

Voice to Empower: A Multilingual AI Solution for Tribal Inclusion

1st Prof. Meghna K
Artificial Intelligence and Data
Science
MEA Engineering College,
Perinthalmanna Kerala, India

2nd Mohammed Rizvin M K
Artificial Intelligence and Data
Science
MEA Engineering College,
Perinthalmanna Kerala, India

3rd Sahwa K
Artificial Intelligence and Data
Science
MEA Engineering College,
Perinthalmanna Kerala, India

4th Wafa
Artificial Intelligence and Data Science
MEA Engineering College,
Perinthalmanna Kerala, India

5th Mohammed Sanin M
Artificial Intelligence and Data Science
MEA Engineering College,
Perinthalmanna Kerala, India

Abstract - The Voice to Empower project introduces a multilingual artificial intelligence platform designed to bridge the communication barriers faced by India's tribal communities.

Many tribal groups lack access to digital services because most platforms do not support their native language or dialect. This system integrates Automatic Speech Recognition (ASR), Machine Translation, and Text-to-Speech (TTS) technologies to deliver information and assistance in the user's language.

The process begins when a tribal user speaks in their native language, such as Santali. The ASR module converts this spoken input into text, which is then translated

into English for information retrieval purposes. The system searches for relevant and accurate responses from trusted sources and then translates the answers into the user's tribal

language. Finally, the TTS module converts the translated text into natural speech output, allowing the users to hear the response clearly. By combining speech, translation, and language processing, the system enables tribal users to access essential information about health, education, and government services with ease. This AI-driven approach promotes digital inclusion, cultural preservation, and empowerment, ensuring that every tribal individual can understand and interact with technology in their language.

I. INTRODUCTION

In today's digital world, most online services are created for globally-common languages such as English or Hindi. Despite this, history of the Santali community and their native language not supporting by existing services disrupts them from using this needed solution alongside the rest of

tribe communities in India. This creates a serious digital gap, where people cannot easily get important information about the health and education sectors, farming or government welfare systems.

To solve this issue, we developed TRIEM (Tribal Responsive Intelligent Empowerment Model) a real-time AI-based voice assistant designed specifically for tribal communities. TRIEM enables users to voice naturally in Santali, and the device recognizes their voice for processing their question, and responds back in clear Santali speech.

The system operates through an integrated pipeline of advanced language technologies that enable seamless voice interaction. It begins with Automatic Speech Recognition (ASR), which converts spoken Santali input into text format. The recognized text is then processed using Machine Translation (MT) to translate Santali into English, allowing the system to analyze and understand the query more effectively. Once translated, a Large Language Model (LLM) generates contextually relevant and accurate information in response to the user's question. Finally, the generated response is translated back into Santali and converted into natural speech using Text-to-Speech (TTS) technology, enabling the user to receive the answer in clear and understandable Santali voice output.

This system works as an integrated pipeline of sophisticated

language technologies that facilitate smooth voice interaction. It starts with Automatic Speech Recognition (ASR), which translates the spoken Santali input into text form. The detected text is further processed by Machine Translation (MT), which translates Santali text into English, making it easier for the system to analyze and interpret the query. After translation, a Large Language Model (LLM) produces contextually relevant and accurate information based on the user's query. Finally, the produced output is translated back into Santali and expressed in natural speech form using Text-to-Speech (TTS) technology, allowing the user to receive the answer in clear and understandable Santali speech output.

By leveraging these technologies, TRIEM facilitates smooth voice interaction in the user's native language. The system also has fallback mechanisms and safe response handling for responsible and reliable output.

Our aim is not only to develop a voice assistant but to empower tribal populations, sustain low-resource languages such as Santali, and promote digital inclusion. With TRIEM, technology becomes more inclusive, accessible, and meaningful to those who have been digitally marginalized.

II. METHODOLOGY

The proposed system, TRIEM (Tribal Responsive Intelligent Empowerment Model), follows a real-time end-to-end speech processing pipeline designed specifically for Santali, a low-resource tribal language. The architecture is modular, allowing each component to function independently while contributing to a unified workflow. The overall process begins with speech input from the user and ends with synthesized voice output in Santali.

Initially, the user provides input through a microphone interface integrated into a web-based platform built using Flask. The recorded Santali speech undergoes basic audio preprocessing techniques such as noise reduction and silence trimming to improve clarity and recognition accuracy. The cleaned audio signal is then passed to the Automatic Speech Recognition (ASR) module. In TRIEM, multilingual transformer-based ASR models (such as IndicConformer/Whisper variants explored during development) are used to transcribe Santali speech into text format. Since Santali is a low-resource language, multilingual and cross-lingual transfer learning approaches are leveraged to improve transcription performance.

Once the Santali text is obtained, the system performs an intermediate translation step. The Machine Translation (MT) module translates Santali text into English using Indic language translation models (such as IndicTrans2). This translation enables effective downstream processing, as most large language models are optimized for English-based

reasoning and information retrieval.

Before submitting the query to the Large Language Model (LLM), the system conducts a similarity based cache check using a local SQLite database. The translated English query is compared with previously cached questions using text similarity methods (such as sequence matching or embedding similarity). If an equivalent question is identified above at a certain threshold, the stored response is directly retrieved, thereby cutting down on latency and API calls. In the absence of a match, the question is routed to the LLM (Gemini/Groq/Ollama fallback plan) for response generation. The LLM produces contextually relevant, safe, and structured data, to ensure sensitive areas like healthcare advice are handled with care.

Once the English response is generated, it is again translated back into Santali using the same MT engine. The final Santali text is then routed to the Text-to-Speech module, where models like Indic Parler-TTS are employed to produce natural-sounding Santali speech. The audio is then streamed back to the UI for playback.

To ensure reliability, the TRIEM system uses fallbacks and carefully designed prompts. If the main cloud-based LLM service is down or has reached quota limits, the system automatically switches to the local model (Ollama-based LLM). This fallback approach ensures continuous operation even in low-connectivity settings.

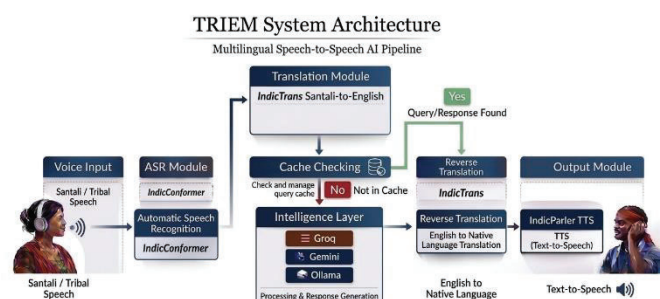


Fig. 1. TRIEM System Architecture

A. ASR: Automatic Speech recognition

Automatic Speech Recognition (ASR) plays a fundamental role in the TRIEM system, as it serves as the primary interface between the user and the AI system. Since the target users are tribal community members who primarily communicate through speech in their native language, enabling accurate voice recognition is essential for natural and accessible interaction. In this work, the ASR module is designed to capture

spoken Santali queries from the user and convert them into textual form so that the system can process and understand the request.

The speech input is first captured through a microphone interface integrated into the system's web application. Before recognition, the recorded audio undergoes basic preprocessing steps such as noise reduction and silence trimming to improve the clarity of the signal. These preprocessing steps help reduce background noise and enhance the overall quality of the speech input, which is particularly important when the system is used in real-world environments where external noise may be present.

Once the audio signal is cleaned, it is passed to the ASR model for transcription. Since Santali is considered a low-resource language with limited publicly available speech datasets, multilingual speech recognition models are utilized to improve recognition capability. These models are trained on multiple languages and can generalize better to languages that have limited training data. Through this approach, the system is able to convert Santali speech into text with reasonable accuracy, enabling further processing in the pipeline.

However, during experimentation it was observed that speech recognition performance can vary depending on factors such as pronunciation differences, dialect variations, and background noise. Tribal languages often contain regional variations and informal speech patterns, which can affect transcription accuracy. Despite these challenges, the ASR module provides a reliable mechanism for transforming spoken Santali queries into machine-readable text, forming the first and most critical step in the TRIEM processing pipeline.

By enabling speech-based input, the ASR component significantly improves accessibility for users who may not be comfortable typing in digital systems. This approach supports inclusive interaction and aligns with the overall goal of the TRIEM system, which is to provide technology that can be easily used by tribal communities in their own language.

B. TTS: Text To Speech

The Text-to-Speech (TTS) module is responsible for delivering the final response of the TRIEM system in spoken Santali, enabling users to receive information in a natural and easily understandable format. Since many users in tribal communities may prefer listening rather than reading text, speech output plays an important role in making the system accessible and user-friendly. The TTS component converts the generated Santali text response into clear and audible speech that can be played back to the user through the application interface.

After the system generates a response using the language model and translates it back into Santali, the resulting text is passed to the TTS module. The TTS system processes the text and synthesizes speech by modeling the pronunciation and rhythm of the language. This allows the system to produce audio responses that resemble natural human speech rather than robotic or mechanical output. The generated audio is then

streamed directly to the user interface so that the user can hear the response immediately.

Developing speech synthesis for Santali presents certain challenges because it is considered a low-resource language in the field of speech technology. Compared to widely spoken languages, there are fewer publicly available datasets and pretrained models specifically designed for Santali speech synthesis. To address this limitation, multilingual or adaptable TTS frameworks are used, which can support multiple Indian languages and generalize better to low-resource settings.

During system testing, it was observed that the clarity and naturalness of the synthesized voice depend on the quality of the translated text and the pronunciation patterns learned by the model. Despite these limitations, the TTS module successfully provides understandable spoken responses that allow users to interact with the system without needing to read text.

By providing voice-based responses in the user's native language, the TTS component completes the conversational loop of the TRIEM system. It ensures that the information generated by the AI model is returned to the user in an intuitive and accessible way, supporting the overall goal of creating an inclusive voice-enabled platform for tribal communities.

C. Machine Translation (MT)

The Machine Translation (MT) component acts as a bridge between the Santali language used by the user and the English language used for processing within the system. After the Automatic Speech Recognition (ASR) module converts the spoken Santali input into text, the MT module translates this Santali text into English so that it can be understood and processed effectively by the language model.

Once the system generates a response in English, the MT module performs a second translation step by converting the English response back into Santali. This ensures that the final output is delivered to the user in their native language. Since Santali is a low-resource language with limited translation datasets, multilingual translation models are used to support this process. Through this approach, the system enables smooth communication between the user and the AI model while maintaining interaction in Santali.

D. Large Language Model (LLM)

The Large Language Model (LLM) is responsible for generating meaningful and informative responses to user queries in the TRIEM system. After the user's Santali speech is converted to text and translated into English, the processed query is sent to the LLM for understanding and response generation. The model analyzes the input question and produces a relevant answer based on its knowledge and reasoning capabilities.

In this system, multiple LLM services are integrated to improve reliability and availability. Cloud-based models such as Grok and Gemini are used to generate accurate responses when internet connectivity and API access are available. In

situations where cloud services are unavailable or API limits are reached, the system automatically switches to a locally running model through Ollama. This fallback mechanism ensures that the system can continue functioning even in limited connectivity conditions.

By combining multiple language models with a fallback strategy, the TRIEM system is able to provide consistent and context-aware responses while maintaining continuous operation for users.

III. TRIBAL HISTORY

India has a Significant number of tribes, with the Scheduled Tribes are found in the states of Jharkhand, Odisha, West Bengal, and Chhattisgarh. The tribes have strong cultural heritage and linguistic identities that have been maintained over the years. Languages such as Santali are essential for communication and expression of culture, but they are underrepresented in digital technology.

Although the cultural richness is high, many tribes face difficulties in accessing education, health care, and government social services. With the increasing use of digital technology for public services, language emerges as a critical issue. Digital technology is mostly in English or popular regional languages, and this creates a problem for native speakers of tribal languages to understand and utilize these services. The lack of digital literacy adds to this problem.

Santali, a language with a substantial number of speakers and its own script, Ol Chiki, is classified as a low-resource language in the field of artificial intelligence because of the lack of data and computational resources. Using voice-based AI technology for such languages can be very helpful. It will help in promoting digital inclusion and will also help in preserving the linguistic identity of those communities that have remained underserved in the digital age.

IV. FALL BACK MECHANISM

To ensure reliable operation, the TRIEM system incorporates a fallback mechanism that allows the system to continue functioning even when the primary language model service becomes unavailable. Since the system relies on external APIs for generating responses, issues such as network instability, API rate limits, or service interruptions can occasionally affect performance. To address this, a multi-layer fallback strategy is implemented within the architecture.

When a user query is processed, the system first sends the translated English text to the primary cloud-based language models such as Gemini or Grok. If the system detects an error, timeout, or quota limitation from these services, the request is automatically redirected to a locally hosted model running through Ollama. This local model acts as a backup processing

layer, ensuring that the system can still generate responses without depending entirely on external services.

This fallback design improves the robustness and reliability of the TRIEM system, particularly in environments where internet connectivity may not always be stable. By maintaining both cloud-based and local processing options, the system is able to deliver consistent responses and provide uninterrupted assistance to users interacting in their native language.

V. OBSERVATIONS AND CHALLENGES

TRIEM combines these technologies to facilitate seamless voice-triggered conversations that take place in the user's native language. The system is equipped with fallback logic and safe-response handling to provide responsible and reliable answers.

The Hypothesis behind was not limited to building a voice assistant, but rather to empowering our tribal communities, conserving low-resource languages like Santali and ensuring everyone's digital incorporation. With TRIEM, technology is accessible, inclusive, and integrated for those most digitally underserved.

Another important observation was related to machine translation quality. Translating Santali to English and back to Santali occasionally resulted in minor semantic shifts or simplified expressions. This highlighted the limitations of current multilingual translation models when handling structurally different and low-resource languages. Ensuring that the meaning of the original query remained intact required careful prompt design and response validation.

Since the system integrates multiple components such as ASR, translation, LLM processing, and TTS, the response time may be affected, especially when using cloud-based large language models. To mitigate this issue, caching systems and fallback solutions were incorporated.

Finally, the entire development process has also brought to light the challenge of developing AI systems for low-resource languages. The fact is that low-resource languages have limited data, fewer models to learn from, and evaluation benchmarks that are more complex to work with compared to high-resource languages.

A. Conclusion and Future Work

TRIEM is a multilingual AI voice assistant developed to support Santali-speaking tribal communities through natural voice interaction. The system integrates speech recognition, translation, AI-based response generation, and speech synthesis to provide smooth speech-to-speech communication. It helps reduce language barriers and improves digital accessibility for tribal users in areas such as education, healthcare, and information access.

Future improvements can focus on increasing speech recognition and translation accuracy, supporting more tribal languages, and integrating Retrieval-Augmented Generation (RAG) for better information retrieval. The system can also be expanded with mobile support, offline processing, and improved multilingual AI models for more reliable real-world usage.

ACKNOWLEDGMENT

We would like to express their sincere gratitude to Ms. Meghna K, Assistant Professor, Department of Artificial Intelligence and Data Science, MEA Engineering College, for her continuous guidance, valuable suggestions, and technical support throughout the development of this project.

We also thank the Department of Artificial Intelligence and Data Science, MEA Engineering College, for providing the necessary infrastructure and academic support for successfully carrying out this work.

Finally, we extend our appreciation to all faculty members, friends, and peers who contributed directly or indirectly to the completion of this project.

REFERENCES

- [1] Author et al., "Character Error Rate Estimation for ASR," 2024.
- [2] A. Radford et al., "Multi-Perspective Transfer Learning for MOS Prediction in Low-Resource Languages," 2024.
- [3] R. Bharadwaj et al., "Telugu ASR Fine-Tuning using Wav2Vec2-XLS-R-300," 2025.
- [4] Author et al., "Speech Translation from Darija to Classical Arabic using Fine-Tuned ST Models," 2024.
- [5] X. Lyu et al., "LLM-Based Zero-Shot Streaming TTS System with CosyVoice," 2025.
- [6] Author et al., "Rule-Based Machine Translation System for Hindi-Tulu," 2023.
- [7] D. Oneata et al., "Comparative Analysis of Machine Translation Approaches for Low-Resource Languages," 2024.
- [8] M. Negri et al., "Automatic Dialect Detection for Santali Using a Supervised Autoencoder," 2024.
- [9] LIMMITS Challenge Team, "LIMMITS'24: Multi-Speaker, Multi-Lingual Indic TTS with Voice Cloning," 2024.
- [10] Author et al., "Hola-TTS: Cross-Lingual Zero-Shot Text-to-Speech System," 2025.