Report on

Speech To Text Recognition

By

Sandali Srivastava(202410116100179)
Sakshi Tripathi(202410116100176)

Session:2024-2025 (II Semester)

Under the supervision of

Mrs. Komal Salgotra

KIET Group of Institutions, Delhi-NCR, Ghaziabad

Department Of Computer Applications
KIET GROUP OF INSTITUTIONS, DELHI-NCR,
GHAZIABAD-201206
(2025)

TITLE: Speech To Text Recognition

INTRODUCTION:

Speech-to-text conversion, also known as automatic speech recognition (ASR), is a core technology in artificial intelligence (AI) that enables machines to convert spoken language into written text. This innovation bridges the gap between human communication and computer understanding, allowing users to interact with devices more naturally through voice.

The process involves several AI techniques, including signal processing, natural language processing (NLP), and machine learning. Modern speech recognition systems use deep learning models, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and more recently, transformers, to improve accuracy and adapt to different accents, languages, and noisy environments.

Speech-to-text technology powers a wide range of applications, from virtual assistants like Siri and Alexa to real-time transcription services, voice-controlled devices, and accessibility tools for individuals with disabilities.

As AI continues to evolve, speech recognition systems are becoming more efficient, accurate, and context-aware, revolutionizing how humans interact with technology in everyday life.

Objective of the Project:

The primary objective of speech-to-text conversion is to accurately and efficiently transform spoken language into written text using artificial intelligence. This technology aims to enable seamless and natural interaction between humans and machines through voice input, eliminating the need for manual typing or physical interfaces.

Key goals include:

- **Enhancing accessibility** for individuals with physical or visual impairments.
- **Improving productivity** by enabling hands-free operation in various environments (e.g., driving, healthcare, customer service).
- **Facilitating real-time communication** and transcription for meetings, lectures, and broadcasts.
- **Supporting multilingual and accent-agnostic recognition** to cater to diverse user groups.
- **Integrating voice interfaces** into devices and applications for more intuitive user experiences.

How the Model Works :

The speech-to-text process involves several stages, combining signal processing and machine learning especially deep learning models to convert spoken audio into written text. Here's how it typically works:

1. Audio Input

- The system captures spoken words using a microphone.
- The input is usually in waveform format (a raw audio signal).

2. Preprocessing

- The audio is cleaned to reduce noise.
- It’s split into smaller segments (called frames).
- Features such as Mel Frequency Cepstral Coefficients (MFCCs) or spectrograms are extracted to represent sound patterns.

3. Acoustic Model

- This model learns the relationship between audio features and phonemes (smallest units of sound).
- Deep learning architectures like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, or transformers are used to model temporal dependencies in speech.

4. Language Model

- This model understands the structure and grammar of a language.
- It helps predict word sequences, improving accuracy and context understanding (e.g., distinguishing "their" vs. "there").

5. Decoder

- Combines the outputs of the acoustic and language models to find the most likely transcription.
- Uses algorithms like **beam search** to select the best word sequence.

6. Text Output

- The final predicted sequence of words is displayed as written text.

Example:

Audio → Feature Extraction → Acoustic Model → Language Model → Text Output

METHDOLOGY:

The methodology for converting speech into text involves a sequence of computational steps, integrating signal processing, machine learning, and natural language processing. Below is a breakdown of the main phases:

1. Audio Acquisition

- Speech input is captured via a microphone or audio file.
- The audio signal is usually sampled at a fixed rate (e.g., 16 kHz).

2. Preprocessing

- **Noise reduction**: Filters are applied to remove background noise.
- **Normalization**: Volume levels are standardized.
- **Segmentation**: Audio is divided into manageable frames (typically 20–40 milliseconds).

3. Feature Extraction

- Converts audio signals into a form suitable for machine learning.
- Common features include:
    - MFCC (Mel Frequency Cepstral Coefficients)
    - Spectrograms

    - Log Mel Filterbanks

- These features represent the frequency and energy patterns of speech.

4. Acoustic Modeling

- Maps audio features to phonemes (basic sound units).
- Uses deep learning models such as:
    - CNNs (for spatial features)
    - RNNs/LSTMs/GRUs (for sequential data)
    - Transformers (for context and long-range dependencies)

5. Language Modeling

- Predicts the most likely word sequences.
- Improves accuracy by incorporating grammar, syntax, and context.
- Can be based on:
    - N-gram models
    - RNN-based models
    - Transformer-based models (e.g., BERT, GPT)

6. Decoding

- Combines outputs from the acoustic and language models.
- Applies search algorithms (e.g., beam search) to generate the most probable text.
- Includes error correction and confidence scoring.

7. Postprocessing

- Converts text into grammatically correct, readable format.
- Adds punctuation, capitalization, and handles special tokens (e.g., numbers, abbreviations).

CONCLUSION :

Speech-to-text conversion is a transformative application of artificial intelligence that enables machines to understand and transcribe human speech with high accuracy. By combining signal processing, deep learning, and natural language processing, this technology has made voice-driven interaction more accessible, efficient, and natural.

From virtual assistants to real-time transcription services, speech-to-text systems are being widely adopted across industries, improving accessibility, productivity, and user experience. As AI continues to evolve, future advancements will bring even greater accuracy, support for multiple languages and dialects, and seamless integration into our everyday lives.

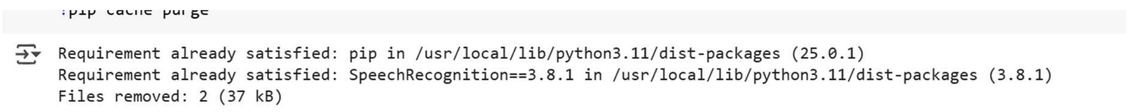
Ultimately, speech-to-text technology is not just about recognizing words—it's about bridging the gap between human expression and digital understanding.

CODE :

```
import speech_recognition as sr

!pip install --upgrade pip

!pip install SpeechRecognition==3.8.1

!pip cache purge

!pip install pydub

from pydub import AudioSegment

def recognize_speech(audio_file):

    # Convert mp3 to wav

    sound = AudioSegment.from_mp3(audio_file)

    sound.export("temp.wav", format="wav")

    # Now use the wav file for speech recognition

    recognizer = sr.Recognizer()

    with sr.AudioFile("temp.wav") as source:

        audio_data = recognizer.record(source)

        try:

            text = recognizer.recognize_google(audio_data)

            return text

        except sr.UnknownValueError:
```

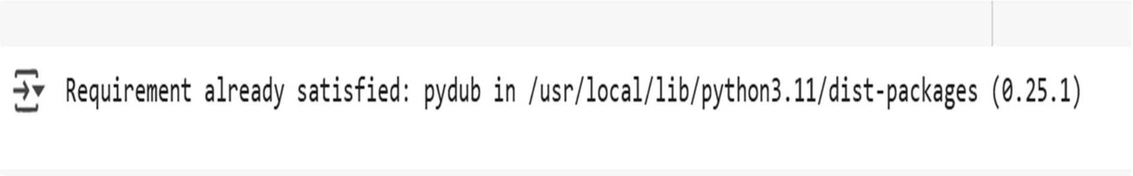
```
return "Could not understand audio"
        except sr.RequestError as e:
            return f"Could not request results from Google Speech
Recognition service; {e}"
from google.colab import files
uploaded = files.upload()
audio_file = list(uploaded.keys())[0]
recognized_text = recognize_speech(audio_file)
print("Recognized Text:", recognized_text)
!pip install librosa
import librosa
!pip install matplotlib
import librosa.display
import matplotlib.pyplot as plt
audio_data, sample_rate = librosa.load(audio_file)
plt.figure(figsize=(12, 4))
librosa.display.waveshow(audio_data, sr=sample_rate)
plt.title('Audio Waveform')
plt.xlabel('Time (s)')
plt.ylabel('Amplitude')
plt.show()
```

SCREENSHOT :

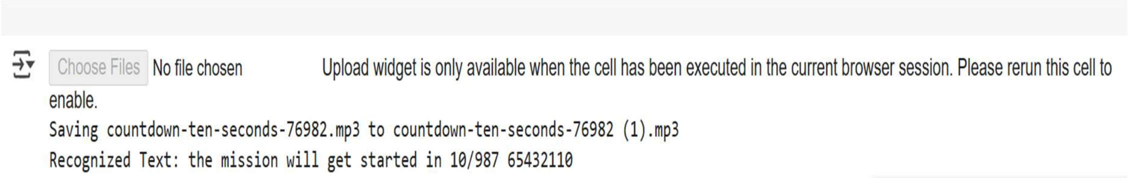


```
!pip cache purge
```

```
Requirement already satisfied: pip in /usr/local/lib/python3.11/dist-packages (25.0.1)
Requirement already satisfied: SpeechRecognition==3.8.1 in /usr/local/lib/python3.11/dist-packages (3.8.1)
Files removed: 2 (37 kB)
```

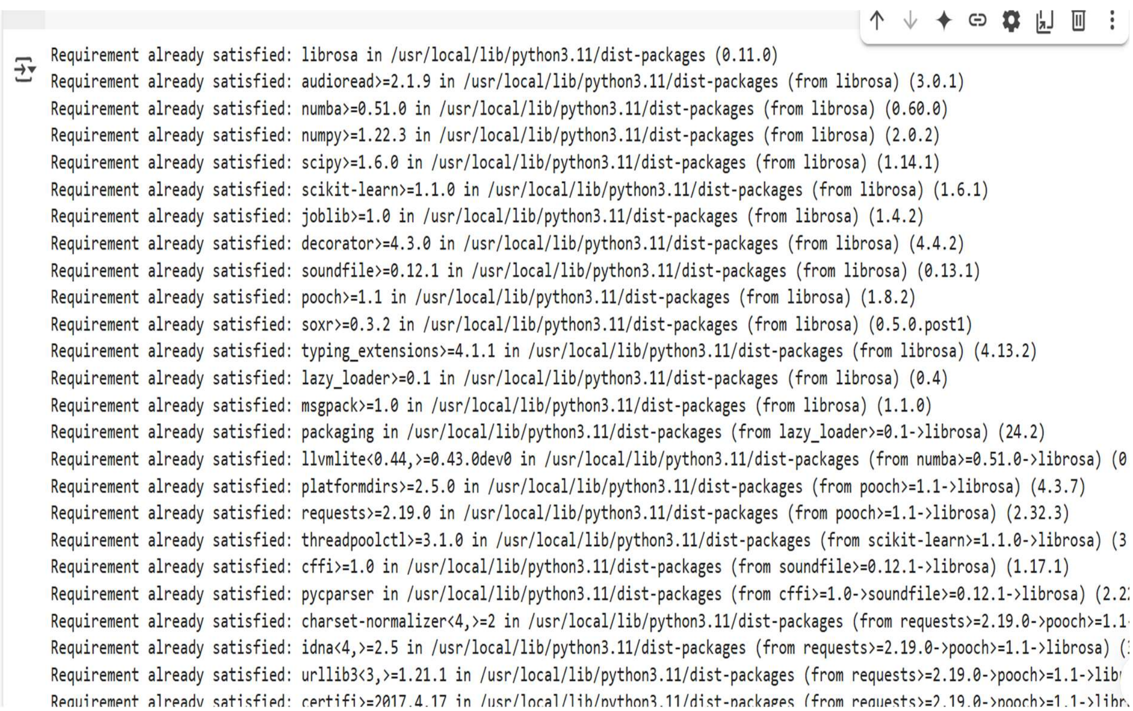


```
Requirement already satisfied: pydub in /usr/local/lib/python3.11/dist-packages (0.25.1)
```

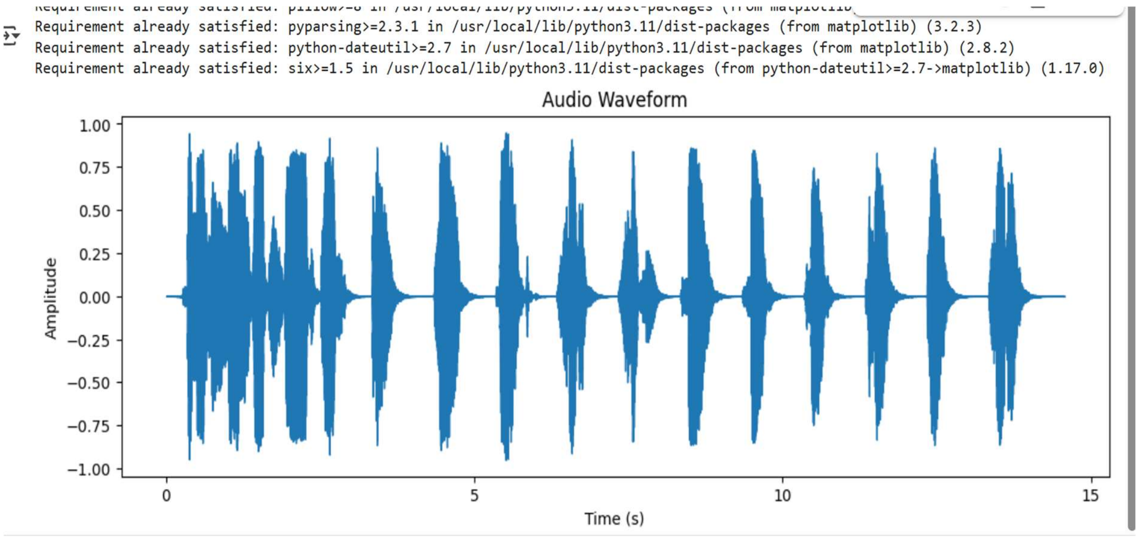


Upload widget is only available when the cell has been executed in the current browser session. Please rerun this cell to enable.

```
Saving countdown-ten-seconds-76982.mp3 to countdown-ten-seconds-76982 (1).mp3
Recognized Text: the mission will get started in 10/987 65432110
```



```
Requirement already satisfied: librosa in /usr/local/lib/python3.11/dist-packages (0.11.0)
Requirement already satisfied: audioread>=2.1.9 in /usr/local/lib/python3.11/dist-packages (from librosa) (3.0.1)
Requirement already satisfied: numba>=0.51.0 in /usr/local/lib/python3.11/dist-packages (from librosa) (0.60.0)
Requirement already satisfied: numpy>=1.22.3 in /usr/local/lib/python3.11/dist-packages (from librosa) (2.0.2)
Requirement already satisfied: scipy>=1.6.0 in /usr/local/lib/python3.11/dist-packages (from librosa) (1.14.1)
Requirement already satisfied: scikit-learn>=1.1.0 in /usr/local/lib/python3.11/dist-packages (from librosa) (1.6.1)
Requirement already satisfied: joblib>=1.0 in /usr/local/lib/python3.11/dist-packages (from librosa) (1.4.2)
Requirement already satisfied: decorator>=4.3.0 in /usr/local/lib/python3.11/dist-packages (from librosa) (4.4.2)
Requirement already satisfied: soundfile>=0.12.1 in /usr/local/lib/python3.11/dist-packages (from librosa) (0.13.1)
Requirement already satisfied: pooch>=1.1 in /usr/local/lib/python3.11/dist-packages (from librosa) (1.8.2)
Requirement already satisfied: soxr>=0.3.2 in /usr/local/lib/python3.11/dist-packages (from librosa) (0.5.0.post1)
Requirement already satisfied: typing_extensions>=4.1.1 in /usr/local/lib/python3.11/dist-packages (from librosa) (4.13.2)
Requirement already satisfied: lazy_loader>=0.1 in /usr/local/lib/python3.11/dist-packages (from librosa) (0.4)
Requirement already satisfied: msgpack>=1.0 in /usr/local/lib/python3.11/dist-packages (from librosa) (1.1.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.11/dist-packages (from lazy_loader>=0.1->librosa) (24.2)
Requirement already satisfied: llvmlite<0.44,>=0.43.0dev0 in /usr/local/lib/python3.11/dist-packages (from numba>=0.51.0->librosa) (0
Requirement already satisfied: platformdirs>=2.5.0 in /usr/local/lib/python3.11/dist-packages (from pooch>=1.1->librosa) (4.3.7)
Requirement already satisfied: requests>=2.19.0 in /usr/local/lib/python3.11/dist-packages (from pooch>=1.1->librosa) (2.32.3)
Requirement already satisfied: threadpoolctl>=3.1.0 in /usr/local/lib/python3.11/dist-packages (from scikit-learn>=1.1.0->librosa) (3
Requirement already satisfied: cffi>=1.0 in /usr/local/lib/python3.11/dist-packages (from soundfile>=0.12.1->librosa) (1.17.1)
Requirement already satisfied: pycparser in /usr/local/lib/python3.11/dist-packages (from cffi>=1.0->soundfile>=0.12.1->librosa) (2.2
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->pooch>=1.1
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->pooch>=1.1->librosa) (
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->pooch>=1.1->lib
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.11/dist-packages (from requests>=2.19.0->pooch>=1.1->libr
```



```
Requirement already satisfied: pillow>=8 in /usr/local/lib/python3.11/dist-packages (from matplotlib
Requirement already satisfied: pyparsing>=2.3.1 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (3.2.3)
Requirement already satisfied: python-dateutil>=2.7 in /usr/local/lib/python3.11/dist-packages (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-packages (from python-dateutil>=2.7->matplotlib) (1.17.0)
```