SPEECH TO TEXT TRANSLATION USING NATURAL LANGUAGE PROCESSING IN TENSORFLOW

A Micro Project Report

Submitted by

MARREVULA KEERTHI SAI

Reg.no: 99220040120

B.Tech – Computer Science & Engineering, Data Science

Kalasalingam Academy of Research and Education

(Deemed to be University)

Anand Nagar, Krishnankoil - 626 126 February,2024

KALASALINGAM
ACADEMY OF RESEARCH AND EDUCATION
(DEEMED TO BE UNIVERSITY)
Under sec. 3 of UGC Act 1956. Accredited by NAAC with "A++" Grade
Anand Nagar, Krishnankoil, Srivilliputtur (Via), Virudhunagar (Dt) - 626126, Tamil Nadu | info@kalasalingam.ac.in | www.kalasalingam.ac.in

SCHOOL OF COMPUTING
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Bonafide record of the work done by MARREVULA KEERTHI SAI – 99220040120 in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Data Science of the Computer Science and Engineering, during the Academic Year Even Semester (2023-24)

Mr. B. Shanmuga Raja
Project Guide
Assistant Professor
Department of CSE
Kalasalingam Academy of
Research and Education
Krishnan kovil - 626126

Mrs. R. D urga Meena
Faculty Incharge
Assistant Professor
Department of CSE
Kalasalingam Academy of
Research and Education
Krishnan kovil - 626126

Dr. J. Bharath Singh
Evaluator
Associate Professor
Department of CSE
Kalasalingam Academy of
Research and Education
Krishnan Kovil - 626126

Abstract

The abstract discusses the implementation of two important technologies: Speech-to-Text (STT) translation and Natural Language Processing (NLP) using TensorFlow, a popular deep learning framework. For STT translation, the process begins with preprocessing steps like noise reduction and feature extraction to improve the quality of audio input. Then, a deep neural network architecture, often using recurrent neural networks (RNNs) or convolutional neural networks (CNNs), is used to convert these acoustic features into text.

In NLP, TensorFlow is utilized for tasks like understanding, interpreting, and generating human language. The process starts with data preprocessing, which involves tokenization and padding. Word embeddings are then used to represent words as numerical vectors. TensorFlow offers various options for embeddings, including pre-trained ones or custom models. Neural networks, ranging from traditional RNNs to advanced Transformer models, are then constructed based on the specific NLP task, such as sequence labeling or sentiment analysis. During training, loss functions and optimizers are defined, and the model is compiled to learn from labeled data. Evaluation metrics like accuracy and precision are used to assess the model's performance.

Contents

List of Figures

Chapter 1

Introduction

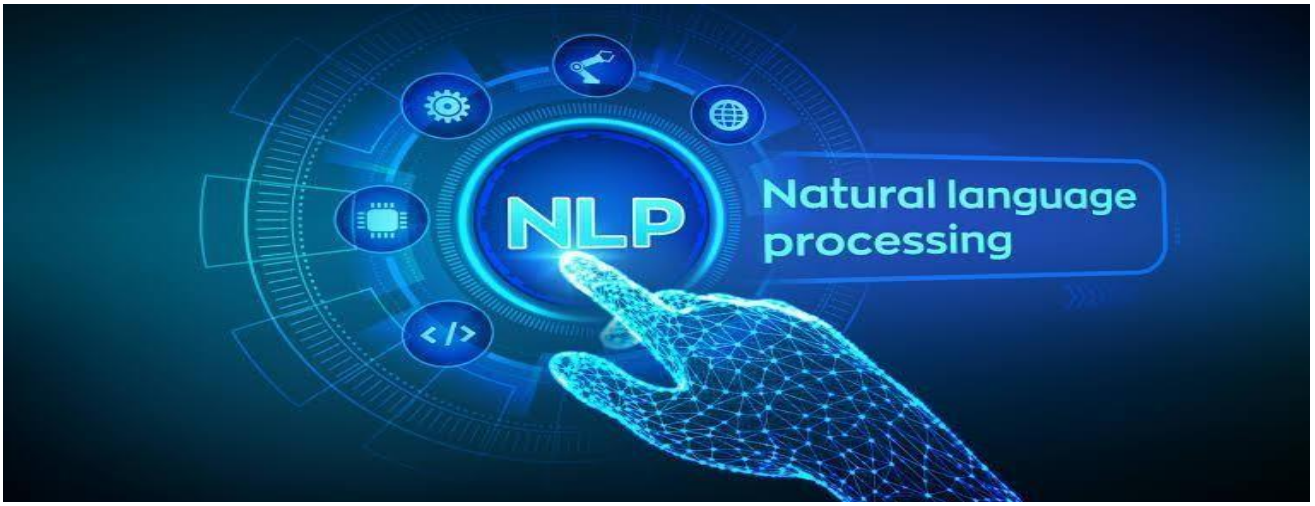
1.1 Natural Language Processing

It enhances communication between computers and human language. Its main objective is to empower computers to comprehend, analyze, and produce human language effectively and meaningfully. Key aspects of NLP include:

Language Comprehension: NLP endeavors to equip computers with the ability to understand human language in various forms, encompassing written text and spoken words. This involves tasks like categorizing text, gauging sentiment, and recognizing speech patterns.

Language Generation: Another focus of NLP is to generate language that resembles human speech, such as creating content automatically, translating between languages, and developing chatbots capable of engaging in natural conversations.

Multilingual Capabilities: NLP is increasingly being applied to multilingual scenarios, enabling systems to function with languages beyond English and adapt to diverse linguistic and cultural environments.



1.1 Natural Language Processing

1.2 Machine Learning

Machine learning(ML) is a branch of artificial intelligence where computers learn from data to make predictions or decisions without being explicitly programmed. It's like teaching a computer to learn from examples rather than giving it specific instructions for every task. It involves algorithms that improve automatically through experience and data.
There are three main types:

1.2.1 TYPES OF MACHINE LEARNING

1. **Supervised Learning :-** It involves models learning from labeled data, where each example has a known outcome. It is akin to teaching with answers provided, commonly utilized in tasks such as recognizing spam emails or predicting house prices.

2. **Unsupervised Learning :-** Involves models looking for patterns or structures in unmarked data It is similar to discovering hidden similarities in a dataset without any guidance, often applied to tasks like customer segmentation or anomaly detection.

3. **Reinforcement Learning :-** An agent learns to make decisions by engaging with its environment, perceiving feedback in the form of rewards or penalties. This method is frequently employed in gaming, robotics, and optimization problems.

Various algorithms and models, including linear regression, decision trees, and neural networks, are utilized in machine learning. Techniques like feature engineering and model evaluation aid in enhancing the performance of these models.

Chapter 2

2 System Work

2.1 Existing Work

Speech-to-text (STT) translation is a technology that transforms spoken words into written text, making communication and document creation easier. Google's Speech-to-Text API is a prime example of this technology, using advanced neural networks to accurately transcribe audio into text. It combines different types of neural networks to understand both the timing and content of speech, even in noisy environments or with different accents. This technology has evolved from simpler methods to complex deep learning models, leading to more precise and adaptable systems useful for various purposes like transcribing services, virtual assistants, and aiding accessibility.

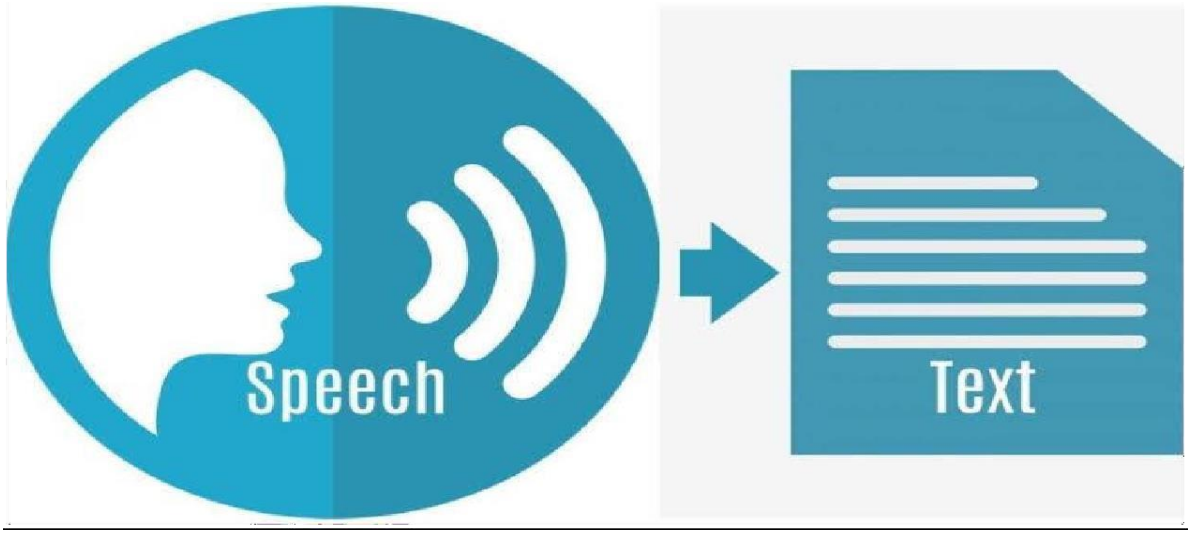


Figure 2.1 : Speech Recognition

2.2 Literature Review

A literature review encompasses a broad range of topics, including various techniques like neural networks, recurrent neural networks, transformers, and attention mechanisms, as well as applications such as sentiment analysis, machine translation, text summarization, and question answering. It also delves into important papers, significant works, recent advancements, challenges, and future directions in the field. NLP systems primarily rely on Deep Learning (DL), a subset of Machine Learning (ML), which provides a flexible and adaptable framework for processing both visual and textual data.

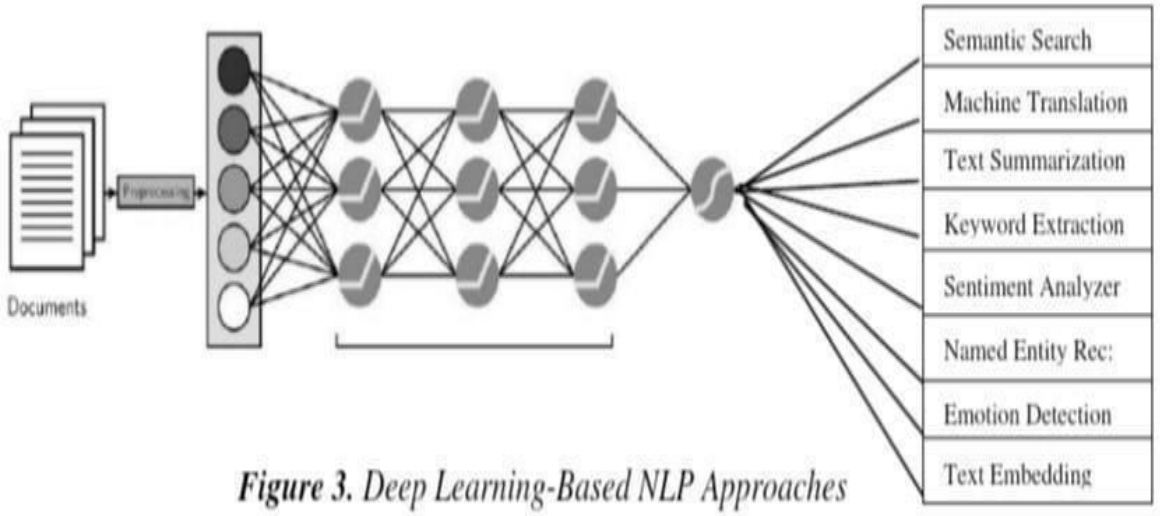


Figure 3. Deep Learning-Based NLP Approaches

2.2 NLP approaches

2.2.1 Advantages of Speech-to-text translation

- **Efficiency:** Instead of typing, you can speak, making transcription much faster.
- **Accessibility:** People with disabilities, especially those with motor impairments, find it easier to communicate through speech-to-text technology.
- **Multitasking:** You can speak while doing other things, boosting productivity.
- **Language Support:** It helps translate languages in r eal - time, bridging co mmunicatio n gaps.
- **Documentation:** It assists in creating accurate records like meeting minutes or lecture notes.
- **Reduced Errors:** Though not perfect, speech recognition tech has improved, leading to fewer mistakes in transcriptions.
- **Hands-free Operation:** Perfect for situations where manual input is difficult or unsafe, such as driving or cooking.

2.2.2 Disadvantages of Speech-to-text translation

- **Accuracy:** Accuracy can be hindered by factors like accents, background noise, and complex speech patterns, leading to errors in transcriptions.
- **Privacy concerns:** Transcribing spoken words into text raises privacy issues, as sensitive information may be recorded and stored.
- **Dependency on internet connection:** Many speech-to-text systems require an internet connection to function, limiting usability in areas with poor connectivity.
- **Limited languages and dialects:** Not all languages and dialects are equally supported, which could be a barrier for users who speak less common languages or dialects.
- **Training requirements:** Some speech recognition systems require extensive training to improve accuracy, which can be time-consuming and resource-intensive.

2.2 How speech to text works using NLP

2.3 Proposed Work

Proposing a system for news classificatio and machine learning involves designing a robust framework in predefined topics or classes. Here's a high-level overview of a proposed system:

Data Collection: Firstly, gather a large amount of audio recordings and their corresponding transcriptions. This forms the basis of your dataset.

Preprocessing: Convert the audio files into a format suitable for input into a neural network. This could involve turning them into spectrograms, which are visual representations of the audio frequencies. Techniques like Fourier transforms and normalization may be used to prepare the data.

Building the Model: Create a neural network architecture, like CN or Transformer models, this model will take the spectrograms (or other audio representations) as input and output the corresponding text transcriptions.

Training: Train the model using the collected dataset. Essentially, you're teaching the model to associate the audio inputs with the correct transcriptions by adjusting its parameters.

Evaluation: Test the trained model on a separate set of data to see how well it performs on new, unseen examples. Common metrics like Word Error Rate or Character Error Rate are used to measure performance.

Fine-Tuning: Based on the evaluation results, make adjustments to the model and its parameters to improve its performance.

Deployment: Once you're satisfied with the model's performance, deploy it in a realworld setting where it can process new audio inputs and provide text transcriptions in real-time.

2.3.1 Features

1. **Text Classification:** Consider it as an automated process of organizing emails into different folders. It enables computers to comprehend and classify text into distinct groups, such as determining whether an email is spam or not, analyzing sentiments in social media posts, or arranging news articles based on their topics.

2. **Named Entity Recognition (NER):** This can be likened to possessing a special ability to extract specific information from a bulk of text. It has the capability to identify and extract significant entities like names of individuals, locations, organizations, or dates from text, which proves useful for tasks like extracting information from news articles or analyzing trends in social mediaposts.

3. **Sentiment Analysis :**Envision a scenario where a computer can analyze thousands of product reviews and inform you whether people liked the product or not. It can achieve this by examining the sentiment expressed in the text, determining whether it is positive, negative, or neutral. This proves highly beneficial in understanding customer feedback, social media trends, or the overall sentiment surrounding a particular topic.

4. **Machine Translation:** It aids in overcoming language barriers by translating text from one language to another. Whether it involves converting a webpage from French to English or facilitating real-time conversation translation, NLP techniques like neural machine translation enable computers to comprehend and produce accurate translations.

3 Chapter 3

Implementation

3.1 Data Collection

1. Begin by collecting speech data, capturing a range of recordings of individuals speaking in the desired language. Ensure diversity in topics, accents, and speech patterns to enhance the model's adaptability to various scenarios.

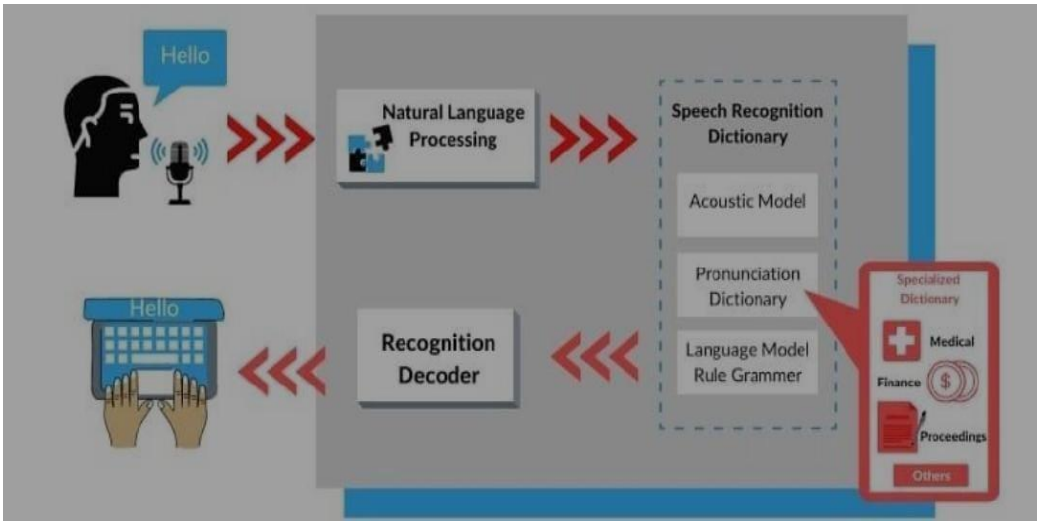
2. Following data collection, transcribe the spoken recordings into written text. This can be achieved through manual transcription or utilizing specialized software for automatic speechto-text conversion.

3. Conduct data cleaning post-transcription to rectify any errors, inconsistencies, or irrelevant segments. This step is crucial to maintain the accuracy and reliability of the data used for model training.

4. Prior to proceeding further, validate and evaluate the quality of the collected data.

5. Divide the data into three distinct sets: one for model training, another for fine-tuning and monitoring performance during training, and a final set for testing the model's overall performance. This segmentation prevents the model from simply memorizing the training data.

6. Throughout the entire process, consider ethical implications such as privacy, consent, and bias. Ensure that data usage is authorized and that the model interacts with individuals in a fair and respectful manner.



3.1 Audio and speech data collection

3.1.1 Experimental Analysis

Data :-

Experimental analysis data for speech-to-text translation involves assessing how accurately and quickly a system converts spoken words into text. Key metrics like word error rate (WER), character error rate (CER), accuracy, and processing time are used. These metrics help evaluate the effectiveness of different speech-to-text systems across different scenarios, like different languages, accents, or background noise levels. Researchers typically rely on established datasets like Libri Speech or Common Voice to conduct experiments and compare the performance of various algorithms or models.

Code :-

```
import tkinter as tk
from tkinter import ttk
import threading
```

```
import speech_recognition as sr
# Language codesfor different languages
language_codes = {
'Telugu': 'te-IN',
'Tamil':'ta-IN',
'English': 'en-US',
'Hindi': 'hi-IN',
'Kannada': 'kn-IN',
'Malayalam': 'ml-IN'
}
def perform_speech_to_text():
language = language_codes[language_var.get()]
# Initialize the recognizer
recognizer = sr.Recognizer()
# Capture audio from the microphone
with sr.Microphone() as source:
print("Please speak something...")
audio_data = recognizer.listen(source)
try:
# Use Google Web Speech API to convert speech to text
text = recognizer.recognize_google(audio_data, language=language)
root.after(100, lambda: result_text.set("You said: " + text))
except sr.UnknownValueError:
        root.after(100, lambda:result_text.set("Could not understand audio"))
except sr.RequestError as e:
        root.after(100, lambda:result_text.set("Error: Could not request
  results;{0}".format(e)))
def speech_to_text():
# Run speech-to-text in a separate thread to keep the GUI responsive
```

```
threading.Thread(target=perform_speech_to_text).start()
# GUI setup
root = tk.Tk()
root.title("Speech to Text Translator")
root.geometry("500x500") # Set window size
# Set the background color to sky blue
root.configure(bg='sky blue')
# Create a style
style = ttk.Style()
style.configure('TLabel',font=('Arial', 14),background='sky blue')
style.configure('TButton', font=('Arial', 14))
# Language selection dropdown
ttk.Label(root, text="Select Language:", style='TLabel').pack(pady=10)
language_var = tk.StringVar()
language_dropdown = ttk.Combobox(root,
textvariable=language_var,
values=list(language_codes.keys()), font=('Arial', 12))
language_dropdown.pack(pady=10)
language_dropdown.set('English')
# Result label
result_text = tk.StringVar()
ttk.Label(root, textvariable=result_text, style='TLabel').pack(pady=10)
# Button to trigger speech-to-text conversion
convert_button = ttk.Button(root, text="Convert", command=speech_to_text,
style='TButton')
convert_button.pack(pady=10)
# Start the GUI main loop
root.mainloop()
```

3.1.2 Result

To summarize, to convert speech to text, you'd first set up a speech recognition system using tools like Python's Speech Recognition module. Then, you capture audio either from a microphone or a file and process it through the system. The system analyzes the audio, turning it into text based on speech patterns. You retrieve this text output, which you can then customize further to match your application's needs.

- **Setting up the system:** First, you need to prepare a system that can understand spoken words.

  You can use tools like the Speech Recognition module in Python.

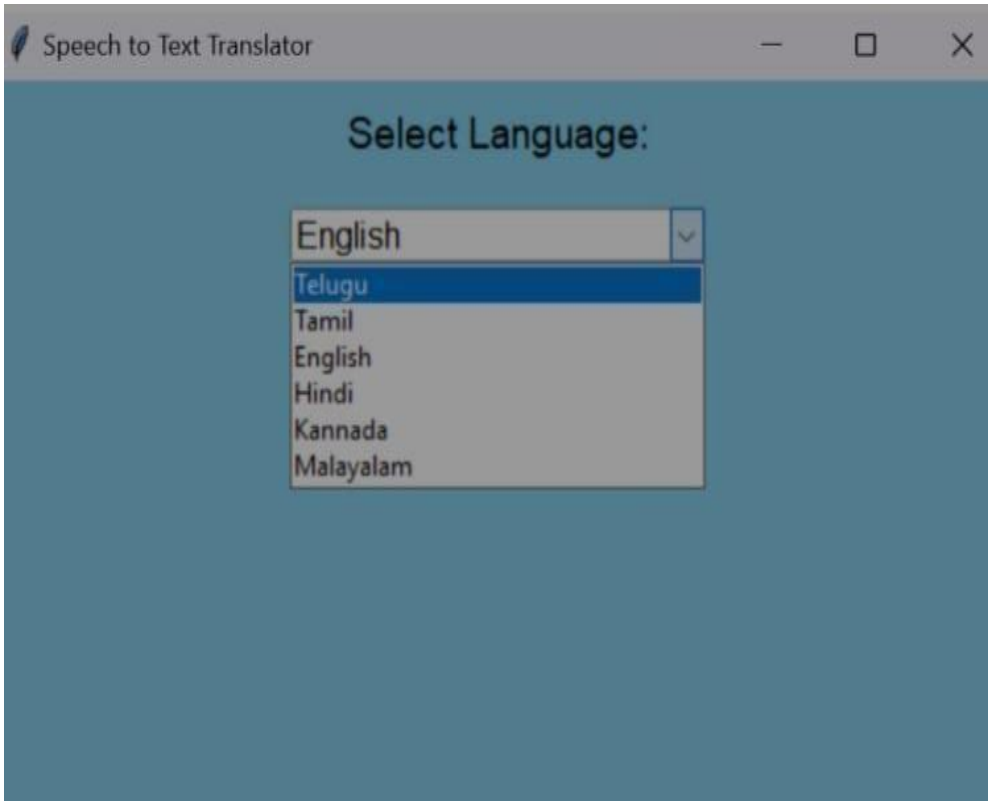
- **Getting the audio:** Next, you gather the spoken words either through a microphone or from a pre-recorded audio file. If you're using a microphone, you need to convert the sound it captures into a format that the system can understand.

- **Processing the audio:** Then, you send this audio to the system you set up earlier. It listens to the audio, recognizes the words, and turns them into text.

- **Getting the text:** The system then gives you the text version of what it heard. You can save this text or use it in your application.

- **Customizing the output:** If you want to change the text to better fit your needs, you can do that here. This might mean simplifying it, summarizing it, or even rewriting it entirely.

By following these steps, you can use speech-to-text technology to convert spoken words into text, and with some additional processing.

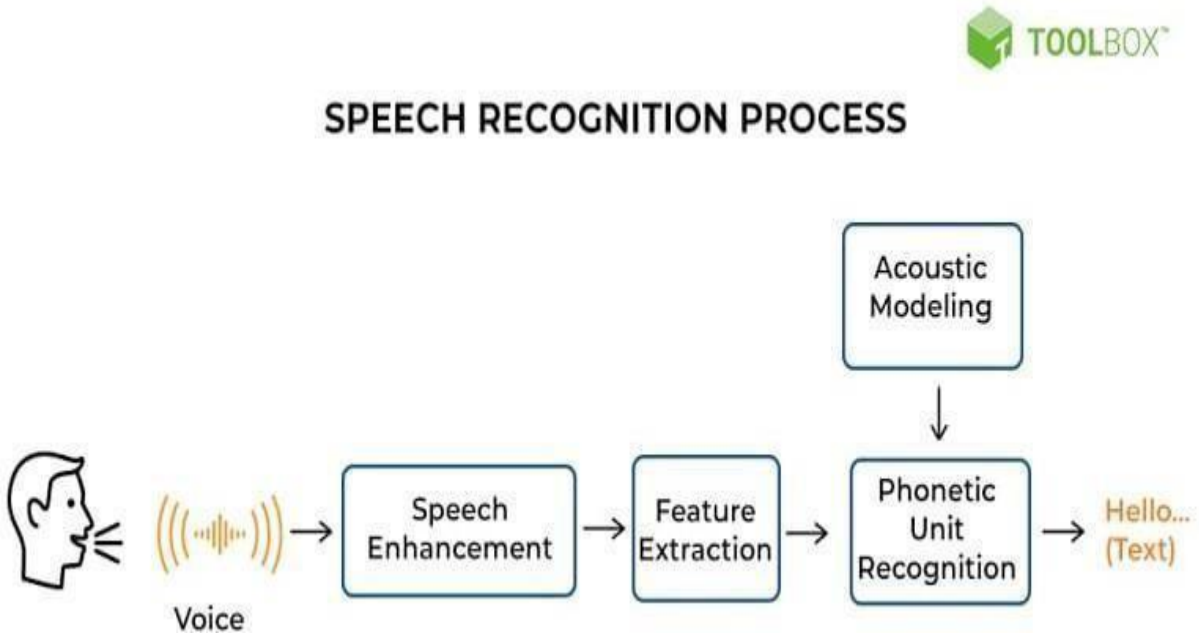


3.1 GUI interface

3.2 Sample Output

3.2 Step by step process of how speech-to-text works

1. **Audio Capture:** First, the system listens to the spoken words using a microphone or similar device.
2. **Preprocessing:** Any background noise or distortion is removed from the audio.
3. **Feature Extraction :**Key aspects of the sound, like pitch and frequency, are identified.
4. **Acoustic Modeling :**A model learns to recognize basic speech sounds (phonemes) from the extracted features.
5. **Language Modeling:** Another model predicts which words are likely based on recognized phonemes and language rules.
6. **Decoding:** Phonemes are transformed into words using algorithms to determine the most probable word sequence.
7. **Postprocessing:** The transcribed text may be refined to correct errors, like spell-checking.
8. **Text Output:** Finally, the system presents the transcribed text to the user, ready for use or further processing.



3.2 Speech recognition process

Chapter 4

Conclusion and Future Work

4.1 Conclusion

It is an impressive technology that transforms spoken words into written text. This process involves analyzing the speech signal, identifying characteristics, and utilizing language modeling techniques to generate the corresponding text. The technology has a wide range of applications and improves communication and accessibility. Over time, speech-to-text translation has advanced in accuracy, language support, and real-time capabilities, simplifying communication, particularly for individuals with disabilities.

In summary, despite the significant progress made in speech-to-text translation with NLP, ongoing research is crucial to overcome challenges and enhance the precision and dependability of these systems. With continuous technological advancements and enhancements in NLP algorithms, the future holds the potential for even more sophisticated speech-to-text translation capabilities.

4.2 Future Work

We are planning to create to the interface which translates the speech from one language to text from another language with some more languages.

In the future, advancements in speech-to-text translation using NLP are likely to focus on improving accuracy, especially in understanding complex sentences and context. One area is improving the accuracy and robustness of speech recognition systems. Researchers are continuously working on developing better algorithms and models to handle different accents, background noise, and variations in speech patterns. This will make speech to text translation more reliable and effective in various real-world scenarios. Another area of focus is multilingual speech to text translation. As technology advances, there is a growing need for speech recognition systems that can accurately transcribe andtranslate speech in multiple languages.

Chapter 5

References

1. Dario Amodei et al. (2016) present Deep Speech 2, a deep learning-based approach for speech recognition, in their paper titled "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin."

2. In their paper titled "Listen, Attend and Spell" (2016), William Chan et al. introduce an attention based neural network model designed for sequence-to-sequence speech recognition.

3. Ashish Vaswani et al. (2017) introduce the Transformer model in their paper "Attention is All You Need," which has gained significant popularity in the field of NLP for tasks like machine translation.

4. Jacob Devlin et al. (2018) present BERT, a pre-trained language representation model that demonstrates exceptional performance in various NLP tasks such as text classification and named entity recognition, in their paper "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding."

Chapter 6

Certification

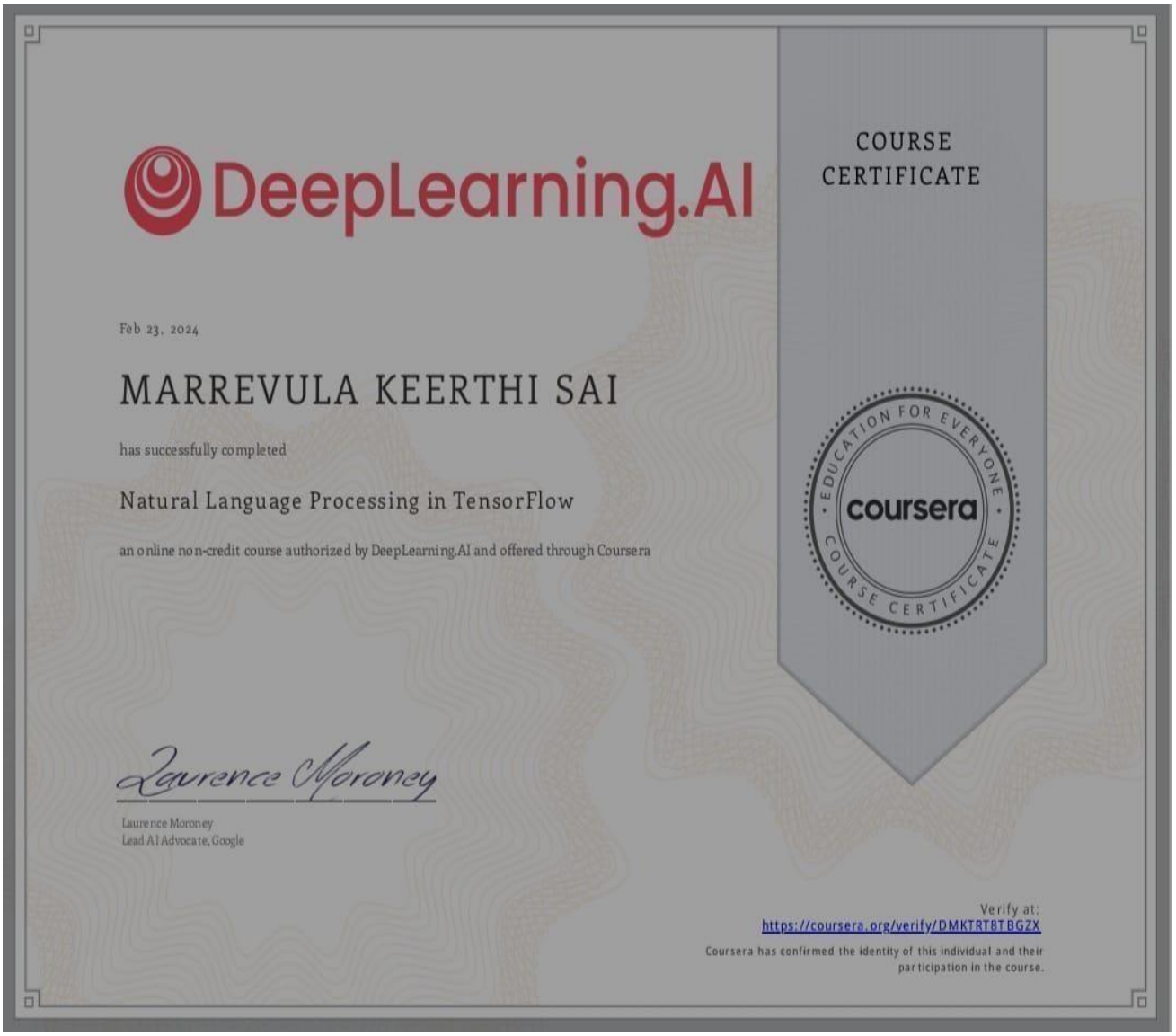


Figure 6.1: Certification details