A Project Report

on

MULTI-LINGUAL SPEECH TO SPEECH TRANSLATION

ARTIFICIAL INTELLIGENCE FOR DATA SCIENCE

by

Navadeep Reddy (2010030313)

Siddharth (2010030475)

Vipul Reddy (2010030502)

Manoj Peravali (2010030503)

under the supervision of

Dr. Arpita Gupta

Assistant Professor

Department of Computer Science and Engineering

K L University Hyderabad,

Aziz Nagar, Moinabad Road, Hyderabad – 500075, Telangana, India.

April, 2022

DECLARATION

The Project Report entitled "**MULTI-LINGUAL SPEECH TO SPEECH TRANSLATION**" is a record of bonafide work of **NAVADEEP REDDY (2010030313), SIDDHARTH (2010030475), VIPUL REDDY (2010030502), MANOJ PERAVALI (2010030503)** submitted in partial fulfillment for the award of B.Tech in the Department of Computer Science and Engineering to the K L University, Hyderabad. The results embodied in this report have not been copied from any other Departments/University/Institute.

NAVADEEP REDDY
(2010030313)

VIPUL REDDY
(2010030502)

MANOJ PERAVALI
(2010030503)

SIDDARTHA SRI SAI

(2010030475)

CERTIFICATE

This is to certify that the Project Report entitled "**MULTI-LINGUAL SPEECH TO SPEECH TRANSLATION"** is being submitted by **Navadeep Reddy (2010030313), Siddharth (2010030475), Vipul Reddy (2010030502), Manoj Peravali (2010030503)** submitted in partial fulfillment for the award of B.Techin CSE to the K L University, Hyderabad is a record of bonafide work carried out under our guidance and supervision.

The results embodied in this report have not been copied from any other departments/ University/institutes.

Signature of the Supervisor

Dr. Arpita Gupta

Assistant Professor

Signature of the HOD

Signature of the External Examiner

ACKNOWLEDGEMENT

First and foremost, we thank the lord almighty for all his grace & mercy showered upon us, for completing this project successfully.

We take a grateful opportunity to thank our beloved Founder and Chairman who has given constant encouragement during our course and motivated us to do this project. We are grateful to our Principal **Dr. L. Koteswara Rao** who has been constantly bearing the torch for all the curricular activities undertaken by us.

We pay our grateful acknowledgment & sincere thanks to our Head of the Department **Dr. Chiranjeevi Manike** for her exemplary guidance, monitoring, and constant encouragement throughout the course of the project. We thank **Dr. Arpita Gupta** of our department who has supported us throughout this project holding the position of supervisor.

We wholeheartedly thank all the teaching and non-teaching staff of our department without whom we won't have made this project a reality. We would like to extend our sincere thanks, especially to our parents, our family members, and our friends who have supported us to make this project a grand success.

Abstract

In today's world language translation is very important, because if any person attends a global meeting/conference the language might be different from what is known, at that time speech translation is very useful. Speech translator is mediator between two languages. In this we have reviews various issues related to Speech translation as well as various difficulties in it. The purpose of this project is to develop a speech translator that can recognize initial language spoken and can translate from one language to other languages sentence by sentence.

I believe that these kinds of developments are making our lives simpler. Area where real-time speech translation can be helpful is in online lectures, while listening to the instructor we can translate to the language we are comfortable with.
This feature will make the learning process easier. This is just a basic example; I am sure there are many crucial areas where real-time speech translation can be implemented.

Introduction

Speech-to-Speech Translation aims at translating a source speech signal into a target speech signal.

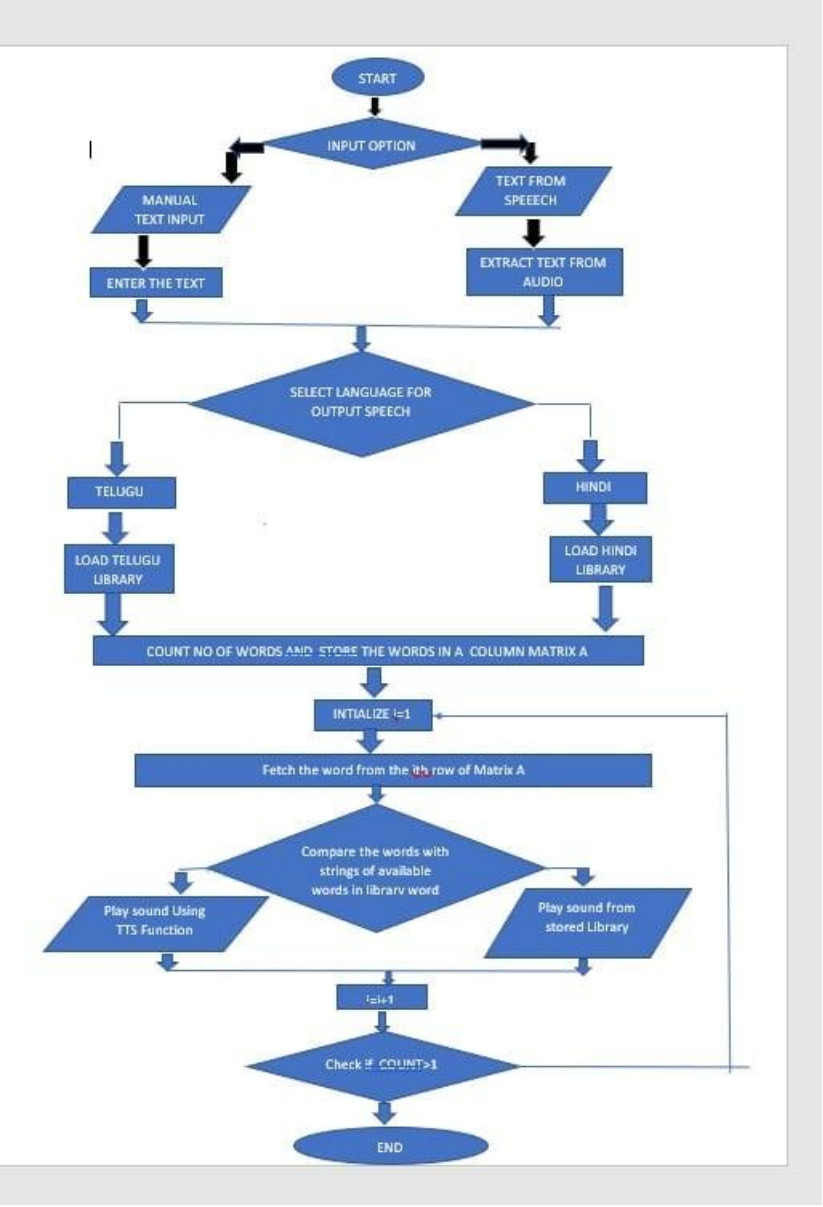
Word translation
The first translation systems identify one-to-one associations between words of target and source languages.

Phrase translation
The human translation is a very complex process which is not only word-based.

Language model
A language model has an important role in a statistical machine translation.

Decoding
The translation issue is treated as an optimization problem. Translating a sentence from English into a foreign language involves finding the best foreign target sentence

Flowchart


START
INPUT OPTION
MANUAL TEXT INPUT
TEXT FROM SPEEECH
ENTER THE TEXT
EXTRACT TEXT FROM AUDIO
SELECT LANGUAGE FOR OUTPUT SPEECH
TELUGU
HINDI
LOAD TELUGU LIBRARY
LOAD HINDI LIBRARY
COUNT NO OF WORDS AND STORE THE WORDS IN A COLUMN MATRIX A
INTIALIZE i=1
Fetch the word from the ith row of Matrix A
Compare the words with strings of available words in librarv word
Play sound Using TTS Function
Play sound from stored Library
i=i+1
Check if COUNT>i
END

Literature review

- The Voice/speech translation system integrates two technologies: Automatic Speech Recognition, Machine Translation. The speaker of language A speaks, and the speech recognizer recognizes the utterance. The input is then converted into a string of words, using dictionary and grammar of language A, by using the massive corpus of text of language A.
- HMMs are mostly used in speaker recognition today. We get the output sequence of symbols from these models. HMMs are used in speech recognition because the audio signal can be considered piece wise stationary signal.
- Neural Networks has come up as nice approach for acoustic modelling in ASR since 1980s. In contract to HMMs, neural networks do not make any assumptions regarding the statistical properties and have several qualities that make them great model for speech recognition.
- Most of the data that we find is from different domains. For example, the texts that are used in the chat rooms are different than those used in the parliaments. The problem in the Neural machine translation can be that it is trained on the data that is not at relevant to the user and hence not getting correct translations. This problem is called as Domain Adaptation. Speech translation is conventionally carried out by cascading an Automatic Speech Recognition System and Machine Translation system. Generally, the factors that are optimized are the language models and the acoustic models along with the word error rate for the ASR system and the BLEU score for the MT system

S.NO	Authors	Title	Publishing	Pros	Cons
1	Mattia Antonino Di Gangi1,2,Roldano, Cattoni1,Matteo Negri and Marco Turch	Must-c:a Multilingual speech Translation corpus	07-06-2019	Scarcity of training corpora.	Scalable to add new data and cover new languages.
2	Hirofumi Inaguma 1Shunkiyono 2 kevin Duh 3 Shigeki Karita 4 Nelson Yalta 5 Tomoki Hayashi 6,7 Shinji Watanable	Espnet-ST:All-in-One Speech Translation	30-09-2020	Quick development of speech-to-speech translation systems in a Single framework	Gap between end-to-end and cascaded approaches
3	**Alexandre Berard; Laurent Bessacier;Ali Can Kocabiyikoglu;Olivier Pietquin**	End-to-End Aytomatic Speech Translation	15-04-2018	Source language transcription is not available	Compact and efficient end-to-end speech translation models
4	Jia* Ron J.Weiss*Fadi Biadsy,Wolfgang Macherey,Melvin Johnson,Zhifeng Chen,Yonghuiwu	Direct speech-to-speech translation with a sequence-to-sequence model	25-06-2019	Translation without relyingon anm intermediate text representation	The voice tranfer does not work as well
5	Parnia Bahar 1,2,Albert Zeyer1,2,Ralf schulter 1 and Hermann Ney	SpecAugment for End-to-End Speech Translation	2005	Low-cost implementation	Effectiveness of the approach.

Software and hardware requirements

OS: Windows XP/7/8/10/11

RAM:4/8 GB

Processor: Intel i7-i9/AMD R3-9

System-type: 64-bit OS

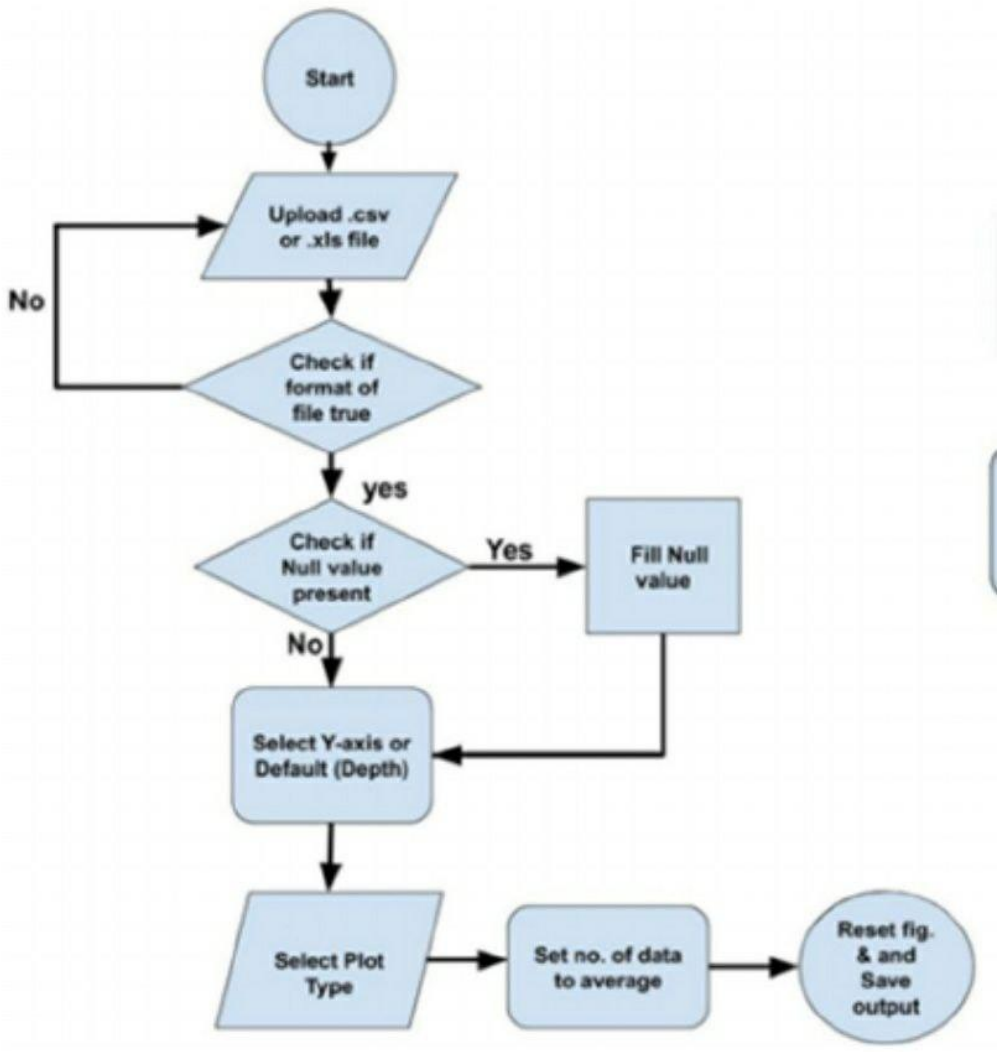
TOOLS: PyCharm

METHODOLOGY & ALGORITHMS

Tkinter – python GUI programming tool

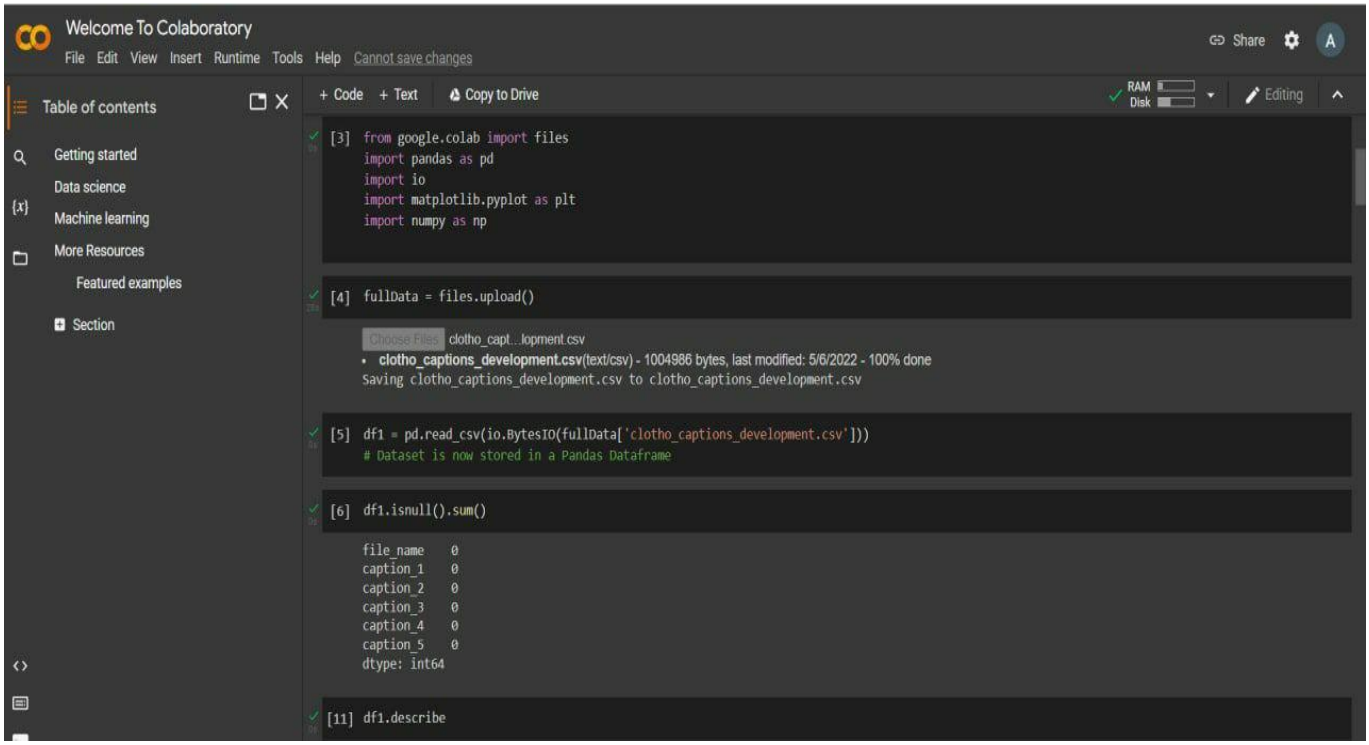
Tkinter is a library written in Python that is widely used to create GUI applications. It is very easy to build GUI using Tkinter and the process is even faster and has several widgets that can be used while developing GUI. These include buttons, radio buttons, checkboxes, etc.

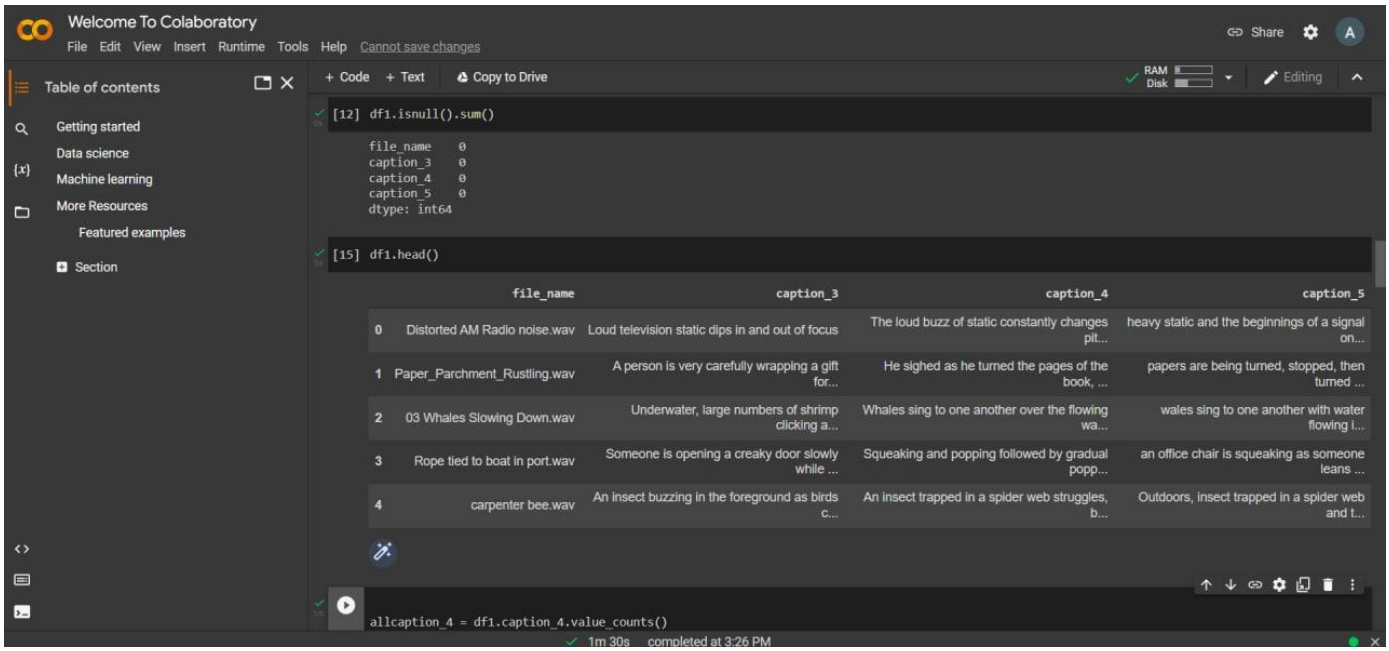
Here we will be using the CVSS Data set that is available on Kaggle.



DATASET

Datasets	Characteristics	Characteristics and models
CVSS	**CVSS** is a multilingual-to-English speech to speech translation corpus, covering sentence-level parallel S2ST pairs from 21 languages into English. CVSS is derived from the Common Voice speech corpus and the CoVoST 2 speech-to-text translation (ST) corpus, by synthesizing the translation text from CoVoST 2 into speech.	i) PnG NAT ii) PnG NAT with VC iii) Speaker Encoder
MuST-C	**MuST-C** currently represents the largest publicly available multilingual corpus (one-to-many) for speech translation. It covers eight language directions(English to German, Spanish, French, Italian, Dutch, Portuguese, Romanian and Russian). The corpus consists of audio, transcriptions and translations of English TED talks, and it comes with a predefined training, validation and test split.	i)Bilingual vs. Multilingual ii)Low-resource scenario Fine tuning
MaSS	MaSS (Multilingual corpus of Sentence-aligned Spoken utterances) is an extension of the CMU Wilderness Multilingual Speech Dataset, a speech dataset based on recorded readings. MaSS extends it by providing a large and clean dataset of 8,130 parallel spoken utterances across 8 languages. (The covered languages are: Basque, English, Finnish, French, Hungarian, Romanian, Russian and Spanish)	i)LAS Network Architectures ii)Learning Rate Schedules iii)Shallow Fusion with Language Models


Welcome To Colaboratory
File Edit View Insert Runtime Tools Help Cannot save changes
Share
Table of contents
+ Code + Text
Copy to Drive
RAM
Disk
Editing
Getting started
Data science
Machine learning
More Resources
Featured examples
Section
[3] from google.colab import files
import pandas as pd
import io
import matplotlib.pyplot as plt
import numpy as np
[4] fullData = files.upload()
Choose Files clotho_capt...lopment.csv
clotho_captions_development.csv(text/csv) - 1004986 bytes, last modified: 5/6/2022 - 100% done
Saving clotho_captions_development.csv to clotho_captions_development.csv
[5] df1 = pd.read_csv(io.BytesIO(fullData['clotho_captions_development.csv']))
Dataset is now stored in a Pandas Dataframe
[6] df1.isnull().sum()
file_name 0
caption_1 0
caption_2 0
caption_3 0
caption_4 0
caption_5 0
dtype: int64
[11] df1.describe


Welcome To Colaboratory
File Edit View Insert Runtime Tools Help Cannot save changes
Share
Table of contents
+ Code + Text
Copy to Drive
RAM
Disk
Editing
Getting started
Data science
Machine learning
More Resources
Featured examples
Section
[12] df1.isnull().sum()
file_name 0
caption_3 0
caption_4 0
caption_5 0
dtype: int64
[15] df1.head()
file_name caption_3 caption_4 caption_5
0 Distorted AM Radio noise.wav Loud television static dips in and out of focus The loud buzz of static constantly changes plt... heavy static and the beginnings of a signal on...
1 Paper_Parchment_Rustling.wav A person is very carefully wrapping a gift for... He sighed as he turned the pages of the book, ... papers are being turned, stopped, then turned ...
2 03 Whales Slowing Down.wav Underwater, large numbers of shrimp clicking a... Whales sing to one another over the flowing wa... wales sing to one another with water flowing i...
3 Rope tied to boat in port.wav Someone is opening a creaky door slowly while ... Squeaking and popping followed by gradual popp... an office chair is squeaking as someone leans ...
4 carpenter bee.wav An insect buzzing in the foreground as birds c... An insect trapped in a spider web struggles, b... Outdoors, insect trapped in a spider web and t...
allcaption_4 = df1.caption_4.value_counts()
1m 30s completed at 3:26 PM

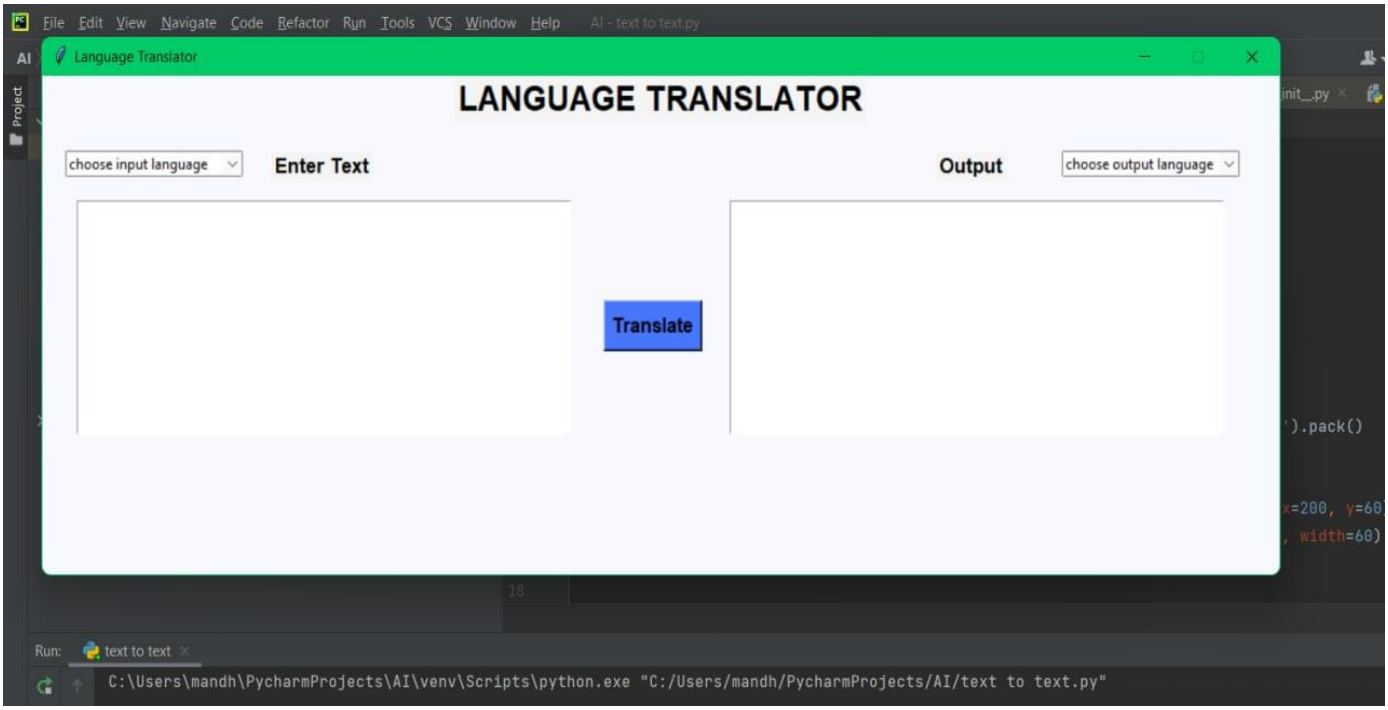
IMPLEMENTATION

```python
from tkinter import *
from tkinter import ttk
from googletrans import Translator, LANGUAGES

root = Tk()
root.geometry('1080x400')
root.resizable(0, 0)
root.title("Language Translator")
root.config(bg='ghost white')

# heading
Label(root, text="LANGUAGE TRANSLATOR", font="arial 20 bold", bg='white smoke').pack()

# INPUT AND OUTPUT TEXT WIDGET
Label(root, text="Enter Text", font='arial 13 bold', bg='white smoke').place(x=200, y=60)
Input_text = Text(root, font='arial 10', height=11, wrap=WORD, padx=5, pady=5, width=60)
Input_text.place(x=30, y=100)
```

```python
Output_text = Text(root, font='arial 10', height=11, wrap=WORD, padx=5, pady=5, width=60)
Output_text.place(x=600, y=100)

##################
language = list(LANGUAGES.values())

src_lang = ttk.Combobox(root, values=language, width=22)
src_lang.place(x=20, y=60)
src_lang.set('choose input language')

dest_lang = ttk.Combobox(root, values=language, width=22)
dest_lang.place(x=890, y=60)
dest_lang.set('choose output language')



#######################################  Define function #######
```

```
def Translate():
    translator = Translator()
    translated = translator.translate(text=Input_text.get(1.0, END), src=src_lang.get(), dest=dest_lang.get())
    Output_text.delete(1.0, END)
    Output_text.insert(END, translated.text)


########## Translate Button ########
trans_btn = Button(root, text='Translate', font='arial 12 bold', pady=5, command=Translate, bg='royal blue1',
                   activebackground='sky blue')
trans_btn.place(x=490, y=180)

root.mainloop()
```

Results


Language Translator
LANGUAGE TRANSLATOR
choose input language
Enter Text
Output
choose output language
Translate
Run:
text to text
C:\Users\mandh\PycharmProjects\AI\venv\Scripts\python.exe "C:/Users/mandh/PycharmProjects/AI/text to text.py"

Conclusion and Future work

Humans can interact with each other through natural language. If both people understood their languages, then interaction between these two people is more complete. In this Speech translation system, there is no need of creation of database manually for matching/converting source text to destination text, due to this translation time will be reduced. From the comparison between techniques in speech recognition, Sphinx model is identified as one of the popular connectionist techniques and suitable to use in speech recognition. We plan to develop a multi-way translation like Hindi to English, Tamil and Telugu. Finally, the Real time voice translation system is done in this way.

At some point in the future, speech recognition may become speech understanding. The statistical models that allow computers to decide what a person just said may someday allow them to grasp the meaning behind the words. Although it is a huge leap in terms of computational power and software sophistication, some researchers argue that speech recognition development offers the most direct line from the computers of today to true artificial intelligence. In our project we would like to completely develop and deploy speech translation feature along with sub-titles directly into live videos and speeches.

REFERENCES:

- A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan, "JANUS-III: Speech-to-speech translation in multiple languages," in Proc. ICASSP, 1997.

- W. Wahlster, Verbmobil: Foundations of speech-to-speech translation. Springer, 2000

- S. Nakamura, K. Markov, H. Nakaiwa, G.-i. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," IEEE Transactions on Audio, Speech, and Language Processing, 2006.

- International Telecommunication Union, "ITU-T F.745: Functional requirements for network-based speech-to-speech translation services," 2016.

- H. Ney, "Speech translation: Coupling of recognition and translation," in Proc. ICASSP, 1999