IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

# Text-to-Speech & Speech-to-Text Conversion

***Shaikh Arbaz Hussain***
*CSE*
*Nalla Narasimha Reddy*
*Group of Institutions*
*Hyderabad, Telangana*

***Shaikh Irshad Baba***
*CSE*
*Nalla Narasimha Reddy*
*Group of Institutions*
*Hyderabad, Telangana*

***Yasa Keerthi***
*CSE*
*Nalla Narasimha Reddy*
*Group of Institutions*
*Hyderabad, Telangana*

***Dr. S. Sree Hari Raju***
*Associate Professor*
*CSE*
*Nalla Narasimha Reddy*
*Group of Institutions*

***Abstract*—Text-to-Speech & Speech-to-Text is the integration of completely two different technologies onto a single platform. These technologies are quite opposite to each other in conversions and deal with highly trained models with Deep Neural Networks (DNNs). Text-to-Speech (TTS) is a technology that helps users to convert any text into speech. The text may be his/her content a document or an email, whatever the text may be TTS can convert it into human-like speech. This technology is used nowadays in Google location tracking Systems, Railway Stations, education, and so on. Corpus-based TTS helps in generating a synthesized speech that resembles human sound. Speech-to-Text (STT) is a technology that deals with speech recognition which converts recognized speech into text. Speech may be human instant voice or streaming content, podcasts, and call recordings. This technology is widespread in today's business world which helps quick notes from words. The converted text is saved as Transcript.txt and speech is saved as .mp3 for user further requirements in his folders. This project and these technologies help humans work of reading, writing, and speaking much more easily and effectively.**

***Keywords—Text-to-Speech, Speech-to-Text, corpus-based TTS, Deep Neural Network, Google location Tracking System, podcasts.***

## I. INTRODUCTION

As we all know text and speech are the communication bridges between humans and also with digital computers. Text-to-Speech synthesizer is a technology that helps computers speak to humans that resembles the human speaker.TTS engines use Deep learning algorithms to analyze, preprocess the text, and synthesize the speech sound. TTS Conversions happen in two stages. First, text is analyzed, preprocessed, and converted into phonemes. Second is the generation of speech sound that takes the help of inbuilt trained datasets. These datasets consists of a huge number of speech sounds for a particular word/sentence with prosody units. Speech-to-Text is the software that has the capability of recognizing the human voice decibels and can convert the recorded voice into text. A computer program draws the auditory signals from human speech and then converts the speech into text which is called Unicode. At last, these characters are combined to form words and words as sentences to generate a transcript. While conversion of speech into text the speech should not have any background noise or inadequate performance these issues may lead to poor quality of transcripts. The advantage is that if the input is text TTS gives speech as output and if the input is speech sound STT gives text as output so the generated text/speech outputs can again be heard as speech or again converted into text as per user requirement. At last, these converted data is securely stored if the user wants it for further changes/ references.

## II. LITERATURE SURVEY

In [1] Dr. S. A. Ubale, Girish Bhosale, Ganesh Nehe, Avinash Hubale, and Avadhoot Walunjkar suggested that the existing system is facing difficulty in processing the text in the images so they have designed an algorithm that is capable of converting the text in images with high background. This feature mainly helps people who have eyesight. In [2] Babu Pandipati and Dr. R. Praveen Sam explained how speech is converted into text by using Deep Neural Network (DNN) methods. Speech is the medium of humans and computers without any sense of touch. The STT conversions are accurate results when dealing with Hidden Markov Models (HMM). The proposed system does the work of sending email with voice as input and at the receiver end the email which is as text is narrated in speech format. Here the sender need not type text and the receiver need not read the text. The total process of email sending & email receiving is done just with STT models. HMM is created by a system for converting speech into text as it acts as an internal database. Convolution Neural Network (CNN) helps in improving the quality of speech recognition software. It is the variant of DNN. In [3] Shivangi Nagdewani and Ashika Jain explained the methods of converting text into speech and speech into text. They suggested that HMM combined with DNN results in the best accuracy of text or speech. The speech conversion feature is supported by Python programming languages which support Pyttsx3 and GTTS for online speech conversions. Also, the conversions may be many languages of user interest. In [4]

Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik and Supriya Agarwal explained in detail about the text and speech conversion systems. They said that Text-to-speech synthesis is performed in 3 phases: Articulator synthesis, Format synthesis, and concatenative synthesis. Speech-to-text conversion uses HMM and Artificial Neural Network (ANN). HMM is best suited for recognition speed and recognition accuracy of quality of sound. ANN algorithm helps in the removal of noisy and unwanted data. They explained Machine Translation (MT) that is a field of Artificial Intelligence. MT uses the MT system to translate one language to another. This is the reason why we need to develop applications that can translate text from one language to another. Acoustic models build a relation between the information and phonetics. This builds a correlation between the text and phonemes. Language models estimate the probability of occurrence of a word after a word. They differentiate the pronunciation of words and phrases that look alike. Therefore all the research papers explained the use of deep learning techniques and, effect of highly trained models to generate the speech or text as output.

## III. METHODOLOGY

The methodology of Text-to-Speech & Speech-to-Text is exactly the opposite with inputs and outputs but both conversions use the same models for conversions.

Text-To-Speech methodology- The TTS engine or software is composed of a front end and a back end. The front end first analyses the raw text from the user which contains the numbers, abbreviations, etc. This process of analyzing the raw data and enhancing the text is called preprocessing or text normalization. Now, these texts are assigned with their phonetic transcriptions. Therefore, assigning phonemes to text is called Text-To-Phoneme or Grapheme-To-Phoneme conversion. The phonemes are then attached with the correct prosody units which generate the symbolic representation of text. This completes the front end. Now the back end generates the speech with the help of a speech synthesizer. The speech sounds which are recorded in the databases are concatenated to generate the synthesized speech. The synthesizer can import a vocal track model into it to generate an accurate human voice as output. The quality of a synthesizer can be said by the quality of sound it gives as output. The system differs in the storage of recorded speech. The huge database gives accurate speech sounds.

Speech-to-Text methodology- The STT engine or the speech-recognizing software composed of two parts. The front end i.e. the first part recognizes the speech and draws the auditory signals from speech. These recorded speeches are assigned with their particular text using characters called Unicode. The back end i.e. the second part generates text by combining the text or phrases from the STT inbuilt databases. The larger the database size the accurate the text transcripts. So to get an accurate and quality transcript by the STT engine we need to ensure that the recorded speech does not have any background noise or inadequate pronunciation and also one person should speak at a single time for quality recording STT uses HMM as internal databases as a dictionary for generating the text.

The conversions of TTS & STT use the TTS & STT engine, language models, Acoustic databases, phonetic databases, prosody databases text preprocessing libraries, and feature extraction process. Acoustic databases are the collection of audio recordings which are used to train the models. The audio recordings are of different linguistics with different accents and pronunciations. Language models are AI based on models that are capable of understanding preprocessing and generating the natural language. The feature extraction process is the crucial step in STT because it recognizes the raw signals of speech and converts the signals in such a way that are further analyzed and preprocessed into text. Phonetic databases consist of phonemes of every word for conversion of text into signals. Prosody databases consist of prosody features. Prosody feature refers to the pitch, duration, intensity, rhythm, annotation, accent, and stress of each word for clarity and quality of speech. Therefore, we can say that for every step in TTS & STT conversion, there is a database connected internally. The generation of speech and text will be dependent on the databases and highly trained models. In deep learning, the models are fed with high training data so that when it comes to testing the new high-level input these models do not lack in information and generate an efficient output. The efficiency of any system depends on the database and the trained models. The pieces of stored speech or text are combined to form the converted speech and text for the user.

## IV. IMPLEMENTATION

To implement this project Text-to-Speech & Speech-to-Text the main step is to integrate the Text-To-Speech module and Speech-To-Text module into a single platform using Python as a source language for the user to have the benefits of both resources on a single program. The process involved in implementation:

1. Importing the Required Libraries and modules
- Import the required libraries such as Tkinter. From Tkinter import the modules such as file dialog, and combo box.
- Import Speech Recognition module, Pyttsx3, and os.

Tkinter (
Tk stands for Toolkit and inter is the interface) is the Graphical User Interface (GUI) that is built in Python. This provides the features for developing a GUI desktop application. Tk provides the tools and different widgets.
The file dialog module in Tkinter allows user to save the files in their folders. The Ttk module provides a combo box widget that allows the user to select a single option from the list of choices. Here in our program, we use a combo box for selecting the voice for generated speech. The voices we are providing are male and female voices. The other combo box allows the user to choose the speed of speech that may be low, high, and normal.
The speech recognition module in Python helps developers to integrate speech engines such as Google Web speech API, and sphinx. In this project, we use Google Web speech API to recognize the voice of the user through the microphone. This only supports online conversion as it deals with Google API
Pyttsx3 is a library in python that converts Text-to-Speech offline. It uses Sapi5 in Windows for the conversion of text. Sapi5 allows 2 voice generations of male & female.
Os module in Python provides a way to interact with the system functions.

2. Loading the VGG16 model

It is the convolution Neural Network architecture that takes the input as visual geometry.

3. Functioning of Text-to-Speech

When the user feels to convert the text into speech he needs to give input which of text format in the textbox. The input may be the users own words or a mail or a document but it should be the text. Now user has to select the speak button to convert the text into speech. The user also has the option of selecting the male or female voice and the speed of speech. After satisfying the user needs user has the option to save the converted speech into the users own file. The file is saved as a .mps file.

4. Functioning of Speech-to-Text

When the user feels to convert his speech into words, the user hast to select the record button in the available GUI interface. Then the program starts recognizing the speech of the user.

After completion of recognition it generates a transcript that transcript can be saved in his folder. If the user wants to convert the generated transcript into speech he has to select the speak button. This process goes on.

5. Working with the download directory

The generated speech and text can be downloaded or saved into a user file in the user own folder. This is provided by the file dialog module in Tkinter in Python.

6. Graphical User Interface

Python provides the Tkinter library for developing GUI applications in Python with a set of tools and widgets. The tools we have used are combo boxes, buttons, text boxes, and so on.

## V. SYSTEM ANALYSIS

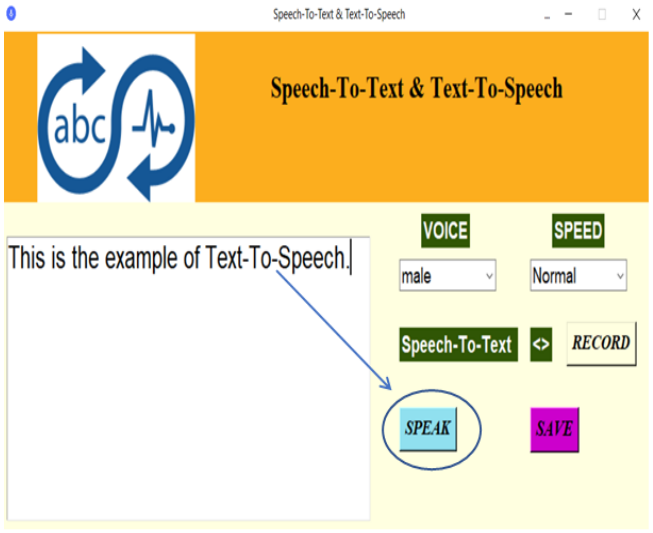


*Fig.5.1.Text-To-Speech Conversion*

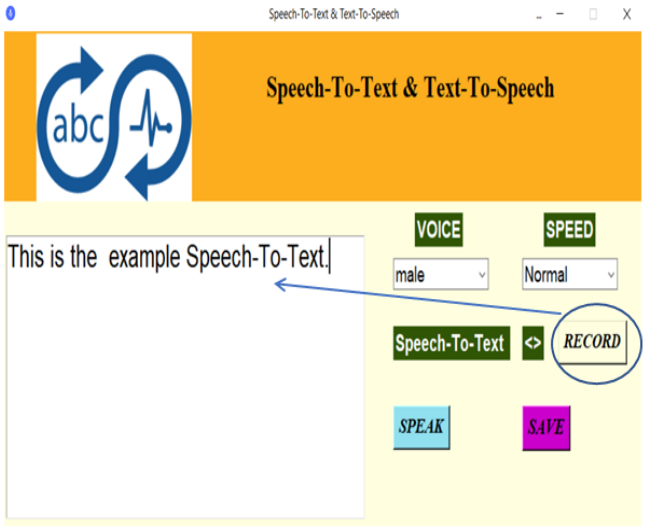


*Fig.5.2.Speech-To-Text Conversion*

- Transcripts are particularly helpful for HR, marketing departments, and event procedures.
- Transcripts can be heard as speech again with the help of TTS simultaneously.

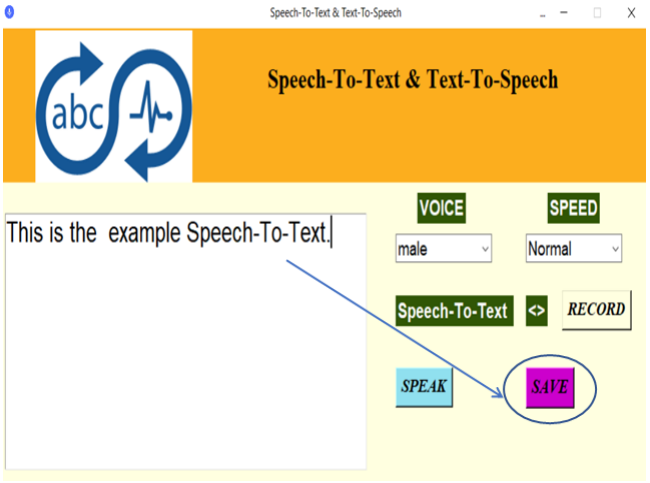


*Fig.5.3.Saving into Folder*

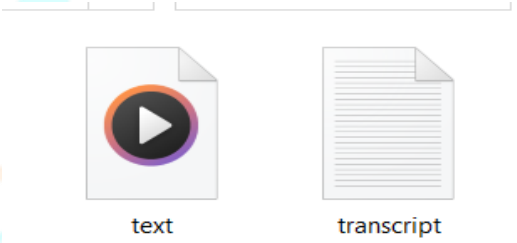


*Fig.5.4.Saved as text.mp3 and transcript.txt*

## VI. ADVANTAGES

- Both TTS & STT are available on a single platform.
- Saves the converted speech or text into the user's folders for further use.
- Speech sounds can be heard in male and female voices with expected speed.
- Speech is saved as a .mp3 file and text is saved as a transcript.txt file.
- TTS & STT offer different voices and accents.
- TTS can convert any text that may be email, documents can be converted into speech.
- TTS allows computers or phones to read the text in a louder voice for its users.
- TTS is especially beneficial for those people who have eyesight disorder.
- TTS helps users with any sort of difficulty in reading and writing.
- TTS supports online users to listen to e-books.
- TTS supports every person in educating themselves easily.
- STT transcribes customer calls, podcasts, and streaming content.
- STT helps to take notes in real-time or have notes to refer to after calls.

## VII. CONCLUSION

The integration of two opposite TTS & STT modules has made user work easier. The advantage of saving converted speech and text also made users to refer content whenever needed in the future. TTS systems help widely in education. Online users can hear the text loudly instead of

reading. TTS helps people who have eyesight disorder and for people who have reading problems and pronunciation problems. This has made education go smoothly for children. Speech-to-Text is widely used for business purposes for writing down notes of streaming videos or call recordings of customers simultaneously while the user is on call. This made the work of specially writing notes easier.

## VIII. REFERENCES


[1] Dr. S. A. Ubale, Girish Bhosale, Ganesh Nehe, Avinash Hubale, Avdhoot Walunjkar, June 2022, "A review on Text-to-Speech Converter", IJIRT| Volume 9| Issue 1| ISSN: 2349-6002.

[2] Babu Pandipati, Dr. R. Praveen Sam, 2021, " Speech to text Conversion using Deep Learning Neural Net Methods", Turkish Journal of Computer and Mathematics Education| Vol.12.

[3] Shivangi Nagadewani, Ashika Jain, 2020, "A Review on methods for speech-to-text and text-to-speech conversion", International Research Journal of Engineering and Technology (IRJET)| e-ISSN: 2395-0056| p-ISSN: 2395-0072| Volume: 07| Issue: 05.

[4] Ayushi Trivedi, Navya Pant, Pinal Shah, Simran Sonik, and Supriya Agarwal, Mar-Apr. 2018, "Speech to text and text to speech recognition system-Areview", IOSR Journal of Computer Engineering (IOSR-JCE) |e-ISSN: 2278-0661| p-ISSN: 2278-8727| Volume 20| Issue 2| Ver. I| pp 36-43.