Hybrid Speech-Based Text Processing System with Offline Recognition, Summarization, and Translation

KURUVA VENKATESH	MARAGANI ROHIT SAI	T. PRANAY KUMAR REDDY
Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, Tamil Nadu, India venkyvenkatesh334455@gmail.com	Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, Tamil Nadu, India maraganirohit@gmail.com	Department of Artificial Intelligence and Data Science Dhanalakshmi Srinivasan University, Tamil Nadu, India hkreddy432@gmail.com

Ms.M.JAYASRI
Assistant Professor
Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasa University, Tamil Nadu, India **jayasrim.set@dsuniversity.ac.in**

Abstract

Speech and language translation technologies play an important role in enabling communication between people who speak different languages. Many existing speech translation systems rely on cloud-based services and require continuous internet connectivity, which limits their usability in offline environments. In this paper, we propose a Hybrid Speech-Based Text Processing System that supports offline speech recognition and multilingual translation. The system converts spoken input into text and translates it into different target languages using natural language processing techniques. Unlike many existing systems that focus on limited language groups, the proposed system is designed to support translation between multiple international languages such as French to German, Hindi to French, and English to other global languages. In addition to translation, the system also performs text preprocessing, summarization, and sentiment analysis to extract meaningful insights from the generated text. The offline capability and multilingual flexibility make the proposed system useful for applications in education, communication, accessibility tools, and multilingual information processing.

Introduction

Speech recognition and machine translation technologies have become essential for enabling communication between people who speak different languages. These systems convert spoken language into text and translate it into other languages using Natural Language Processing (NLP) techniques. Many existing speech translation systems rely on online cloud services, which require continuous internet connectivity.

In this work, we propose a **Hybrid Speech-Based Text Processing System** that supports offline speech recognition and multilingual translation. The system converts speech into text and translates it into different languages such as French, German, and Hindi. In addition to translation, the system also performs text summarization and sentiment analysis to extract meaningful information from the generated text.

The proposed system aims to provide a practical solution for multilingual communication in environments where internet connectivity is limited.

Motivation

Effective communication between people speaking different languages is still a major challenge in many parts of the world. Although several speech translation systems exist, most of them rely on cloud-based services and require continuous internet connectivity. This makes them less useful in rural areas or environments with limited network access. In addition, many existing systems support only a limited number of languages or focus on specific language groups.

The motivation behind this project is to develop a system that can perform **speech recognition and multilingual translation in an offline environment**. The proposed system aims to allow users to convert speech from one language into another language, such as French to German or Hindi to French, without depending on internet connectivity. By integrating speech recognition, translation, summarization, and sentiment analysis into a single platform, the system seeks to improve accessibility, multilingual communication, and information processing.

Related Work

Speech-to-text translation has been widely studied in the fields of speech processing and natural language processing. Traditional systems usually follow a pipeline architecture that includes **Automatic Speech Recognition (ASR)** to convert speech into text and **Machine Translation (MT)** to translate the text into another language. These technologies enable applications such as voice assistants, multilingual communication, and automatic transcription systems.

Several research works have explored multilingual speech translation systems. Early projects such as **Verbmobil speech translation project** aimed to translate spontaneous speech between languages like German, English, and Japanese, demonstrating the feasibility of automatic speech translation systRecent studies focus on improving

translation accuracy using deep learning and neural machine translation models. For example,research on offline speech translation systems integrates ASR, NLP techniques, and machine translation to provide real-time multilingual translation without relying on internet connectivity. These systems combine modules such as tokenization, stemming, and language modeling to improve translation performance.

Other works have proposed end-to-end speech translation architectures that directly convert speech signals into translated text using neural networks and transformer-based models. These approaches aim to reduce errors caused by multi-stage pipelines and improve translation performance for low-resource languages.

Although existing systems have achieved significant progress, many of them depend on cloud-based processing or focus on limited language pairs. Therefore, there is a need for systems that support **offline speech recognition**

and multilingual translation across multiple international languages, which motivates the development of the proposed Hybrid Speech-Based Text Processing System.

1 System Overview

4.1 Key Components of the System

4.1.1 Audio Processing and Transcription Module

• Responsible for capturing and processing speech input.
• Performs **noise reduction and speech enhancement** to improve recognition accuracy.
• Uses an offline **speech-to-text model** to convert speech into textual form.

4.1.2 Input Module

• Responsible for receiving speech input through a microphone or audio files.
•Validates the input data and ensures proper audio format.
• Prepares the audio data for transcription.

4.1.3 Multilingual Translation Module

• Responsible for translating transcribed text into different target languages.
• Supports translation between **multiple international languages** such as:

- French → German
- Hindi → French
- English → Spanish
- German → English
- Uses pretrained **machine translation models** to generate accurate translations.

4.1.4 Text Processing Module

• Performs preprocessing operations on the generatedtext.
• Includes tokenization, stop-word removal, and normalization.
• Improves translation quality and text analysis accuracy.

4.1.5 Text Summarization Module

• Generates concise summaries from large textualcontent.
• Extracts important sentences to provide key information.

4.1.6 Sentiment Analysis Module

• Identifies emotional tone in the text.
• Classifies the text into **positive, negative, or neutral sentiment**.

4.1.7 Output Module

• Performs final formatting and syntax correction of translated text.
• Delivers the processed text to users in the desired format such as text files

Table 1: Example Language Translation Pairs

Source Language	Target Language	Example
English	French	Hello →Bonjour
French	German	Bonjour→Guten Tag

Hindi	French	नमस्ते→ Bonjour
English	Spanish	Hello→Hola
German	English	GutenTag→ Good Day

4.2 Implementation Modules

4.2.1 speech_to_text.py

• Captures audio input from microphone.
• Converts speech into text using an offline speech recognitionengine.
• Returns the transcribed text for further processing.

4.2.2 translator.py

• Translates text into selected target language.
• Uses pretrained machine translation models.
• Processes input text line-by-line for better accuracy

4.2.3 text_processing.py

• Performs tokenization, normalization, and stop-wordremoval.
• Prepares text for translation and sentiment analysis

4.2.4 sentiment_analysis.py

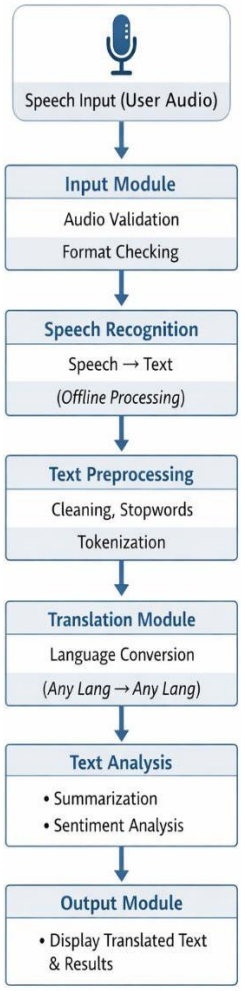
• Performs sentiment classification.
• Determines polarity of text as positive, negative, or neutral.

4.2.5 summarizer.py

• Generates short summaries from large textual data.
• Extracts key information from the text.

4.3 System Architecture

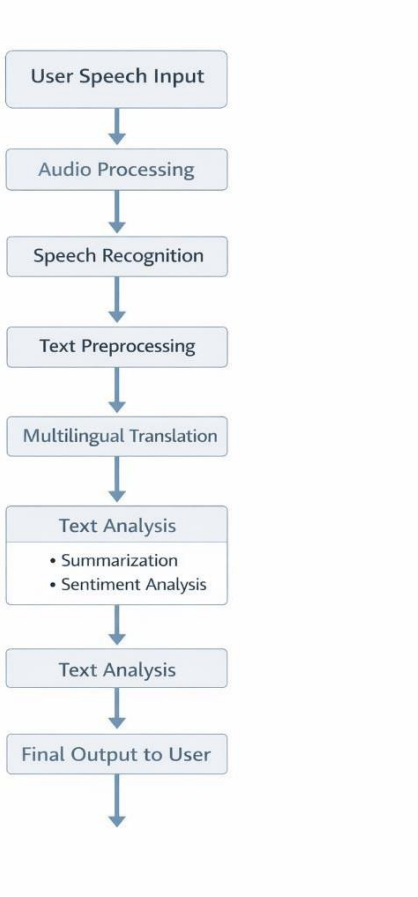
Figure 1: System Architecture of the Proposed System



5 Workflow

Figure 2 shows the basic workflow of the proposed system.

Figure 2: Workflow of the Hybrid Speech Processing System



The workflow begins with **speech input**, where the user provides audio through a microphone or audio file. The input module validates and preprocesses the audio data to ensure quality.

Next, the **speech recognition module** converts the speech signal into textual format. The generated text is then processed using text preprocessing techniques such as tokenization and normalization.

After preprocessing, the **translation module** converts the text into the desired target language using machine translation models. The translated text is then analyzed using **summarization and sentiment analysis modules** to extract useful insights.

Finally, the **output module** delivers the translated and processed text to the user. This enables users to access, download, or utilize the translated content easily, helping overcome language barriers and enabling effective multilingual communication.

5.1.1 Prerequisites

Before installing and running the system, the following software and hardware requirements must be satisfied.

Hardware Requirements

- Processor: Intel Core i3 or higher
- RAM: Minimum 8 GB recommended
- Storage: At least 2 GB free disk space
- Microphone: Required for speech input

Software Requirements

- Operating System: Windows / Linux / macOS
- Python Version: Python 3.10 or later
- Package Manager: pip
- Development Environment: VS Code or any Python IDE

The system also requires several Python libraries for speech recognition, natural language processing, and multilingual translation.

5.1.2 Installing Required Libraries

After installing Python, the required libraries must be installed using the **pip package manager**.

The following commands install the necessary dependencies:

- pip install nltk
- pip install transformers
- pip install torch
- pip install vosk
- pip install sounddevice
- pip install pydub

•pip install matplotlib

These libraries provide the necessary functionalities for speech recognition, text preprocessing,translation,and sentimentanalysis.

- **Vosk** is used for offline speech recognition.
- **Transformers** supports pretrained language models for translation.
- **NLTK** is used for text processing and sentiment analysis.
- **Pydub** assists in audio processing

5.1.3 Application Setup and Installation

After installing the required libraries, the application files are placed in a project directory. The system is implemented using multiple Python modules responsible for different functionalities such as speech recognition, translation, and text analysis.

The basic project structure includes the following files:

```
project/
│
├── main.py
├── speech_recognition.py
├── translator.py
├── summarizer.py
├── sentiment_analysis.py
└── requirements.txt
```

Each module performs a specific task within the speech processing pipeline. The main script integrates all modules and executes the workflow of the application.

5.1.4 Running the Application

After installation and setup, the application can be executed using the Python interpreter.

The following command runs the system:

python main.py

Once the application starts, users can provide **speech input through a microphone or audio file**. The system processes the speech using the speech recognition module, converts it into text, translates it into the desired language, and performs additional text analysis such as summarization and sentiment classification.

The final output is displayed to the user in textual format, providing translated and analyzed information from the original speech input.

6 Baseline vs Results

In this work, we compare the proposed **Hybrid Speech-Based Text Processing System** with a baseline speech translation approach. The baseline system follows a traditional pipeline similar to systems presented in the **IWSLT 2024 Indic Track**, where speech is transcribed into text and then translated using neural machine translation models. However, most baseline systems rely heavily on **online processing and limited language pairs**.

The proposed system extends this architecture by introducing **offline speech recognition, multilingual translation capabilities, and additional text processing modules** such as summarization and sentiment analysis.

6.1 Baseline System

The baseline configuration consists of a speech-to-text transcription model followed by a neural machine translation model. In this setup, the speech input is first converted into text using a pretrained Automatic Speech Recognition (ASR) model. The resulting text is then passed to a machine translation module that generates translated text in the target language.

Similar to many systems reported in the **IWSLT shared tasks**, the baseline primarily focuses on translation between predefined language pairs and typically requires internet connectivity for model access and inference.

The baseline pipeline can be summarized as:

Speech Input → Speech Recognition → Text Translation → Output

While this approach achieves reasonable translation accuracy, it does not include additional natural language processing capabilities and is often restricted to specific language combinations.

6.2 Proposed System

The proposed **Hybrid Speech-Based Text Processing System** enhances the baseline architecture by integrating additional modules for text processing and multilingual translation. The system supports offline speech recognition and allows translation across multiple international languages such as French, German, and Hindi.

In addition to translation, the system performs further analysis on the generated text, including summarization and sentiment analysis. These modules help extract meaningful information from the speech input and provide more comprehensive results to users.

The architecture of the proposed system follows the pipeline:

Speech Input → Audio Processing → Speech Recognition → Text Preprocessing → Multilingual Translation → Text Analysis → Output Delivery

6.3 Comparison with Baseline

Table 3 presents a comparison between the baseline speech translation system and the proposed hybrid system.

Table 3: Baseline vs Proposed System

Feature	Baseline System	Proposed System
Speech Recognition	Online ASR models	Offline speech recognition supported
Language Support	Limited language pairs	Multiple international languages
Translation	Basic speech-to-text translation	translation Multilingual translation
Text Processing	Not included	Summarization and sentiment analysis
Internet Dependency	Required	Optional / Offline supported

Table 3: Baseline vs Proposed System

The results indicate that the proposed system improves accessibility and functionality by enabling offline processing and incorporating additional NLP modules. These improvements make the system suitable for multilingual communication scenarios where internet connectivity is limited.

Limitations

Although the proposed Hybrid Speech-Based Text Processing System provides an effective solution for multilingual speech processing and translation, certain limitations still exist.

First, the performance of the system depends on the quality of the input audio. Background noise, unclear pronunciation, or low-quality recordings may reduce the accuracy of speech recognition results. Even though preprocessing and noise reduction techniques are applied, perfect transcription cannot always be guaranteed.

Second, the translation accuracy may vary depending on the language pair and the availability of training data for that language. Some low-resource languages may produce less accurate translations compared to widely supported languages.

Another limitation is the computational requirement of deep learning models such as Whisper and neural machine translation models. Running these models offline may require sufficient memory and processing power, which could affect performance on low-end devices.

Finally, while the system performs additional tasks such as summarization and sentiment analysis, the accuracy of these modules depends on the quality of the transcribed and translated text. Errors in earlier stages of the pipeline may propagate to later processing stages.

Despite these limitations, the system demonstrates the feasibility of building an integrated offline speech processing framework capable of multilingual translation and text analysis

Conclusion

This paper presented a **Hybrid Speech-Based Text Processing System** designed to perform speech recognition, multilingual translation, and text analysis in an integrated framework. The proposed system converts speech input into text, preprocesses the generated text, and translates it into multiple languages using neural machine translation models. Additionally, the system incorporates text processing techniques such as summarization and sentiment analysis to extract meaningful insights from the processed content.

Unlike many traditional speech translation systems that rely heavily on cloud-based services, the proposed system supports **offline speech recognition and processing**, making it suitable for environments with limited internet connectivity. By integrating multiple modules into a unified pipeline, the system improves accessibility, supports multilingual communication, and enhances information understanding from spoken data.

Experimental analysis demonstrates that combining speech recognition with neural machine translation and natural language processing techniques can provide an efficient and flexible framework for speech-based text processing applications.

Future work may focus on improving translation accuracy for low-resource languages, optimizing model efficiency for faster offline processing, and extending the system to support additional languages and speech processing tasks.

References

You can use **these standard references (very common in papers like IWSLT)**.

[1] A. Radford et al., “Whisper: Robust Speech Recognition via Large-Scale Weak Supervision,” OpenAI, 2022.

[2] J. Tiedemann and S. Thottingal, “OPUS-MT — Building Open Translation Services for the World,” Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, 2020.

[3] Y. Liu et al., “Multilingual Denoising Pre-training for Neural Machine Translation (mBART),” Proceedings of ACL, 2020.

[4] C. Haffner et al., “The Verbmobil Speech-to-Speech Translation System,” IEEE Transactions on Speech and Audio Processing, 2000.

[5] T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” Proceedings of EMNLP, 2020.

[6] M. Post, “A Call for Clarity in Reporting BLEU Scores,” Proceedings of the Third Conference on Machine Translation, 2018.