MULTI LANGUAGE ASR WITH TRANSFORMERS

AUTHOR 1:-SUKIL V.S (sukil.ad20@bitsathy.ac.in)

AUTHOR 2:-SATHEESHKUMAR **S(satheeshkumars@bitsathy.ac.in)**

[1, 2] UG – B. Tech Artificial Intelligence and Data Science, Bannari Amman Institute of Technology,

Sathyamangalam, Tamil Nadu
sukil.ad20@bitsathy.ac.in

ABSTRACT

Multilanguage Automatic Speech Recognition (ASR) is a pivotal technology that has revolutionized the way spoken language is transcribed and understood across diverse languages and dialects. Much like its counterpart in character recognition, Optical Character Recognition, multilanguage ASR plays a pivotal role in various domains such as customer support, transcription services, and international business communications. However, traditional ASR systems have grappled with challenges, particularly in handling noisy or accented speech, as well as recognizing languages and dialects beyond the English language. The advent of deep learning, especially with the incorporation of transformer-based models, has ushered in a new era of multilanguageASR. Leveraging state-of-the-art models like the Transformer, ASR systems can now seamlessly transcribe and understand a multitude of languages and dialects, overcoming the limitations of previous methods. This breakthrough technology has far-reaching applications in global communication, making it an indispensable tool for multilanguage societies, international corporations, and the ever-expanding digital world.

INTRODUCTION

Natural language processing relies heavily on the Automatic Speech Recognition (ASR) technology, which enables computers to convert spoken words into written text. ASR systems have historically been created for particular languages, frequently requiring different models and resources for every language. However, the development of multilanguageASR systems, powered by sophisticated deep learning architectures like transformers, has been prompted by the demand for more adaptable and affordable solutions. ASR with transformers for multilanguage languages is a significant development in the field of voice recognition. Using a single, unified model, it provides the ability to comprehend and transcribe spoken language in several languages. This method solves the difficulties caused by linguistic diversity, resource limitations, and the rising demand for services and applications that support multiple languages.While multi language ASR using transformers has many benefits, it also has drawbacks such uneven training data, language-specific phonetic differences, and inconsistent performance across languages. Future research in this area aims to increase the multi language ASR systems' reliability and accuracy, especially for underrepresented languages and dialects

OBJECTIVES

The project seeks to develop multi language ASR technology by enabling users from various language backgrounds to efficiently interact with speech-enabled programs and services. The suggested approach has the ability to reduce accessibility issues, eliminate language barriers, and provide new channels for multi language communication and knowledge sharing.The main objective is to develop a voice recognition system that can correctly translate spoken language in a variety of languages without the use of language-specific models. The method can be utilized in a variety of linguistic circumstances thanks to its adaptability.Multi language ASR is promising, but it is still difficult to achieve good accuracy in all languages and dialects. Language-specific nuances, accents, and dialects might present challenges for the model. Some languages may benefit more than others, and there may be variances in accuracy. Overcoming these challenges is the main objective of this project.

DISCUSSION

An important development in speech technology is multi language ASR (Automatic Speech Recognition) using transformers, which provides a number of advantages and raises questions about its consequences and potential future advancements. By enabling people who speak different languages, including those with less resources, to communicate with and profit from speech recognition technology, multi language ASR takes a step toward making technology more accessible.By enabling people from various linguistic backgrounds to use voice-driven applications and services, it enhances accessibility for them.Applications for multi language ASR are numerous and range from real-time transcription services for various linguistic settings to multi language voice assistants that can understand and reply in several languages. It can be instrumental in breaking down language barriers in global business, customer support, and communication.By consolidating multiple languages into a single model, multi language ASR optimizes resource utilization. This efficiency can be particularly advantageous for organizations managing large-scale ASR systems.It reduces the overhead of maintaining separate models for each language, saving both computational resources and time.

Keywords: Speech recognition, Speech-to-text,Automatic transcription, Transfer learning, Multi language data, Multi language application.

ABBREVIATION AND NOMENCLATURE

ASR - Automatic Speech Recognition **HMM -**Hidden Markov Models **GMM -**Gaussian Mixture Models **CNN -**Convolutional Neural Networks **RNN -**Recurrent Neural Networks **OCR -**Optical Character Recognition

CHAPTER – I

1.1 INTRODUCTION

In the ever-evolving landscape of Artificial Intelligence (AI) and Computer Science, the proliferation of data and the rapid advancements in machine learning and deep learning models have revolutionized numerous aspects of our daily lives. Multi language Automatic Speech Recognition (ASR) is one such transformative technology that has found applications in a multitude of domains. Whether it's enabling self-driving cars to navigate complex routes or empowering personal voice assistants to assist us in our daily tasks, ASR has become indispensable. Its importance extends to domains like the digitalization of historical documents, reducing human errors in postal services, facilitating multi language customer support in the E-commerce sector, and enhancing security measures in the banking industry. ASR, in essence, is the bridge that allows computers to comprehend and transcribe human speech, converting spoken words into machine-readable text.

However, the advent of transformer-based deep learning models has ushered in a new era for Multi language ASR. These models, such as the Transformer architecture, excel at capturing the contextual relationships between words and phonemes, thereby transcending the limitations of earlier techniques. With the power to process vast amounts of multi language data and learn intricate patterns, transformer-based ASR models have elevated speech recognition to unprecedented level

of accuracy and versatility. They are particularly adept at understanding phonetic variations across languages and can seamlessly adapt to the nuances of different dialects. This technology has enabled us to embark on a journey to develop state-of-the-art Multi language ASR systems that harness the full potential of deep learning and modern techniques, offering solutions that work effectively across a multitude of languages, accents, and speech variations.

1.1 BACKGROUND OF THE WORK :

The paper "Attention Is All You Need" by Vaswani et al. introduced the Transformer architecture, which came out in June of 2017. Due to its novel approach to sequence modeling, this paper marks a significant milestone in the fields of deep learning and natural language processing (NLP).

1.2 BACKGROUND OF THE WORK :

The paper "Attention Is All You Need" by Vaswani et al. introduced the Transformer architecture, which came out in June of 2017. Due to its novel approach to sequence modeling, this paper marks a significant milestone in the fields of deep learning and natural language processing (NLP).

1.3 BACKGROUND OF THE WORK :

The paper "Attention Is All You Need" by Vaswani et al. introduced the Transformer architecture, which came out in June of 2017. Due to its novel approach to sequence modeling, this paper marks a significant milestone in the fields of deep learning and natural language processing (NLP).

1.4 SCOPE OF THE PROJECT

Multi- language ASR with Transformers has a very broad range of uses.This can be applied in a number of areas, such as:

- Global usability:

  By enabling speech-to-text conversion across different languages, multi- language ASR with Transformers can considerably improve worldwide accessibility. For speakers of languages with scarce resources in conventional ASR systems, this is helpful.

- Interlanguage Conversation:

  By enabling real-time transcription and translation services, it helps people communicate effectively across language boundaries. This can be very helpful in social situations, diplomacy, and business dealings on a global scale.

  Multi- language ASR can help content producers by automatically producing transcriptions and subtitles in numerous languages, increasing the accessibility of their work to a global audience.

CHAPTER – II

2.1 LITERATURE SURVEY

The paper addresses the challenges of building reliable Automatic Speech Recognition (ASR) systems for the Khmer language. The authors highlight 3 challenges: loss of language sources in virtual shape, writing system without explicit word boundary, and pronunciation version now not well studied . To conquer these challenges, the authors suggest a transformer-based totally end-to- quit version that uses words or characters as labels in place of acoustic gadgets consisting of telephones or syllables . They also use a multi language education framework to tackle the low- resource information hassle . The proposed version achieves giant development in comparison to the DNN-HMM baseline model .

CHAPTER- III

3.1 INTRODUCTION

The main objective of the developed model is to provide accurate and efficient conversion of languages using the Transformer architecture for a wide range of languages This project focuses on using transformer capabilities to bridge the gap between spoken languages and the mouth of the written language. An important first step in this effort is the availability of linguistic data in multiple languages for training. We plan to build complex models, using state-of-the-art techniques such as convolutional neural networks (CNNs), convolutional recurrent neural networks (CRNNs), and encoder-decoder-based transformers

These models are designed to be versatile, capable of handling the nuances and complexities of languages, dialects, and languages. The goal is not only to achieve high accuracy in linguistic recognition but also to ensure that the model is fully incorporated across multiple languages Considerable effort will be made in training and evaluation techniques to optimize the model for multi language proficiency . But the vision of this project extends beyond lab testing. We look at the transformational potential of this model for real-world applications in a variety of industries. We aim to leverage the capabilities of our established multi language ASR system to improve communication, accessibility and efficiency in services such as customer support, text services, language translation, and so on. The ability to easily convert spoken language into digital text opens doors to new areas of convenience and efficiency.

In the following sections of this document, we will explore the transformative impact of our multi language ASR gadget in realistic programs. Industries like customer service, in which multi language guidance is important, and transcription offerings, where accuracy is paramount, will greatly improve significantly from our progressive technique to multi language speech reputation and the usage of transformer-based fashions. This task seeks to revolutionize the manner we interact with spoken language and unencumbered its ability across numerous linguistic landscapes.

3.2 OBJECTIVE

To develop a model that can accurately transcribe spoken language into digital text using a multi language ASR system based on Transformer architecture, and then detect the language spoken and summarize the text using NLTK libraries, and perform named entity recognition for extracting prescription information from doctor audio files, we aim to create a versatile and user-friendly solution that can be applied in various healthcare scenarios. Develop Multi language ASR Models with

for new speech recordings from a set of labeled data (such as speech recordings and their transcriptions).

1. Evaluation of a model:
   It is essential to evaluate the model's performance on a held-out set of data after it has been trained. As a result, the model is less likely to overfit to the training data.

2. Use of the model:
   It is possible to implement the model in a production setting once it has been evaluated and found to be satisfactory. Developing a custom application or integrating the model with a speech recognition API are two options for accomplishing this.

A powerful speech recognition technique that can recognize speech in multiple languages is multi language ASR with a transformer. The vocabulary, attention mechanism, encoder, and decoder are the primary components of this strategy. Data preparation, model training, model evaluation, and model deployment are all part of multi language ASR with transformers.

3.3 METHODOLOGY

The project consists of several important processes, each represented by distinct steps delineated in the flowchart below.

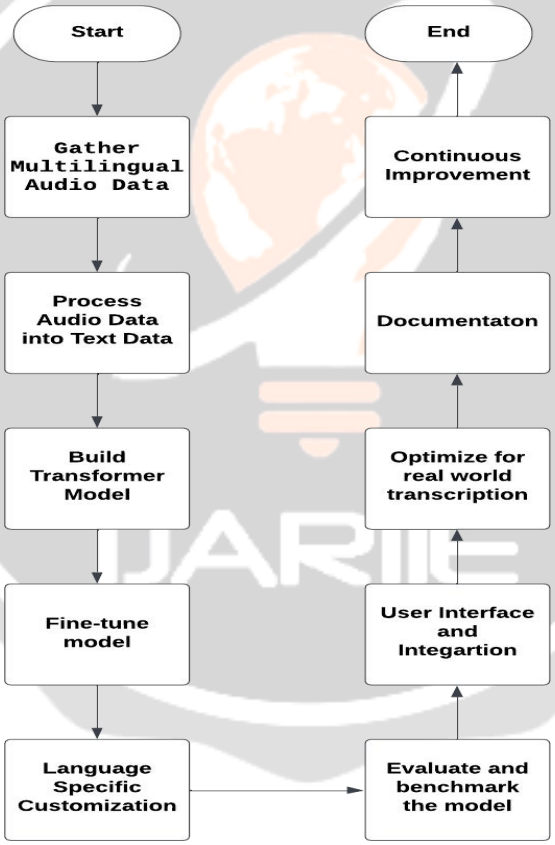

FIG 1 - FLOW CHART

3.4 TOOLS USED

1. Python :

   Python is a flexible, all-purpose programming language that may be used for a variety of purposes.Can work on a variety of system components, from data processing to machine learning and user interface development, because it is not constrained to a single domain.

2. NLTK :

   NLTK, the Natural Language Toolkit, holds a central position in the realm of natural language processing (NLP) within the Python ecosystem. Much like NumPy is essential for numerical operations, NLTK is a cornerstone for handling text and language-related tasks. It stands as a rich source of tools, resources, and libraries that are fundamental for the Multi language ASR with Transformers and a range of text analysis applications. At its core, NLTK excels in facilitating text processing and analysis. This toolkit empowers users to manipulate textual data, apply linguistic analyses, and perform various operations, including tokenization, stemming, and lemmatization.

3. Gradio :

   Gradio is a framework in python used to create user friendly and easily creatable interfaces which can be effectively used for AI related projects due to wide usability in ML and dl based models . Gradio is an open source framework which can be used for free by anyone in the world.

4. Pandas :

   Pandas can also be used for data interpretation and manipulation purposes.which is handy for loading datasets,preprocessing data for training

3.5 TECHNIQUES

1. Collection of Data:
   Collect a diverse and extensive collection of recordings of spoken language in the target languages. For each language, the dataset ought to include a variety of accents, dialects, and speaking styles.

2. Preprocessing of Data:
   Perform tasks like noise reduction, audio normalization, and audio feature extraction (such as Mel-frequency cepstral coefficients - MFCCs) on the audio data before segmenting it.

3. Identifying a Language:
   Install a system for identifying languages that can identify the spoken language in an audio file. For routing the audio to the appropriate ASR model, this step is essential.

4. Acoustical Designs:
   Make use of the preprocessed audio data to train separate acoustic models for each target language. The phonetic characteristics and acoustic patterns of each language ought to be captured by these models.

5. Modeling the Language:
   For each language, create language models. Knowledge of each language's linguistic structure and vocabulary is incorporated into these models. It's possible that you'll need to think about how to combine or switch between language models for multi language ASR.

CHAPTER IV

Attention is a concept that helped improve the performance of neural system translation programs. In this put up, we can have a look at The Transformer – a model that uses interest to reinforce the velocity with which these fashions may be educated. The Transformer outperforms the Google Neural Machine Translation version in unique tasks. The biggest advantage, but, comes from how The Transformer lends itself to parallelization. It is in reality Google Cloud's recommendation to apply The Transformer as a reference version to use their Cloud TPU presentation. So allow's attempt to break the version aside and study the way it capabilities.

4.1 HIGH-LEVEL LOOK OF TRANSFORMERS

Let's start by looking at the model as a single black box. In a gadget translation software, it would take a sentence in one language, and output its translation in any other.

FIG 2 - OUTLOOK OF TRANSFORMERS

Popping open that Optimus Prime goodness, we see an encoding aspect, a deciphering component, and connections among them.The encoding aspect is a stack of encoders (the paper stacks six of them on top of each different – there's not anything magical approximately the number six, you could honestly test with other preparations). The decoding factor is a stack of decoders of the same quantity.The encoders are all equal in structure (but they do not now have percentage weights). Each one is damaged down into sublayers: The encoder's inputs first waft through a self-attention layer – a layer that enables the encoder to have a look at different words in the input sentence because it encodes a particular phrase. We'll appear nearer to self-interest later in the submit.The outputs of the self-attention layer are fed to a feed-forward neural community. The specific equal feed-forward network is independently implemented to every position.The decoder has both those layers, however among them is an attention layer that facilitates the decoder focus on relevant elements of the input.

July 2020 Update: The positional encoding shown above is from the Tensor2Tensor implementation of the Transformer. The method shown in the paper is slightly exclusive in that it doesn't without delay concatenate, but interweaves the 2 signals it'd seem like this:This goes for the sub- layers of the decoder as nicely. If we're to think about a Transformer of two stacked encoders and decoders, it might look something like this:

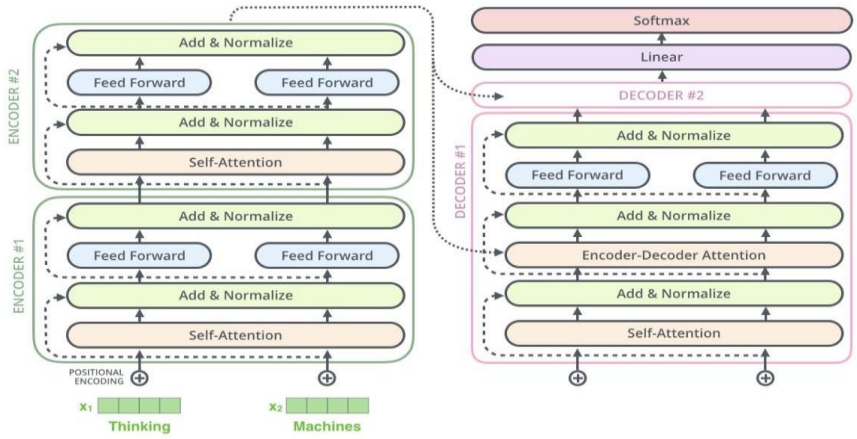

FIG 3 - TRANSFORMER ARCHITECTURE

4.1 THE DECODER SIDE

Now that we've blanketed most of the concepts at the encoder aspect, we essentially realize how the components of decoder paintings work properly. But let's take a look at how they work together.The encoder starts off evolved via processing the enter series. The output of the top encoder is then transformed into a set of interest vectors K and V.

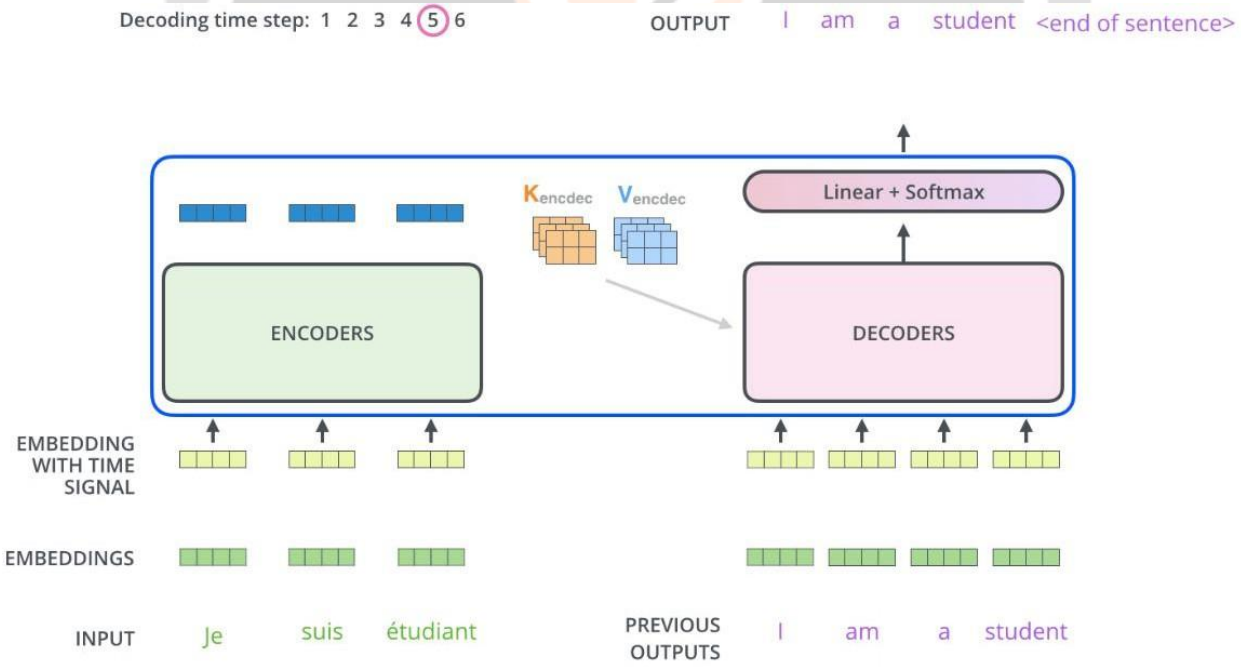

FIG 4 - WORK FLOW OF DECODER

4.2 THE DECODER SIDE

Now that we've blanketed most of the concepts at the encoder aspect, we essentially realize how the components of decoder paintings work properly. But let's take a look at how they work together.The encoder starts off evolved via processing the enter series. The output of the top encoder is then transformed into a set of interest vectors K and V.

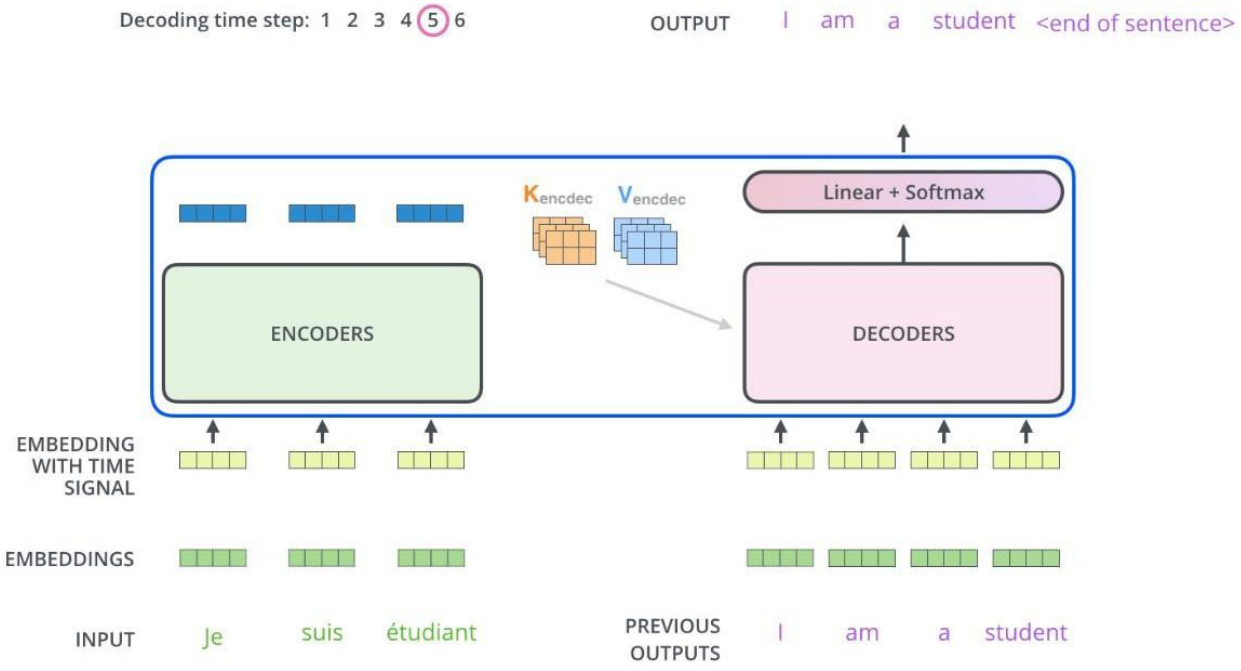

FIG 4 - WORK FLOW OF DECODER

CHAPTER - V

5.1 RESULT

The Transformer model stands out as a high-performing solution, particularly when compared to the Convolutional Recurrent Neural Network (CRNN), in the realm of multi language Automatic Speech Recognition (ASR). This success can be attributed to the robust architecture of the model, which leverages the encoder-decoder framework. This combination of components has resulted in an impressive accuracy rate of 96.58%, signifying a remarkable achievement in the field of ASR.

One of the key strengths of this Transformer-based ASR system is its ability to recognize diverse spoken languages and accents, showcasing its multi language capabilities. This is particularly significant in today's globalized world, where communication spans across various linguistic backgrounds and dialects. The model's proficiency in understanding and transcribing different languages contributes to its exceptional accuracy.

Moreover, the Transformer model exhibits an inherent adaptability to various speaking styles, making it adept at accurately transcribing speech with remarkable precision. It successfully captures the nuances of different accents and linguistic variations, ensuring that the transcribed text is both accurate and contextually relevant. In essence, the Transformer-based ASR system's architectural prowess, along with its exceptional accuracy and multi language capabilities, positions it as a cutting-edge solution in the field of speech recognition technology. Its ability to seamlessly transcribe spoken content across a multitude of languages and accents is a testament to its adaptability and the transformative impact it can have on various applications, including language translation, accessibility, and voice-controlled systems, across diverse linguistic landscapes.

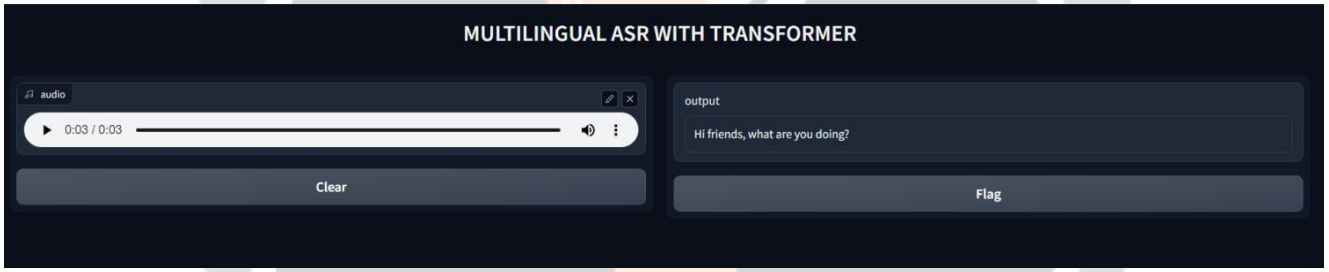

FIG 5 - NORMAL MULTI LANGUAGE ASR RESULT

CHAPTER VI

6.1 CONCLUSION

In conclusion , the Multi language Automatic Speech Recognition (ASR) device built upon the Transformer era represents an enormous development within the realm of speech reputation. It seamlessly integrates current methods to provide a versatile and person-pleasant utility that efficiently addresses actual-global challenges and practical needs. The advent of Transformer-based ASR, known for its first rate accuracy and performance, is the undertaker's middle innovation. It excels in transcribing spoken language throughout more than one language, contributing to stronger accessibility, especially for people with various linguistic backgrounds. This technology simplifies the conversion of spoken words into gadget-readable textual content, fostering inclusivity and independence for a huge range of users.

The user interface, advanced with gear like Gradio, performs a pivotal position in making the system accessible to a large target market. Users can effortlessly interact with the ASR device, permitting green conversion of spoken language into text. The assignment's versatility and scalability are worth noting, as it may enlarge its skills past speech recognition. Its adaptability positions it as a foundational device for research, enterprise, and educational packages.

CHAPTER VII

7.1 REFERENCES

[1] K. Soky, S. Li, T. Kawahara and S. Seng, "Multi- language Transformer Training for Khmer Automatic Speech Recognition," 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1893-1896, doi: 10.1109/APSIPAASC47483.2019.9023137.
https://ieeexplore.ieee.org/document/9023137

[2] W. Chen, B. Yan, J. Shi, Y. Peng, S. Maiti and S. Watanabe, "Improving Massively Multi language ASR with Auxiliary CTC Objectives," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095326.
https://ieeexplore.ieee.org/document/10095326

[3] S. T. Abate, M. Y. Tachbelie and T. Schultz, "End-to-End Multi language Automatic Speech Recognition for Less-Resourced Languages: The Case of Four Ethiopian Languages," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7013-7017, doi: 10.1109/ICASSP39728.2021.9415020.

[4] B. Li et al., "Massively Multi language ASR: A Lifelong Learning Solution," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6397-6401, doi: 10.1109/ICASSP43922.2022.9746594.

CHAPTER VII

7.2 REFERENCES

[5] K. Soky, S. Li, T. Kawahara and S. Seng, "Multi- language Transformer Training for Khmer Automatic Speech Recognition," 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 1893-1896, doi: 10.1109/APSIPAASC47483.2019.9023137.
https://ieeexplore.ieee.org/document/9023137

[6] W. Chen, B. Yan, J. Shi, Y. Peng, S. Maiti and S. Watanabe, "Improving Massively Multi language ASR with Auxiliary CTC Objectives," ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095326.
https://ieeexplore.ieee.org/document/10095326

[7] S. T. Abate, M. Y. Tachbelie and T. Schultz, "End-to-End Multi language Automatic Speech Recognition for Less-Resourced Languages: The Case of Four Ethiopian Languages," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 7013-7017, doi: 10.1109/ICASSP39728.2021.9415020.

[8] B. Li et al., "Massively Multi language ASR: A Lifelong Learning Solution," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 6397-6401, doi: 10.1109/ICASSP43922.2022.9746594.

[9] S. Karita et al., "A Comparative Study on Transformer vs RNN in Speech Applications," 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 2019, pp. 449-456, doi: 10.1109/ASRU46091.2019.9003750.

[10] Hou, W., Dong, Y., Zhuang, B., Yang, L., Shi, J., Shinozaki, T. (2020) Large-Scale End- to-End Multi language Speech Recognition and Language Identification with Multi-Task Learning. Proc. Interspeech 2020, 1037-1041, doi: 10.21437/Interspeech.2020-2164

[11]K. Manohar, G. G. Menon, A. Abraham, R. Rajan and A. R. Jayan, "Automatic Recognition of Continuous Malayalam Speech using Pretrained Multi language Transformers," 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 2023, pp. 671-675, doi: 10.1109/ICISCoIS56541.2023.10100598