International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Speech -To -Text Translation Using Hugging Face Model

Ms. P. SANDHUSTA
Department of
Artificial Intelligence
and Machine Learning
Sri Shakthi Institute of
Engineering and TechnologyCoimbatore, India

Ms. A. K. SANDHIYA
Department of Artificial
Intelligence and Machine
LearningSri Shakthi Institute of
Engineering and Technology
Coimbatore, India

Ms. M. LAKSHITHA
Department of Artificial
Intelligence and Machine
LearningSri Shakthi Institute of
Engineering and Technology
Coimbatore ,India

Ms. S. KAVI PRIYA
Department of Artificial
Intelligence and Machine Learning
Sri Shakthi Institute of
Engineering and Technology
Coimbatore, India

Ms.B. A. SRI SWETHA
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*—The internet has revolutionized communication, but it's particularly challenging for those with hearing impairments. A multilingual speech-to-text conversion system using Hugging Face algorithms is being developed to assist these individuals in various tasks. This system, which uses advanced neural network models and natural language processing algorithms, will translate spoken language into text, enabling seamless communication and engagement in various contexts, thereby enhancing the quality of life for those with hearing impairments.**

***Keywords*—Speech-to-Text (STT)**

INTRODUCTION

Automated voice recognition and speech translation have become major areas of study due to advancements in machine learning and deep learning algorithms. Speech recognition converts spoken words into written text using advanced ASR technology, while machine translation uses sophisticated algorithms and pre-trained models. These systems are trained on vast amounts of bilingual data, allowing them to understand nuances of different languages and produce accurate translations. However, challenges remain, such as recognizing and transcribing speech with accents, background noise, or speech variations. Speech-to-text conversion has practical applications in transcription of interviews, lectures, meetings, and enabling individuals with hearing impairments to access spoken information. It also plays a crucial role in voicecontrolled systems, virtual assistants, and speech-enabled technologies. Hugging Face is an open-source library and platform that offers a variety of pre-trained models for natural language processing (NLP) tasks, including text classification, sentiment analysis, named entity recognition, machine translation, and text generation. The library offers model variants and sizes to accommodate different computational resources and requirements, multilingual models, domain-specific models, and architecture variants. It also provides resources and tutorials for fine-tuning and transfer learning on domain-specific or task-specific datasets. Hugging Face has a vibrant community of developers, researchers, and NLP enthusiasts, which provides a central repository for sharing, discovering, and downloading pre-trained models. The library supports various deployment options, including integration with deep learning libraries, model serving frameworks, and cloud-based platforms. Hugging Face models have gained popularity due to their ease of use, versatility, and performance, making them widely adopted in research and industry.

LITERATURE REVIEW

The study "Implementation of Efficient Speech Recognition System on Mobile Device for Hindi and English Language" by Gulbakshee Dharmale, Dipti D. Patil, and V. M. Thakar in the International Journal of Advanced Computer Science and Applications (IJACSA) discusses the importance of speech recognition in real-time communication. The researchers propose a system that combines the speech recognition approach of existing software with language processing to improve the overall accuracy of speech to text conversion. The proposed Phonetic Model supports multi-lingual speech recognition and achieves 90% accuracy for Hindi and English speech to text recognition. The study also highlights the importance of context-sensitive support in automatic speech recognition (ASR) systems, which can respond to user commands based on existing and previous interactions. P Deepak Reddy, Chirag Rudresh, and Adithya A S from PES University, India, have published an article titled "Multilingual Speech to Text using Deep Learning based on MFCCFeatures. " The paper proposes a novel approach to solving the problem of multilingual speech recognition, which currently has low accuracy in translating sentences containing a mixture of two or more languages. The approach uses a next Word Prediction model in combination with a Deep Learning speech recognition model to accurately recognize and convert the input audio query to text.

.

The authors also propose using cosine similarity between audio features for fast and accurate recognition. The DL speech recognition model in combination with the Word Prediction model achieves an accuracy of 71% when tested on an in-house multilingual dataset. This method aims to lift the language barrier and improve communication between people. In this paper, we apply unsupervised pre-training to improve This paper presents a method for improving supervised speech recognition using unsupervised pre-training. The model, wav2vec, is a convolutional neural network that uses raw audio data to compute a general representation for speech recognition. The goal is contrastive loss, distinguishing true audio samples from negatives. The model uses a fully convolutional architecture for parallelization over time.

PROPOSED SYSTEM

The proposed speech-to-text conversion system uses the Hugging Face model, a leading natural language processing technology, to translate spoken words into written text. The model, trained on extensive datasets, recognizes speech patterns and converts them into accurate text representations. The system is user-friendly, allowing users to speak into a device equipped with the Hugging Face model. Its transformer architecture efficiently processes sequential data, capturing nuanced speech patterns and dependencies. Real-time processing is key, allowing the system to convert each spoken phrase into text on the fly. The system may also incorporate features like punctuation prediction and correction to enhance readability and context. The system is designed for easy integration into various applications and devices, making it accessible to individuals with varying technical expertise.
This paper presents a method for improving supervised speech recognition using unsupervised pre-training. The model, wav2vec, is a convolutional neural network that uses raw audio data to compute a general representation for speech recognition. The goal is contrastive loss, distinguishing true audio samples from negatives. The model uses a fully convolutional architecture for parallelization over time.

METHODOLOGY

i) Import libraries:

The initial step involves installing the necessary libraries, specifically transformers libraries, for working with the Hugging Face model.

ii) Transformers:

Hugging Face Model utilizes transformers, a neural network architecture for natural language processing tasks. The transformers library offers pre-trained models and fine-tuning tools, enhancing speech recognition systems' accuracy and reliability due to its large dataset of spoken language.

iii) Librosa:

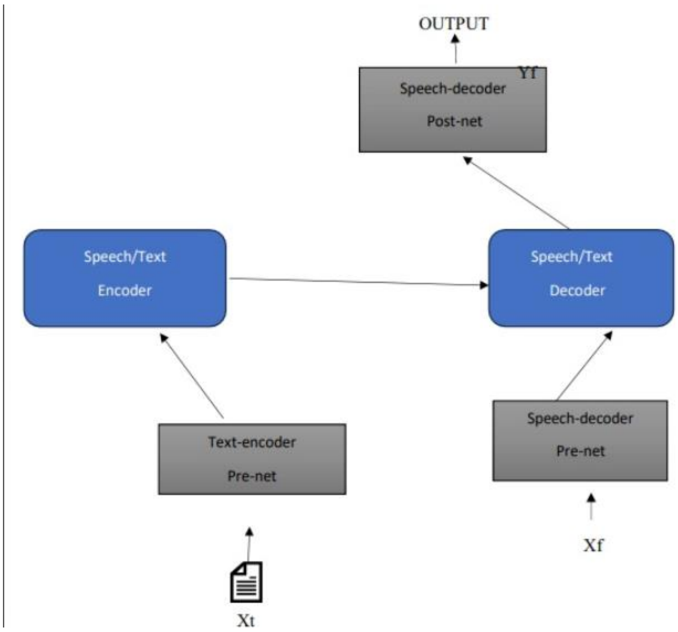
Librosa is a Python library for audio and music analysis. It provides tools and functions to work with audio data, making it particularly useful in the context of speech-to-text conversion.Librosa simplifies the process of loading audio files in various formats(WAV,MP3) provides functions for basic audio processing tasks such as resampling, time-stretching, and pitch-shifting.

iv) PyTorch:

PyTorch is a popular choice for speech-to-text conversion due to its flexibility, dynamic computational graph, ease of experimentation, GPU acceleration support, and pre-trained models and tools.

v) Model Selection:

A pre-trained language model from the Hugging Face Transformers library that is suitable for your task and matches the language of your transcripts. Models such as BERT, GPT, or ROBERT, a depending on the nature of your task (e.g., text classification, text generation).If necessary, fine-tune the selected model on a labelled dataset of speech transcripts to further optimize its performance for your specific task.



vi) Load Model :

Load the pre-trained Wav2Vec2 model and tokenizer from Hugging Face using the from pretrained function. Using Wav2Vec2 model and tokenizer from the transformers library. The model is pre-trained on a large dataset and can be easily loaded using Hugging Face's.

vii) WAV2VEC2:

Wav2Vec2 is a speech model that accepts a float array corresponding to the raw waveform of the speech signal.Wav2Vec2 model was trained using connectionist temporal classification (CTC) So the model output has to be decoded using Wav2Vec2CTCTokenizer.

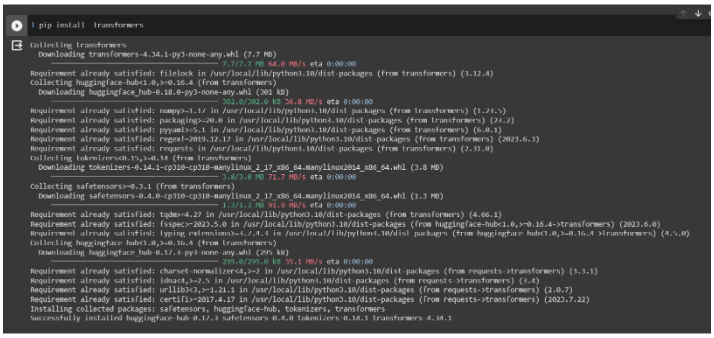
viii) Tokenization:

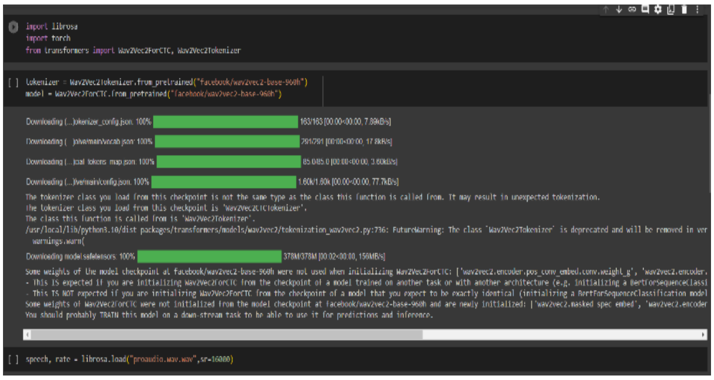
Use the Hugging Face tokenizer to tokenize the speech transcripts into individual tokens. Select an appropriate tokenizer based on your specific requirements and the

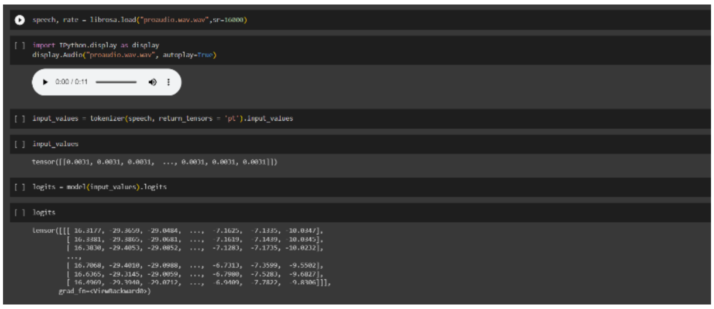
ix) Load and Preprocess Audio data:

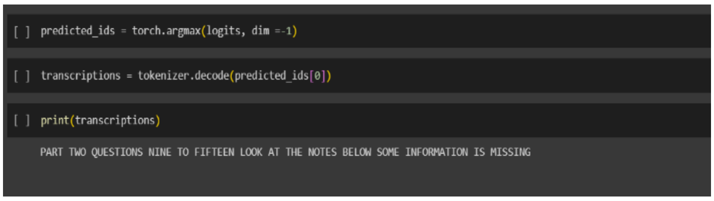
The librosa.load function is used to load an audio file, with a sample rate set to 16000 samples per second for transcription and preprocessing.

x) Fine-Tuning and Iteration:
The evaluation results may necessitate fine-tuning the model, adjusting hyperparameters, or iterating on data preparation. This can be done by training on more labelled data or using domain adaptation techniques. Refer to Hugging Face Transformers documentation for guidance and stay updated.









EXPERIMENTAL OUTPUT

CONCLUSION:

The study focuses on using hugging face models to convert speech into text for deaf people. The model, trained on 3500 audios, achieved a word error rate of 32.42%. The integration of hugging face models into speech-to-text conventions is a significant advancement in language processing capabilities, enhancing accuracy and efficiency. This technology has implications for accessibility, communication, and information dissemination, demonstrating the potential of innovative solutions in shaping the future of language technologies.

Our work helps the deaf people in understanding others speech by converting it to text by using hugging face. The motivation for this kind of study and experiments is the fact that whisper is a simple ASR system for its perfect understandability in human face-to-face communication. Speech Recognition for Hindi Language and Conversion into Hindi Text is done.

The implemented model is trained on a dataset of 3500 audios and achieved a word error rate of 32.42%. It is shown that whisper signals, can give high scores in word recognition for speech. For Open AI whisper small the WER is 87.30% and now our work fine-tuned the model and reduced the Word Error rate to 32.42%. In conclusion, the integration of Hugging Face models into speech-to-text conventions marks a significant leap forward in language processing capabilities. This technology not only enhances accuracy and efficiency but also has profound implications for accessibility, communication, and information dissemination. As we navigate the evolving landscape of artificial intelligence, Hugging Face models stand as a testament to the potential of innovative solutions in shaping the future of language technologies.

REFERENCES

[1] Gulbakshee Dharmale, Dipti D. Patil and V. M. Thakare, “Implementation of EfficientSpeech Recognition System on Mobile Device for Hindi and English Language” InternationalJournal of Advanced Computer Science and Applications(IJACSA), 10(2), 2019.

[2] Multilingual Speech to Text using Deep Learning based on MFCC Features P DeepakReddy, Chirag Rudresh and Adithya A S, PES University, India.

[3] M. T. Tausif, S. Chowdhury, M. S. Hawlader, M. Hasanuzzaman and H. Heickal, "DeepLearning Based Bangla Speech-to-Text Conversion," 2018 5th International Conference onComputational Science/ Intelligence and Applied Informatics (CSII), Yonago, Japan, 2018, pp.49-54, doi: 10.1109/CSII.2018.00016.

[4] A. U. Nasib, H. Kabir, R. Ahmed and J. Uddin, "A Real Time Speech to Text ConversionTechnique for Bengali Language," 2018 International Conference on Computer,Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi,Bangladesh, 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465680.

[5] Wern-Jun Wang, Yuan-Fu Liao, Sin-Horng Chen, RNN-based prosodic modeling for mandarin speech and its application to speech-to-text conversion,Speech Communication,Volume 36, Issues 3–4,2002,Pages 247-265,ISSN 0167- 6393,https://doi.org/10.1016/S0167-6393(01)00006-1.

[6] Software Engineering - A Practitioner’s Approach, Roger S. Pressman, McGraw-Hil

[7] Machine Learning, Tom M. Mitchell, McGraw-Hill International Edition, 6th Edition, 2001

[8]https://huggingface.co/datasets/mozillafoundation/common_voice_11_0

[9] https://huggingface.co/blog/fine-tune-whisper

[10] https://huggingface.co/docs

[11] https://openai.com/research/whisper