INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 08 ISSUE: 05 | MAY - 2024 SJIF RATING: 8.448 ISSN: 2582-3930

SEAMLESS M4T-POWERED SPEECH TO TEXT WITH GRADIO UI

Mr. AJAY R
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology
Coimbatore, India

Mr. DHANA SEKARAN D
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology
Coimbatore, India

Mr. GOWTHAM TJ
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology
Coimbatore, India

Mr. RANJITH KUMAR A
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology
Coimbatore, India

Mr. VASANTH M
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology
Coimbatore, India

Ms. S. NIVEDHA
Assistant Professor
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology
Coimbatore, India

***Abstract*—Speech-to-text translation is pivotal in overcoming language barriers and enhancing communication across cultures. This paper introduces a novel application that leverages the Seamless M4T model and Gradio UI for accurate and efficient speech-to-text translation. The system supports multiple languages, with a focus on English and Tamil, and integrates audio input from diverse sources. Gradio UI's intuitive interface ensures the application's accessibility to a broad audience. The system's performance, evaluated through various test cases, demonstrates its robustness and accuracy, marking a significant advancement in communication technology.**

Keywords*— *Artificial Intelligence, Multilingual Communication, Speech Recognition, Speech-to-Text Translation, User Interface Design, Gradio UI, Seamless M4T.

I. INTRODUCTION

This project is about creating an application that changes spoken words into written text. It's a comprehensive speech-to-text application that uses Seamless M4T and Gradio UI. This application is particularly useful as it allows people to understand each other's thoughts and cultures without the need to learn a new language. It aims to fill the communication gap.

The application uses Seamless M4T, a powerful speech-to-text model that provides exceptional quality and accuracy for transcribing spoken language. It also uses Gradio UI, a user-friendly interface framework that enables a seamless user experience with interactive features. The application accepts audio input through various methods, such as live microphone recording and uploading pre-recorded audio files. Once the speech is analyzed and the corresponding text is extracted, it is displayed on the Gradio UI, allowing users to easily review and interact with the transcribed content.

II. LITERATURE REVIEW

The literature review section expands on the evolution of speech-to-text technologies, emphasizing the challenges faced by previous models in terms of language diversity and real-time processing. It highlights the limitations of existing systems and sets the stage for the introduction of Seamless M4T, which overcomes these hurdles with its advanced language recognition capabilities.

traditional polling data can significantly improve the accuracy of election forecasts, particularly in dynamic political environments. However, challenges such as data

noise, bias, and the evolving nature of social media discourse present ongoing areas of concern. Addressing these challenges requires robust methodologies that can adapt to changing sentiment trends and mitigate the impact of irrelevant or misleading data. By leveraging tools like Text Blob and integrating them with advanced machine learning techniques, researchers can continue to refine election prediction models and provide valuable insights for political stakeholders.

III. EXISTING SYSTEM

The evolution of Speech-to-Text Translation (STT) has been marked by significant technological advancements. Initially, STT systems were based on a cascaded approach where Automatic Speech Recognition (ASR) was used to transcribe speech into text, which was then translated into another language using Machine Translation (MT). This two-step process was often inefficient and prone to errors due to the compounding of inaccuracies at each stage.

With the advent of Direct Speech-to-Text Translation (DSTT), the landscape began to change. DSTT systems aim to translate spoken language directly into text in a different language, bypassing the need for an intermediate transcription step. This direct approach has shown promise in reducing errors and improving the speed of translation.

In addressing the challenges faced by STT systems, several key areas have been the focus of recent developments:

- Modeling Challenges:
  Complex speech signals and the intricacies of linguistic translation require robust model architectures. Encoder-decoder frameworks, including variations of Transformer models, have been at the forefront of tackling these challenges. Multitask learning frameworks have also been explored to enhance the system's ability to handle multiple aspects of the translation task simultaneously.

- Data Scarcity:
  The scarcity of parallel corpora for speech and translation poses a significant hurdle. Innovative solutions such as data augmentation, transfer learning, and the use of multilingual models have been employed to overcome this limitation, enabling the systems to learn from a broader range of data sources.

- Practical Considerations:

Real-world applications of STT systems must contend with issues like real-time translation, accurate segmentation, and the handling of diverse linguistic phenomena such as named entity recognition, gender bias, and code-switching. These practical considerations are critical for the successful deployment of STT systems in various settings.

- End-To-End Models:
  The shift towards End-to-End (E2E) models represents a paradigm change in speech translation. E2E models streamline the translation process by directly converting speech into translated text, which can lead to more contextually aware translations and reduced latency.

The field of STT continues to progress, with ongoing research dedicated to enhancing the robustness, accuracy, and applicability of these systems. As the technology matures, it holds the potential to revolutionize the way we interact across language barriers.

IV. PROPOSED SYSTEM

The application's system architecture is designed with modularity and scalability in mind. The separation of front-end and back-end responsibilities promotes clear code organization, facilitating independent development and testing. This modularity also allows for easy integration with other applications.

At the heart of the application lies the Seamless M4T model, responsible for the core speech-to-text processing. The Gradio UI framework forms the interactive front-end, providing a user-friendly interface for input, output, and control. Communication between the front-end and back-end is handled through well-defined protocols, ensuring seamless data exchange and optimal performance.

FRONT-END:

The front-end serves as the user interface for the application, handling user interactions and presenting the translated text. It is primarily written in Python and utilizes the Gradio UI framework to build a user-friendly interface. The front-end's responsibilities include:

- Accepting audio input from various sources: It allows users to either record audio directly through the microphone or upload pre-recorded audio files.
- Displaying translated text: It presents the translated text in a clear and readable format, enabling users to review and interact with the transcribed content.

- Providing language selection: It offers users the option to choose the target language for translation, catering to multilingual communication needs.

BACK-END:

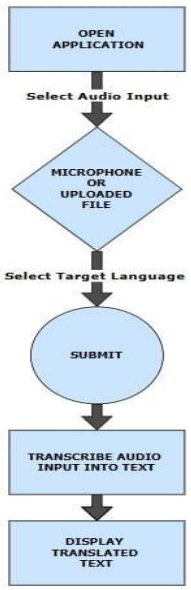
The back-end serves as the processing engine for the application, handling the core tasks of speech transcription and translation. It is primarily written in Python and relies on the Seamless M4T speech-to-text model for accurate transcription and translation. The back-end's responsibilities include:

- Speech recognition: It employs the Seamless M4T model to transcribe the audio input into its corresponding text representation.
- Text translation: It utilizes the Seamless M4T model to translate the transcribed text into the selected target language.

COMMUNICATION BETWEEN FRONT-END AND BACK-END:

The front-end and back-end components communicate directly, enabling seamless data exchange. The front-end sends the audio input to the back-end, and the back-end sends the translated text back to the front-end. This communication mechanism allows for asynchronous processing, ensuring a responsive user experience

USER INTERACTION FLOW DIAGRAM



V. DESIGN AND IMPLEMENTATION

The design and implementation of a Speech-to-Text (STT) system using the SeamlessM4T model and a Gradio UI involves several key steps. First, the SeamlessM4T model, which is a transformer-based model specifically designed for multi-lingual Automatic Speech Recognition (ASR), is used to convert speech input into text. This model is trained on a diverse dataset to accurately transcribe speech in multiple languages.

The implementation process starts with preparing the SeamlessM4T model, which involves loading the pre-trained weights and setting up the model for inference. Next, the Gradio UI is created, defining the input audio widget and the output text display. The audio input is processed and fed into the SeamlessM4T model for transcription. The transcribed text is then displayed in the Gradio UI in real-time.

Finally, the system is tested and evaluated for its accuracy and usability. The SeamlessM4T model's performance is evaluated on a test dataset, measuring metrics such as word error rate (WER) and accuracy. User feedback is also collected to assess the usability of the Gradio UI. The system's performance and user experience are then analyzed to identify any areas for improvement.

VI. DETAILED EXPLANATION OF LIBRARIES USED

- Torch (2.0.1): A deep learning library that accelerates the SeamlessM4T model’s performance, enabling efficient training and inference of neural networks.

- Fairseq2 (0.1): A sequence modeling library that facilitates the use of the SeamlessM4T model for accurate speech and text translation.

- Seamless communication: The core code and data for the SeamlessM4T model, providing access to the Translator class and model files necessary for speech-to-text conversion.

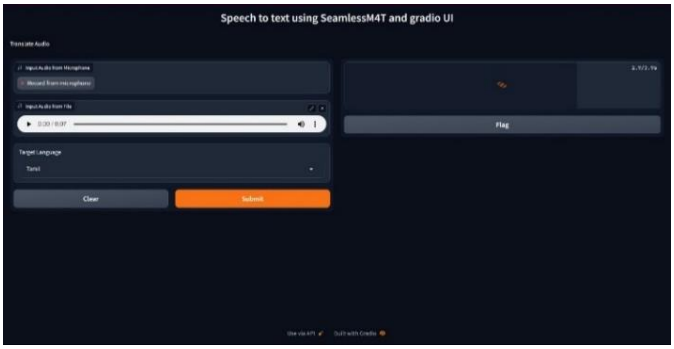
- Gradio (3.50.0): A user interface library that simplifies the creation of web interfaces, allowing users to interact with the speech-to-text function through audio input and output.

- Librosa: An audio analysis and processing library that adjusts the input audio file's speed to match the SeamlessM4T model's requirements.

VII. RESULT AND DISCUSSION

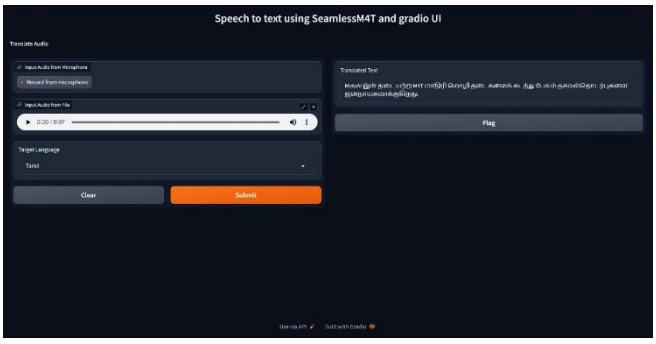
INPUT

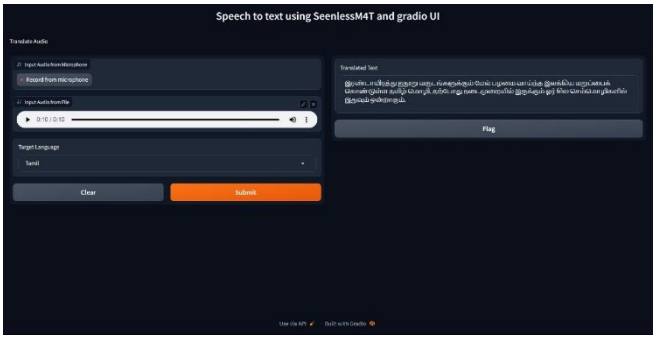
Can interact with the application by recording their voice or uploading audio files. The application supports the option to transcribe speech in either tamil or english.



OUTPUT

Once the speech is transcribed, the text is displayed on the screen via the gradio ui





USER INTERACTION FLOW

The user interaction flow for the speech-to-text translation application is as follows:

- The user opens the application and selects the source of their audio input, either from a microphone or an uploaded audio file.
- The user selects the target language for the translation.
- The user clicks the "**Submit**" button.
- The application transcribes the audio input and displays the translated text.

VIII. CONCLUSION

The speech-to-text translation application developed using Seamless M4T and Gradio UI effectively bridges communication gaps by transcribing spoken language into text. The application seamlessly integrates Seamless M4T's powerful speech-to-text capabilities with Gradio UI's intuitive interface, providing users with a user-friendly and accurate translation tool. The application's support for multiple languages, including Tamil and English, enhances its versatility and caters to a wider range of users.

IX. REFERENCES

[1] **SeamlessM4T: Massively Multilingual & Multimodal Machine Translation** - This paper introduces SeamlessM4T, a unified model supporting speech-to-speech translation, speech-to-text translation, text-to-speech translation, text-to-text translation, and automatic speech recognition for up to 100 languages. It's a significant step forward in the field, especially with its robust performance against background noises and speaker variations.

[2] **Meta AI Blog on SeamlessM4T** - This blog post provides an overview of SeamlessM4T, highlighting its capabilities in automatic speech recognition, speech-to-text translation, and more for nearly 100 languages. It also discusses the challenges of creating a unified multilingual model and the breakthroughs achieved with SeamlessM4T.

[3] **Gradio Real-Time Speech Recognition Guide** - Gradio offers a tutorial on deploying a pretrained speech-to-text model with a Gradio interface. It

includes instructions for creating both full-context and streaming speech recognition systems.

[4] **Speech Recorder and Translator using Google Cloud Speech-to-Text and Translation** **Speech Recorder and Translator using Google Cloud Speech-to-Text and Translation** - This study proposes a speech recorder and translator that combines Automatic Speech Recognition (ASR) and translation technologies to recognize video content and translate it into other languages

[5] **End-to-End Speech-to-Text Translation**: A Survey **End-to-End Speech-to-Text Translation: A Survey**: This survey paper provides an extensive review of end-to-end speech-to-text translation models, discussing the employed models, metrics, and datasets used for speech-to-text tasks.

[6] End-to-End Speech Recognition with Transformers" by Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, and others.

[7] Yasuhisa Fujii, Y., Yamamoto, K., Nakagawa, S., "AUTOMATIC SPEECH RECOGNITION USING HIDDEN CONDITIONAL NEURAL FIELDS", ICASSP 2011: P-5036-5039.

[8] Mohamed, A. R., Dahl, G. E., and Hinton, G., "Acoustic Modelling using Deep Belief Networks", submitted to IEEE TRANS. On audio, speech, and language processing, 2010.