Mini Project Synopsis
On
“TransSummarize: Leveraging NLP for Transcription and Summarization”

SMIT SIKKIM MANIPAL UNIVERSITY

INSPIRED BY LIFE

SIKKIM MANIPAL INSTITUTE OF TECHNOLOGY

SUBMITTED BY

Hari Om (Reg No. 202000004)

Sushma Oinam (Reg No. 202000048)

Aryan Raj Pradhan (reg No. 202000012)

Department of Information Technology

SESSION: 2023 - 2024

Under the guidance of:

Dr. Saumya Das

Department Of Information Technology, Assistant Professor

DEPARTMENT OF INFORMATION TECHNOLOGY

SIKKIM MANIPAL INSTITUTE OF TECHNOLOGY

Majitar, Rangpo, East Sikkim-737136

Contents

1.Aim of the Project

The project for TransSummarize aims to help in taking information from a conversation or an original audio data.

It will reduce the time and effort of manual documentation.

Sometimes, the main points from a communication in the same language tends to be missed due to accent according to the person.

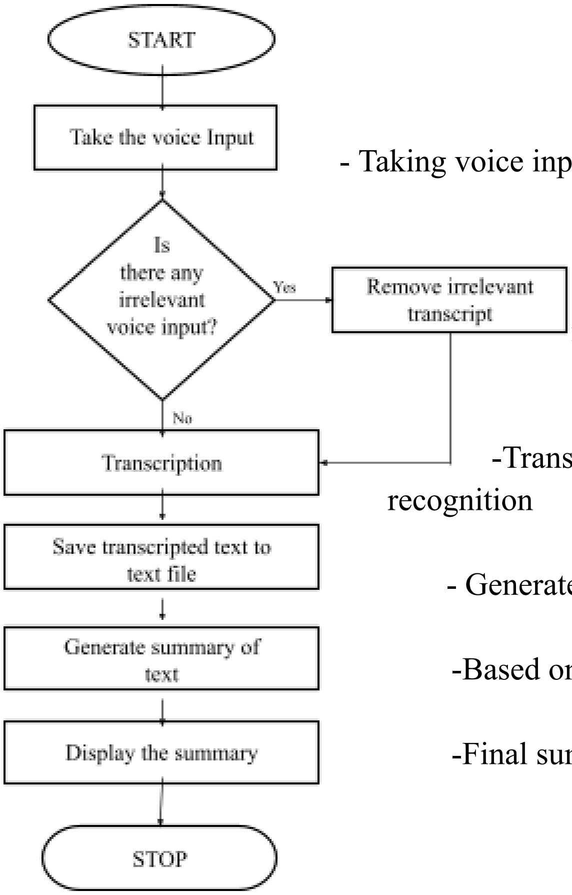
This project will combine speech recognition and text summarization features.

In small scenarios like class, this feature will be helpful as it can provide summarizing lectures into a document without losing vital information.

2.Problem definition

Currently, there are many speech-to-text tools available, but most of them simply transcribe the speech into text without providing any summary or analysis. This can be time-consuming and inefficient, as the user has to manually read through the entire text to understand the main points.

3.Solution Strategy


START
Take the voice Input
Is there any irrelevant voice input?
Yes
Remove irrelevant transcript
No
Transcription
Save transcripted text to text file
Generate summary of text
Display the summary
STOP

- Taking voice input from microphone using Pyaudio

- If irrelevant information generated, it will be removed

-Transcription is done by Speech to Text using Speech-recognition

- Generated transcription is saved in text file using python file-handling

-Based on a trained model, Summary is generated.

-Final summary is displayed.

Figure 1: Flowchart Of Solution Strategy

4.Workflow Diagram

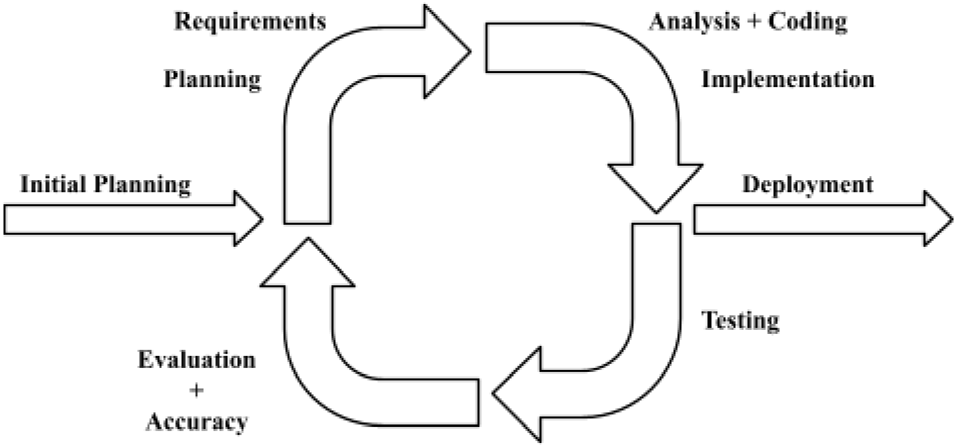


Figure 2: Iterative Model

5. Work Progress

Text Summarization
Summarization is a process of going through a corpus and discovering and identifying the important topics, points, or keywords in it and generating a brief paragraph or text that allows us to get the gist of the corpus.

- Short, fluent and accurate summary

Two main Summarization techniques:

1. Extractive Text Summarization
   Extracts original sentences from the corpus directly if they are deemed important or ranked so. i.e subset of the corpus.

2. Abstractive Text Summarization
   Attempts to discover the important topics, points, or keywords and tries to understand their context and generate the summary intelligently. More difficult to implement than extractive text summarization.

1. Extractive Text summary
   Algorithms used: TextRank, nlargest
   - Text Rank
     - TextRank is based on the PageRank algorithm used on Google Search Engine.
     - It is a graph-based ranking model.
     - It prefers pages with a higher number of pages hitting it.
     - Originally "TextRank" algorithm used the percentage of words appearing among two sentences as the weights of an edge between them.
     - The algorithm then creates a graph with sentences as the nodes and overlapped words as the links.

   - nlargest
     - nlargest is a function that finds n number of elements that is the largest or have the largest value for a given key.
     - it can be used to find the rank of sentences in the corpus for extractive text summarization.

2. Abstractive text summary
   Algorithms: T5 (Text-To-Text Transfer Transformer),Seq2Seq

   - Seq2Seq Model

   Seq2Seq model is a deep learning model that takes in a sequence of inputs and gives a new sequence of outputs.
   Takes into account the context words surrounding the input.

   The basic idea behind Seq2Seq is to use a neural network to encode the input sequence into a fixed-length vector and then use another NN to decode the vector into the output sequence.

   Two main components:
   1. Encoder - That creates hidden layers from input words.
   2. Decoder - That takes a hidden vector from the encoder, and the current word to produce the next hidden vector and predict the upcoming word.

   Characteristic:
   - Sequences as a corpus
   - Word Embedding mechanism
   - Encoder-Decoder
   - Different model for training and testing

6. Literature Review

Author Name , Journal Name, Vol., Year	**Title of the Paper**	**Inference**	**Research Gap**	**Relevance with the present work**
Shaikh Naziya S, R.R.Deshmukh	Speech Recognition System - A Review	● Techniques in Speech Recognition Systems(SRS)	● Problems about noise, echoes,	● Vocabulary of HMM is very high

IOSR Journal of Computer Engineering(IOSR-JCE), Volume 18 (July- Aug, 2016) [1]		● Various SRS modeling techniques are listed. ● The different models have their pros and cons.	background noise and how to counter them are not mentioned.	- can be used for large amounts of data for speech recognition. ● Machine learning techniques for speech recognition.
Nenkova A., & McKeown, K. Mining text data, 43-76.(2012) [2]	A survey of text summarization techniques	● Topic Representation extracts topics discussed in the input document. ● Indicator representation scores the importance of each sentence which will come in the summary.	● Generation of summary for multiple documents ● Complexity of human language	● Provides ideas about how text summarization works.

Kågebäck, M., Mogren, O., Tahmasebi, N., & Dubhashi, D. . Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) (pp. 31-39). (2014, April)	Extractive summarization using continuous vector space models	● evaluate different compositions for sentence representation on a standard dataset using the ROUGE evaluation measures.	● It returns the exact phrase from the information provided just in shortened form.	● the effects of using phrase embeddings for summarization, and showed that these can significantly improve the performance of the state-of-the-art summarization method
Gupta, S., & Gupta, S. K. Expert Systems with Applications, 121, 49-65. (2019).	Abstractive summarization: An overview of the state of the art.	● Its approaches are broadly divided into structure based and semantic based. ● sentence compression, concept fusion, calculation of path scores and summary generation are few common parts of an abstractive summarization system	● tend to generate false information ● Need of quantitative measures	● abstractive summarization systems consist of 3 steps namely pre-processing, inferencing and Natural Language Generation.

7. Requirements

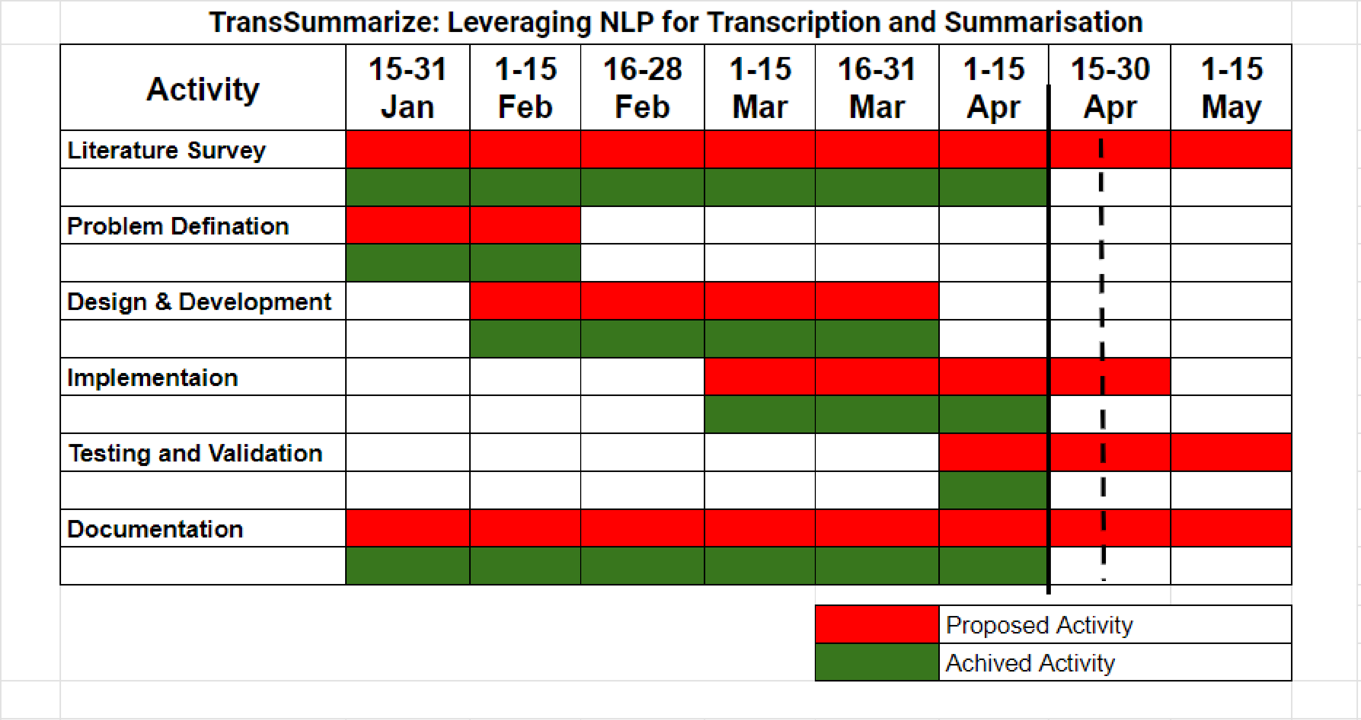
Hardware :

- RAM : 8 GB
- Hard Disk : 750 GB SSD
- Processor : Intel i5
- GPU : Nvidia RTX 3050 ti

Software :

- Programming Language : Python 3.10
- Packages : nltk, T5Tokenizer, pyaudio, SpeechRecognition, spacy, pytorch, etc
- Operating System : Windows 11

8. Gantt Chart



TransSummarize: Leveraging NLP for Transcription and Summarisation

Activity	15-31 Jan	1-15 Feb	16-28 Feb	1-15 Mar	16-31 Mar	1-15 Apr	15-30 Apr	1-15 May
Literature Survey	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	🟩	🟩	🟩	🟩	🟩	🟩		
Problem Defination	🟥	🟥						
	🟩	🟩						
Design & Development		🟥	🟥	🟥	🟥			
		🟩	🟩	🟩	🟩			
Implementaion				🟥	🟥	🟥	🟥	
				🟩	🟩	🟩		
Testing and Validation						🟥	🟥	🟥
						🟩		
Documentation	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
	🟩	🟩	🟩	🟩	🟩	🟩		

🟥	Proposed Activity
🟩	Achived Activity

9. References

[1] Shaikh Naziya, S., & Deshmukh, R. R. (2016). Speech recognition system—a review. *IOSR J. Comput. Eng*, *18*(4), 3-8.

[2] Nenkova, A., & McKeown, K. (2012). A survey of text summarization techniques. *Mining text data*, 43-76.

[3]Kågebäck, M., Mogren, O., Tahmasebi, N., & Dubhashi, D. (2014, April). Extractive summarization using continuous vector space models. In Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC) (pp. 31-39).

[4]Gupta, S., & Gupta, S. K. (2019). Abstractive summarization: An overview of the state of the art. Expert Systems with Applications, 121, 49-65.