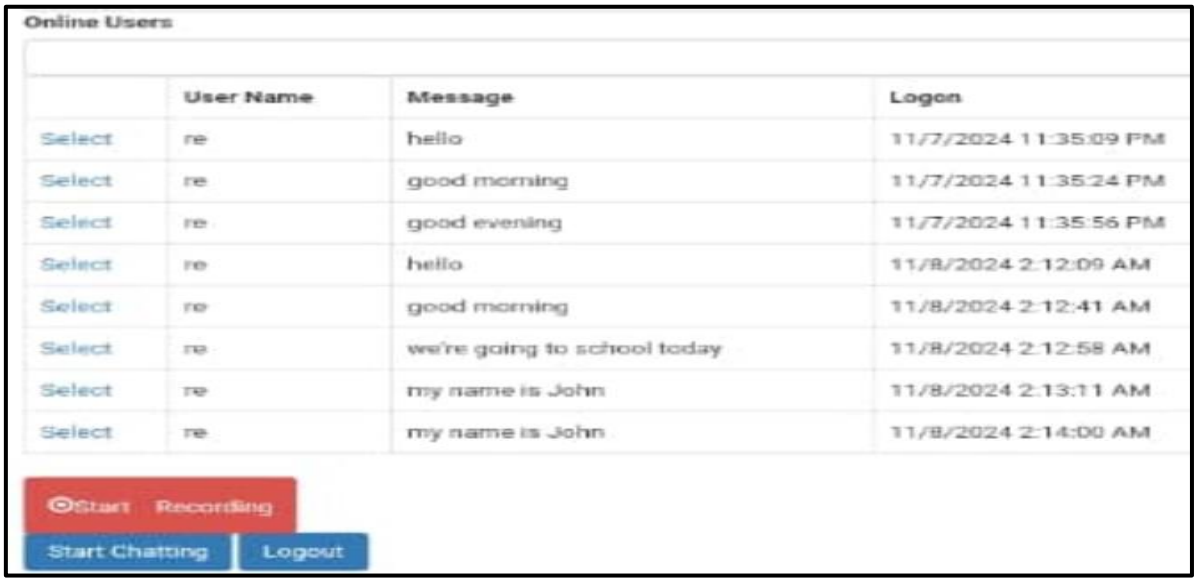

Figure 9: Speech Translation to Text in English language

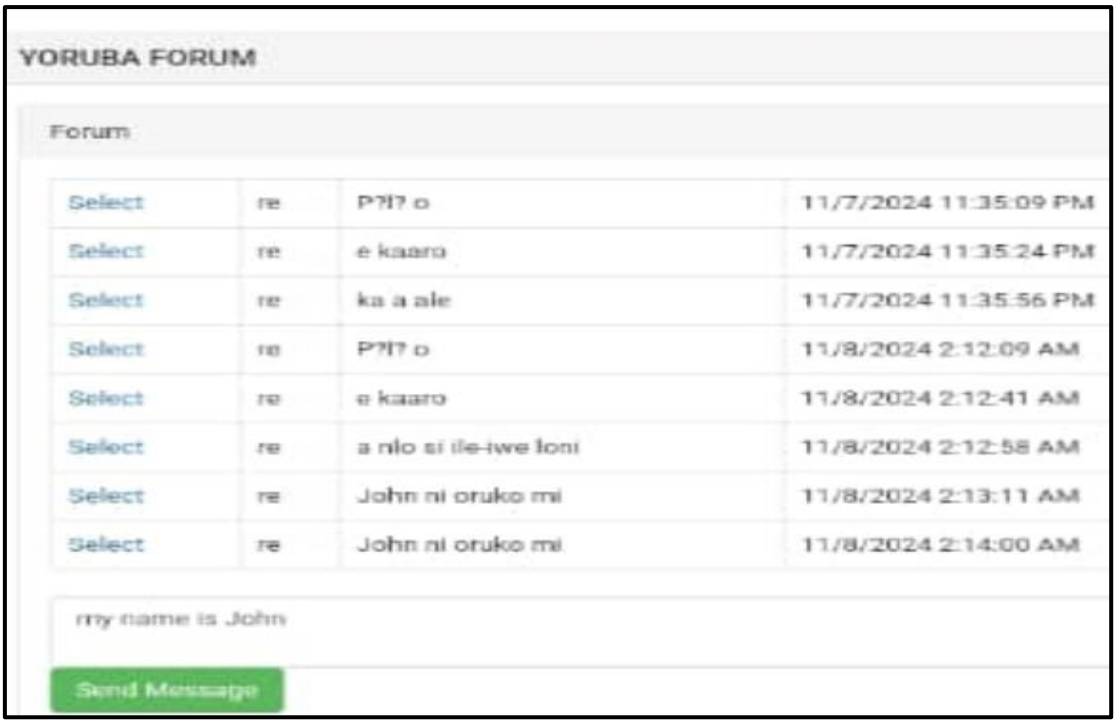

Figure 10: Text in English Language to Yoruba Language

CONCLUSION

The aim of this article is to develop an intuitive web-based platform that facilitates the real-time translation of spoken words to typed English and Yoruba language during a chat. By leveraging the robust capabilities of the Google Language API, the system improves communication by converting users' audio input into text, which is then translated. The software design utilizes Object-Oriented Methodology (OOM), ensuring the system is modular, scalable, and improved maintainability. The workflow starts with users providing input through speech or text, which is processed for clarity and converted to text when needed. This text is then cleaned and tokenized before entering the translation phase, where an encoder generates a context vector that represents the sentence. An optional attention mechanism enhances accuracy, allowing a decoder to translate the context vector into Yoruba language. The final translation is post-processed for coherence and grammatical correctness before being presented as text or converted back into speech for the user. This holistic approach guarantees a seamless and accurate translation experience from English language speeches (audio) to written English and Yoruba text-based languages. This platform marks a significant leap forward in language translation technology, effectively bridging communication barriers between English and Yoruba communicators. This model did not only enhance communication across languages but also lays the groundwork for future advancements in multilingual translation systems, promoting greater inclusivity and accessibility in the digital age.

REFERENCES

Adebara, I., Abdul-Mageed, M. and Silfverberg, M. (2022). Linguistically-motivated Yorùbá-English machine translation. In Proceedings of the 29th International Conference on Computational Linguistics (pp. 5066-5075). (Bird et al., 2020).

Adekunle, O., Agbonifo, O. and Olaniyan, J. (2020). Development of Bi-Directional English to Yoruba Translator for Real-Time Mobile Chatting. International Journal of Computational Linguistics, 11(1).

Adewole, L. B., Adetunmbi, A. O., Alese, B. K. and Oluwadare, S. A. (2017). Token Validation in Automatic Corpus Gathering for Yoruba Language. FUOYE Journal of Engineering and Technology, 2(1), 4.

Adigun, O., Rufai, M. M., Okikiola, F. M. and Olukumoro, S. (2024). Machine Learning Techniques for Prediction Of Covid-19 In Potential Patients, Vol. 7 No. 4 / DOI: https://doi.org/10.33003/fjs-2024-0804-2579

Akintola, A. and Ibiyemi, T. (2017). Machine to Man Communication in Yorùbá Language. Annal. Comput. Sci. Ser, 15(2).

Akintola, A. and Ibiyemi, T. (2017). Machine to Man Communication in Yorùbá Language. Annal. Comput. Sci. Ser, 15(2).

Akinwale, O. I., Adetunmbi, A. O., Obe, O. O. and Adesuyi, A. (2015). Web-based English to Yoruba Machine Translation. International Journal of Language and Linguistics, 3(3), 154-159.

Ajao, J., Yusuff, S. and Ajao, A. (2022). Yorùbá Character Recognition System Using Convolutional Recurrent Neural Network. Black Sea Journal of Engineering and Science, 5(4), 151-157.

Ajibade, B. and Eludiora, S. (2021). Design and Implementation of English to Yor\ub\'a Verb Phrase Machine Translation System. arXiv preprint arXiv: 2104.04125.

Amin, E. A. R. (2022). Using repeated-reading and listening–while-reading via text-to-speech apps in developing fluency and comprehension. World Journal of English Language, 12(1).

Benjamin, A. and Eludiora, S. (2020). Design and Implementation of English to Yorùbá Verb Phrase Machine Translation System. ACL 2020 Submission.

Chauhan, V., Dwivedi, S., Karale, P. and Potdar, S. M. (2016). Speech to Text Converter Using Gaussian Mixture Model (GMM). International Research Journal of Engineering and Technology (IRJET), 3(2), e-ISSN: 2395-0056.

Eludiora, S. I. and Odejobi, O. A. (2016). Development of English to Yorùbá Machine Translator. International Journal of Modern Education and Computer Science, 8(11), 8.

Esan, A., Oladosu, J., Oyeleye, C., Adeyanju, I., Olaniyan, O., Okomba, N. and Adanigbo, O. (2020). Development of a recurrent neural network model for English to Yorùbá machine translation. Development, 11(5).

Graves, A., Liwicki, M., Fernandez, S., Bertolami, R., Bunke, H. and Schmidhuber, J. (2013). Handwriting Recognition With Recurrent Neural Networks. Advances in Neural Information Processing Systems.

Madahana, M. (2022). A Proposed Artificial Intelligence-Based Real-Time Speech-To-Text to Sign Language Translator for South African Official Languages for The Covid-19 Era and Beyond: In Pursuit Of Solutions For The Hearing Impaired. South African Journal of Communication Disorders, 69(2).

Oloruntoyin, S. T. (2014). Development of Yoruba Language Text-to-Speech e-Learning System. International Journal of Scholarly Research Gate, 2(1), 19–36.

Padmane, P. (2022). Multilingual Speech and Text Recognition and Translation. International Journal of Innovations in Engineering and Science, 7(8).

Prachi, K. and Bhope, V. (2015). Implementation of Speech to Text Conversion. International Journal of Innovative Research in Science, Engineering and Technology, 4(7).

Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z. and Liu, T. Y. (2019). Fastspeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems, 32.

Sanchit, C., Aniket, S. and Tanvi, G. (2022). Real-Time Direct Speech-To-Speech Translation. International Research Journal of Engineering and Technology, 9, Jan 2022.

Sawai, R., Paik, I. and Kuwana, A. (2021). Sentence augmentation for language translation using gpt-2. Electronics, 10(24), 3082.

Sneha, B., Himanshi, A., Shreya, J., Shilpa, G., Mrinal, B., Biswajeet, P. and Mazen, A. (2023). Challenges and Limitations in Speech Recognition Technology: A Critical Review of Speech Signal Processing Algorithms, Tools and Systems, Computer Modeling in Engineering and Sciences 2023, 135(2), 1053-1089. https://doi.org/10.32604/cmes.2022.02175

Siddique, L., Aun, Z., Heriberto, C., Fahad, S., Moazzam, S. and Junaid, Q. (2023). Transformers in Speech Processing: A Survey. Computation and Language (cs.CL); Sound (cs.SD); Audio and Speech Processing (eess.AS).

Totare, Y., Sarthak, B., Sandesh, C., Sumit, D., Prathamesh, G. and Anurag, T. (2023). Speech Translation Using Machine Learning. International Research Journal of Modernization in Engineering Technology and Science, 5, May 2023.

Vasilakes, J., Zhou, S. and Zhang, R. (2020). Natural Language Processing. Elsevier Institute for Health Informatics Surgery, ISBN: 9780128202739.

Xiao, T. and Zhu, J. (2023). Introduction to Transformers: an NLP Perspective. arXiv preprint arXiv:2311.17633.