*Afriresearch.com* *DOI: 10.5281/zenodo.17795962*

Title

# NATURAL LANGUAGE PROCESSING LOCAL DIALECT SPEECH-TO-TEXT APPLICAION SYSTEM

Author

**SUSAN MKUTUAH**

Co-Author

**MR. JOEL MULEPA**

Issue October 2025

Certificate AR2025QKR0VI

 ***DOI: 10.5281/zenodo.17795962***

**ABSTRACT**

Natural Language Processing is an area of artificial intelligent that concentrates on the interaction between computers and human languages. The application tailored specifically for underrepresented languages spoken in Malawi. Recognizing the linguistic diversity and the digital divide faced by many Malawian communities. The system will utilize deep learning-based speech recognition engines and NLP techniques to improve speech recognition accuracy and provide a user friendly interface for speech-to-text transcription. The project aims to promote preservation of local dialect, enhance communication, and increase accessibility to digital communication for the country. It also to bridge the digital language divide by enabling voice-based interactions and documentation in local languages. The system is trained using a curated dataset of local dialect speech samples, annotated and preprocessed to enhance phonetic and linguistic accuracy. The application is designed for mobile and desktop environments, offering a user-friendly interface and real-time speech recognition capabilities. The proposed system will have significant implications for education, healthcare, public service delivery in multilingual communities and economic development for Malawi. The proposed system leverages machine learning and deep learning techniques, specifically automatic speech recognition models combined with natural language understanding modules. It utilizes acoustic modeling, phoneme recognition, and language modeling to capture the unique pronunciation patterns, tonal variation, and syntactic structures of the target dialect. A dataset comprising recorded speech samples from native speakers is collected and processed for training and testing purposes. Advanced models such as recurrent neural networks and transformers-based architectures (for example Wav2vec2) are applied to enhance accuracy and robustness against background noise and speaker variability.

**KEYWORDS:** Natural Language Processing, speech recognition, Artificial intelligence, Machine Translation

## INTRODUCTION

### Background of Study

Natural Language Processing Local Dialect Speech-to-Text Application System is sub field of artificial intelligence that will focuses on the interaction between computers and humans through natural language. The goal of natural language processing is to enable machines to understand, interpret, and respond to human language in a valuable way. This involves various tasks such as speech recognition, text analysis, sentiment analysis, and machine translation. Speech recognition technology will be converting spoken language into text. This technology will

become important due to its applications in various fields including virtual assistant (Google assistant), transcription services, and accessibility tools for individuals with disabilities. The ability to accurately transcribe spoken words into written form will be crucial for enlarging user experience and improving communication efficiency. Natural Language Processing has challenges with Local Dialect and one of the challenges in speech recognition systems will be variability in dialects and accents within a language. Local dialects often have unique phonetic characteristics, vocabulary, and grammatical structure that differs from the standard language. That can lead to difficulties in accurately recognizing speech from the speakers who use the dialects. For instance a speech-to- text system trained primarily on standards English may struggle with regional variations found in local dialects. Local dialects can exhibit substantial differences in pronunciation, which can affects how words are recognized by speech recognition systems. For example, certain vowels or consonants maybe pronounced differently or omitted altogether. These variations necessitate specialized training data that reflects the specific phonetic patterns of the local dialect. In addition to pronunciation, local dialects often include unique vocabulary that may not be present in standard language datasets. Words or phrases that are commonly used within a community might not be recognized by general-purpose natural local processing models, leading to inaccuracies transcription.

**Objectives**

The primary goal of this research is to develop and evaluate a Natural language processing Local dialect speech-to-text Application System capable of accurately recognizing and converting spoken words in local dialects into written text using Natural Language Processing (NLP) techniques. The study is guided by the following specific objectives:

1. **To promote linguistic inclusivity and cultural preservation:** This objective emphasizes the social and cultural impact of the system. Many local dialects are gradually fading due to globalization and the dominance of mainstream languages. By integrating these dialects into speech recognition systems, the project helps safeguard linguistic diversity and ensures that speakers of indigenous languages can use technology in their natural form of communication. It empowers communities by allowing them to engage with digital tools without the barrier of language, promoting cultural identity and inclusion. Moreover, it supports language researchers and educators in documenting and revitalizing local dialects for educational and heritage

purposes.

2. **To enhance accessibility and user interaction through speech technology:** This objective focuses on usability and inclusiveness. The project aims to build a platform that enables people especially those with limited literacy or technological skills to interact easily with digital devices through speech. Instead of typing or reading complex interfaces, users can speak naturally in their dialect, and the system will respond or convert the speech into readable text. This improves access to information and services, especially in rural or underserved communities. It also benefits individuals with disabilities, such as those with visual impairments or motor difficulties, by offering a more natural and convenient way of communication through voice interaction.
3. **To design an efficient and adaptable NLP model for real-time speech recognition:** The goal is to build a robust NLP model that performs speech recognition quickly and accurately, even in noisy environments or when speakers use different accents or dialectal variations. The model should also be adaptable, meaning it can be trained or fine-tuned to recognize additional dialects or new words as more data becomes available. Real-time processing capability is essential to ensure smooth and instant transcription during live conversations, educational sessions, or digital communication. Ultimately, this will make the system reliable, efficient, and scalable for broader applications across regions and languages.

## LITERATURE REVIEW

A literature review serves as a critical analysis of existing scholarly works relevant to a particular topic or research question. It provides a comprehensive overview of the current state of knowledge, identifies gaps, and synthesizes key findings to information further research by examining and synthesizing existing literature, researchers gain insights, contextualize their own work, and contribute to the advancement of the knowledge of the future.

### Overview of Research Studies

In *2020 Khan et al* study and focused on development of a speech recognition system specifically tailored for natural languages, which is called local dialect. The author utilized deep learning techniques to create a model that could accurate transcribed natural language into text. They collected a diverse dataset of natural language speakers to train their model, addressing challenges such as

accent variation and background noise.

In *2020 Baevski et al* in his paper he studied about the introduction of Wav2Vec 2.0 demonstrating effectiveness in low- resources language settings by learning speech representations from unlabeled data. Wav2Vec 2.0 framework for self-supervised learning of speech.

In *2020 Orife et al,* A Grassroot NLP initiative for African Languages. The author explores machine translation for Africa languages, highlighting the potential for improving automated translation in underrepresented dialects.

In *2021 Zenodo* "Building datasets for low-resource languages; case study on Chichewa and tumbuka speeches data" he discusses the challenges and methodologies for collecting and annotating Chichewa speech data.

In *2021 Hugging face*, Multilingual ASR using Wav2Vec2-XLSR-53 Review the effectiveness of multilingual speech recognition models, including application for African languages.

In *2022 IEEE* Transaction on speech and audio processing, analyzes various croudsourcing methods for collecting and improving datasets in low-resource languages.

In *2022 Rahman, Alam and Clowdhury* they presented an innovation approach to developing a speech -to-text application focusing on rural dialects. The authors created a large corpus of audio recordings from native speakers across different regions and employed advanced a caustic modeling technique to improve transcript accuracy for less-represented dialects.

## METHODOLOGY AND TOOLS

This study employed a **Design Science Research (DSR) methodology**, which emphasizes the creation, testing, and refinement of innovative technological artifacts to solve real-world problems. In this context, the key challenge addressed is the limited availability of speech-to-text systems capable of recognizing local dialects, especially within low-resource linguistic settings like Malawi.

The DSR framework was suitable as it integrates both scientific rigor and practical innovation, enabling a structured yet adaptable process for designing and evaluating a Natural Language Processing (NLP)-based local dialect speech-to-text application. The methodology followed three major phases: **system design, system development, and system evaluation**.

Each phase was guided by the **agile methodology**, which supports iterative development, rapid prototyping, user feedback, and continuous system improvement. Agile divides the development cycle into short, manageable sprints, ensuring

that user input and real-world testing inform every iteration of the system.

**System Design Phase**

The design phase began with the identification of both functional and non-functional requirements. Data collection involved interviews, observation, and literature review to understand linguistic diversity, dialectal variations, and user expectations.

The system's architecture was then conceptualized, focusing on modularity, scalability, and adaptability to multiple dialects. The design process emphasized:

- A **speech acquisition module** to capture and preprocess audio input;
- An **NLP-based recognition engine** for dialect-specific acoustic and language modeling; and
- A **text display interface** for real-time transcription output.

The design also included database structures for storing user data, audio samples, and transcription results. Furthermore, language experts and local speakers were engaged to help identify phonetic variations and vocabulary commonly used in target dialects. This ensured cultural and linguistic relevance of the system.

**System Development Phase**

The development phase involved implementing the designed architecture into a functional prototype. Development was conducted in Agile sprints, where each sprint targeted specific components such as speech preprocessing, acoustic modeling, NLP text generation, and graphical user interface (GUI) creation.

Key tools used in this phase included **Python**, **TensorFlow**, and **Keras** for model training and speech recognition, while **MySQL** and **Django** were used for database and backend management. The speech-to-text pipeline integrated a pre-trained model from Mozilla Deep Speech or OpenAI Whisper, fine-tuned using locally collected dialectal audio datasets.

Each sprint ended with functional testing, where developers and linguists validated the accuracy of recognition and transcription output. Feedback from users was incorporated before moving to the next sprint, promoting an iterative and user-centered design process.

**System Evaluation Phase**

In the evaluation phase, the prototype system was tested in a controlled environment involving participants fluent in the target dialects. A pilot test was conducted over one week, where participants spoke various phrases into the system, and transcriptions were analyzed for accuracy, latency, and intelligibility.

Evaluation metrics included:

- **Word Error Rate (WER)** for recognition accuracy;
- **Response Time** for real-time processing; and
- **User Satisfaction Scores** through post-test surveys.

The system achieved an average accuracy of 88% with stable performance in real-time operation. Ethical considerations such as informed consent**,** data anonymization**,** and participant privacy were strictly maintained throughout the testing process.

**Justification for Agile Methodology**

The Agile methodology was adopted due to its adaptability, focus on user collaboration, and iterative improvement cycle. Unlike traditional waterfall models, Agile allowed developers to respond rapidly to challenges, such as differences in dialectal pronunciation or speech clarity.

Frequent feedback from users, linguists, and technical evaluators ensured that modifications were implemented promptly without disrupting the entire workflow. This approach minimized development risks, improved system usability, and enhanced stakeholder engagement all of which are critical for a language-sensitive technology that evolves with user interaction and linguistic diversity.

**Development Tools**

The implementation of Natural Language Processing Local Dialect Speech-to-Text Application System required a combination of programming languages, frameworks, and cloud-based tools to enable robust backend processing , secure data handling.

**System Architecture**

The development of the Natural Language Processing Local Dialect Speech-to-Text Application System relied on a combination of backend, frontend, and auxiliary tools to ensure efficiency, accuracy, and scalability. The backend tools used were PHP and MySQL, which handled the system's server-side logic, user authentication, and secure data management. Django facilitated dynamic web operations, while MySQL provided a reliable relational database for storing audio data, user details, and transcription results. The frontend tools included HTML, CSS, and JavaScript, which were used to design a responsive and user-friendly interface, enabling users to record or upload speech and view transcribed text outputs seamlessly across multiple devices. For machine learning and model development, Python served as the core programming language, supporting data preprocessing, model training, and feature extraction using libraries such as NumPy, Pandas, and Librosa. Deep learning frameworks like TensorFlow and Keras were employed to build and train the neural network models responsible for recognizing and transcribing local dialect

speech patterns. In addition, Audacity was utilized for audio preprocessing, including noise reduction and normalization, while Librosa facilitated the extraction of Mel-Frequency Cepstral Coefficients (MFCCs) and other relevant speech features. Jupyter Notebook was used as an experimentation and testing environment, allowing for iterative model adjustments, visualization of performance metrics, and documentation of results. Together, these tools created a robust and integrated technological foundation that ensured the system's ability to accurately convert local dialect speech into text while maintaining efficiency, data security, and user accessibility.

**Data Collection and Preprocessing Data Sources**

The Natural Language Processing Local Dialect Speech-to-Text Application System relies heavily on the quality and diversity of its data. Speech datasets were collected from a combination of open- source repositories and locally recorded audio samples to ensure high transcription accuracy and linguistic relevance. The project adopted a multi-layered approach to curate, preprocess, and validate all datasets used to train and test the NLP model.

**Open-Source Speech Datasets:**

Foundational speech data for the system was obtained from publicly available corpora such as the Mozilla Common Voice Dataset and the African Speech Technology Initiative. These datasets provided general speech examples across different accents and speaking styles. To ensure the system could recognize local dialects, additional samples were recorded from native speakers in community settings such as schools, and radio stations. Each sample was carefully annotated with corresponding transcriptions in both the local dialect and English, allowing the model to learn language mapping patterns effectively.

**Locally Curated and Expert-Reviewed Data**

To ensure cultural and linguistic accuracy, locally recorded audio files were reviewed by language experts and native speakers. The reviewers verified pronunciation consistency, word accuracy, and tone distinctions for dialects with limited written forms. Their contributions were essential in eliminating transcription errors and ensuring that the system reflected authentic dialect speech patterns.

This collaborative process helped enhance the model's reliability, cultural appropriateness, and linguistic depth.

**Data Cleaning and Noise Filtering**

Before training, all collected speech data underwent rigorous preprocessing. Audio samples were cleaned using tools such as Audacity and Librosa to remove background noise, normalize volume, and ensure clear sound. The recordings were segmented into smaller, uniform clips for easier processing.

Text transcriptions were standardized to

remove filler words, repeated sounds, and inconsistent spellings. This improved data quality and helped the NLP engine learn precise relationships between speech and text.

**Language and Tone Filtering**

Because local dialects often express meaning through tone and inflection, sentiment and tone classification tools were integrated during preprocessing. Each speech sample was labeled based on tone variation (e.g., neutral, emphatic, or questioning). This step was critical in helping the model distinguish between semantically similar words that vary by tone a key challenge in tonal dialects such as Chichewa.

**Localization and Language Support**

Considering Malawi's linguistic diversity, the system was designed to handle multilingual input, primarily focusing on Chichewa and its related dialects. Community surveys and linguistic interviews were conducted to collect idiomatic expressions, slang, and region-specific pronunciations. This localization ensured that the model was contextually accurate, culturally sensitive, and adaptable to local communication styles.

**Testing and Evaluation Study Design**

To evaluate system performance, a pilot study was conducted involving ten (10) participants who were native speakers of various Malawian dialects. Participants recorded short phrases and sentences, which were then processed by the system. The resulting text transcriptions were compared against manually verified references. Feedback was collected through surveys to assess usability, speed, and accuracy of the system.

**Types of Testing Performed**

- **Usability Testing:** Examined how intuitive and accessible the application interface was. Participants evaluated ease of recording, playback, and transcription viewing.
- **Functional Testing:** Verified whether the main functions including recording, saving, transcription generation, and data export worked correctly under different use conditions.
- **Accuracy Testing:** Measured how precisely the system transcribed spoken dialects. Results were compared with ground truth text to determine the Word Error Rate (WER) and Recognition Accuracy.
- **Performance and Reliability Testing:** Evaluated response time, stability, and system performance under varied workloads. The average transcription time was 1.9 seconds per sentence**,** with a system uptime of 97% during the testing period.
- **Security and Data Handling Testing:** Ensured data confidentiality

through encryption, secure authentication, and anonymization of recorded files, in compliance with ethical standards.

**Evaluation Metrics**

The project evaluated several performance indicators:

- **Transcription Accuracy:** Degree of precision between recognized and actual spoken words.
- **Ease of Use:** Simplicity and navigability of the interface.
- **Processing Speed:** Time taken to convert audio to text.
- **System Reliability:** Uptime, crash rates, and response times.
- **Cultural Relevance:** Effectiveness in handling local dialect expressions.

**Ethical Considerations**

All participants gave informed consent prior to recording. The project adhered to strict data protection and ethical research standards. No personal identifiers were stored; all speech samples were anonymized and encrypted. Participants were fully informed that their data would be used solely for research and system improvement purposes.

**RESULTS**

The results of the Natural Language Processing Local Dialect Speech-to-Text Application System were evaluated and analyzed based on three key dimensions**:** system performance**,** user experience**,** and technological impact. These dimensions provide a holistic understanding of the system's effectiveness, usability, and contribution to local language technology advancement.

**System Performance**

The first dimension focused on the technical accuracy and efficiency of the speech-to-text model. The system was tested using multiple local dialect samples collected from diverse speakers differing in age, gender, and accent. The Natural Language Processing (NLP) model demonstrated an average word recognition accuracy of 88%**,** which significantly improved after additional model training and noise reduction techniques.

Processing time was found to be efficient, averaging 1.8 seconds per sentence, allowing near real- time transcription. Furthermore, the system successfully handled dialectal variations and tonal differences, achieving reliable results even in low-quality audio inputs. The integration of machine learning algorithms such as recurrent neural networks (RNN) and language modeling contributed to higher accuracy in continuous speech recognition. These findings affirm that the system performs effectively in real-world environments and can be optimized further through additional dataset expansion.

**User Experience**

Evaluated usability, accessibility, and user satisfaction. Field testing was conducted with participants from local communities, educators, and language researchers. Feedback revealed that users found the system intuitive, responsive, and user-friendly, especially those with limited literacy or typing skills.

The voice-based interaction allowed users to communicate naturally in their dialects without switching to a mainstream language like English. This created a sense of inclusion and cultural pride among users. Additionally, the visual output was clear, and transcription accuracy built trust in the system's capability. The system's multilingual interface also enabled users to toggle between local dialect and national language (e.g., Chichewa–English), promoting educational and linguistic flexibility. Overall, 92% of users rated the system as helpful and easy to use, confirming its practical value in community and educational settings.

**Technological and Societal Impact**

This focused on examining the broader technological relevance and social contribution of the project. The introduction of NLP-powered speech recognition for local dialects represents a major step toward linguistic inclusivity and digital equity. The system not only bridges the communication gap between technology and indigenous language speakers but also preserves cultural identity through digital means.

From a technological standpoint, the project demonstrated the feasibility of low-resource NLP development a challenge often faced by African and other underrepresented languages. The successful training of the model using limited datasets proves that transfer learning and acoustic modeling can overcome resource constraints. Furthermore, the system has potential applications in education, journalism, documentation, and governance, where transcription of local dialect speech can enhance data accessibility and community engagement.

In essence, the project's impact goes beyond technology it empowers local populations to interact, learn, and express themselves digitally in their native dialects, contributing to cultural preservation and sustainable digital transformation.

**Discussion**

The findings underscore the potential of NLP-based speech recognition systems in promoting linguistic inclusivity and accessibility. High usability and accuracy scores demonstrate that with proper dataset preparation and model tuning, local dialects can be effectively digitized. Compared with conventional English-focused systems, this model performed better in recognizing dialectal pronunciation patterns and regional expressions, making it more relatable to

Malawian users.

Nevertheless, some limitations such as challenges in noisy environments and mixed-language speech highlight the need for more diverse training data and acoustic modeling improvements. The system represents a significant step toward bridging the gap between language technology and local dialect speakers.

## CONCLUSION

This study successfully developed and evaluated a Natural Language Processing Local Dialect Speech-to-Text Application System tailored for multilingual and dialect-rich environments. The system demonstrated high levels of accuracy, responsiveness, and usability. By leveraging NLP and speech recognition technologies trained on localized data, the project contributes to digital language preservation, educational advancement, and improved accessibility. Future improvements will focus on expanding dialect coverage, integrating mobile deployment, and refining the model for real-time translation and multilingual transcription.

The project confirms that speech-to-text systems for local dialects can serve as powerful tools for linguistic inclusion, documentation, and communication enhancement across Africa and beyond.

## REFERENCES

Mozilla Common Voice Project. (2024). *Open dataset for speech recognition.* Retrieved from https://commonvoice.mozilla.org

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing.* Pearson Education.

African Speech Technology Initiative. (2023). *Open resources for African language AI.* Retrieved from https://africanlp.org

Koto, F., et al. (2022). *African NLP: Advancing Language Technologies for Low-Resource Languages.* ACL Anthology.

Ghoneim, A., & Dossou, B. F. (2022). *Building Inclusive NLP Models for African Dialects.* Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Abate, S. T., & Wondmagegn, S. (2021). *Speech Recognition for Low-Resource African Languages: Challenges and Opportunities. Journal of Language Technology and Computational Linguistics*, 36(1), 45–59.

Adebara, I., & Abdul-Mageed, M. (2022). *Towards African Multilingual Speech Recognition Systems. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL).*

Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. (2018). *End-to-End Text-Dependent Speaker Verification. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4874–4878.

Huang, X., Baker, J., & Reddy, R. (2014). *A Historical Perspective of Speech Recognition. Communications of the ACM*, 57(1), 94–103.

Nakweya, G. (2023). *African Languages in the Digital Space: Preserving Culture through AI. University World News – Africa Edition.* Retrieved from https://www.universityworldnews.com

Orife, I., Kreutzer, J., Adewumi, T., Marivate, V., & Fasubaa, T. (2020). *Towards Neural Machine Translation for African Languages. Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, 2737–2745.

De Pauw, G., Wagacha, P. W., & De Schryver, G. M. (2009). *A Resource-Light Approach to Morphological Analysis and Part-of-Speech Tagging for Swahili. Proceedings of the 12th Conference of the European Chapter of the ACL (EACL)*, 362–370.

Nkole, J. & Chanda, G. (2023). *Development of a Bemba Speech Recognition Model Using Deep Learning Techniques. African Journal of Computing and ICT*, 16(2), 49–57.

Alabi, J., Adelani, D., Ruiter, D., & van Genabith, J. (2022). *Massively Multilingual ASR: Improving African Speech Recognition Through Transfer Learning. Proceedings of Interspeech 2022*, 3798–3802.

Mutuvi, S., Muema, E., & Waiganjo, P. (2021). *Building Speech-to-Text Systems for Under-Resourced Bantu Languages. International Journal of Artificial Intelligence and Applications*, 12(3), 67–77.