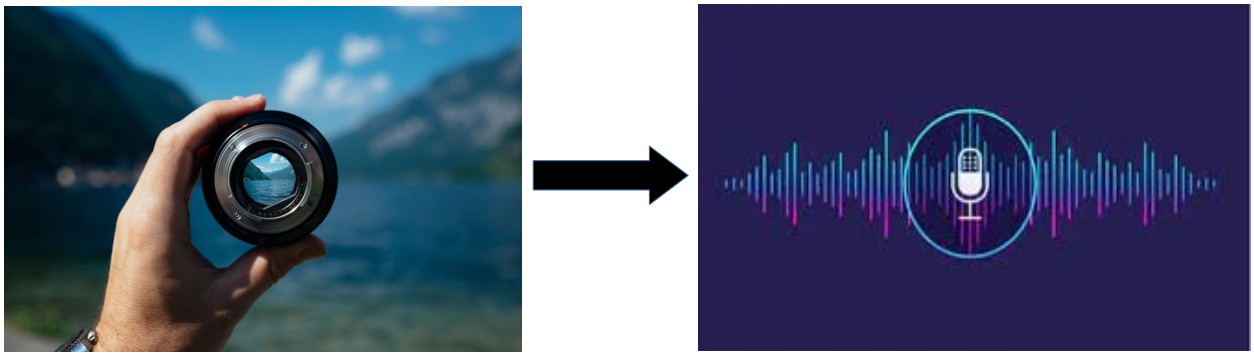
ARTIFICIAL INTELLIGENCE
MINI-PROJECT SYNOPSIS

TOPIC: IMAGE TO SPEECH CONVERSION

T3 Batch Div1
TY COMP
COEP'23

NAME	MIS	EMAIL ID
Isha Rathi	111903037	rathiir19.comp@coep.ac.in
Kimaya Abhyankar	111903041	abhyankarkn19.comp@coep.ac.in
Mihir Malani	111903046	malanimy19.comp@coep.ac.in

INDEX:



Introduction

In our planet of 7.4 billion humans, 285 million are visually impaired out of whom 39 million peopleare completely blind, i.e. have no vision at all, and 246 million have mild or severe visual impairment (WHO,2011).As reading is of prime importance in the daily routine (text being present everywhere from newspapers, commercial products, sign-boards, digital screens etc.) of mankind, visually impaired people face a lot of difficulties. Our device assists the visually impaired by reading out the text to them.

The purpose behind this project is to provide a web-based platform which converts any textual snippet into an .mp3 audible format, which the end user can listen to whenever they want .The main objective is take an input image from the user in the web interface and process it to extract text, which then shall be converted into speech.

Technologies used:

1.HTML
HyperText Markup Language is used to create the website it also provides two forms out of which user can submit query in any one.

2.CSS
Cascading Style Sheet is the collection of all the styles used to develop the website.

3.FLASK
Flask is micro framework used to render web applications and websites using python. Flask is light weight module and requires no additional libraries.

4.OPENCV
Open Computer Vision is a library of functions to perform real time computer vision. OpenCV is used to read nd write the images provided by user and to convert the image to grey scale.

5.TESSERACT-OCR
Tesseract Optical Character Recognition engine is considered to be the most accurate engine for character recognition. Tesseract OCR is the heart of the project it converts the image to a readable text.

6.PYTHON
Python framework is used for building the entire backend and the flask server. Python is the base programming language used for the project.

- **OCR ENGINE**

The extraction of the text in the image is done using optical character recognition (OCR). OCR is a field of research in pattern recognition, artificial intelligence and computer vision. It is the conversion of the images of typed, handwritten or printed text into a digital text or computer format text. Earlier OCR versions had to be trained in each character of a text with its specific font. Today, advanced OCRs are available that have a high degree of accuracy, support a wide variety of image formats, languages and fonts. For our project, we have used Tesseract OCR. It is the most accurate open source OCR engine and is powered by google. It can be used on the Linux, mac and windows platform. The newest Tesseract version supports a hundred languages.However, images must undergo a number of pre-processing stages like noise removal, scaling etc. otherwise the
output will be of low quality.

- **TTS SOFTWARE**

The process of converting text to speech by a computer is called speech synthesis. A text to speech system(TTS) is used to perform speech synthesis. A TTS is composed of two parts: front end and back end. The front end converts the text to a symbol, for example, a number. Each symbol generated is assigned a phonetic. The back end then converts the phonetic into sound.

Feasibility Study

Our device is designed for people with mild or moderate visual impairment by providing the capability to listen to the text. It can also act as a learning aid for people suffering from dyslexia or other learning disabilities that involve difficulty in reading or interpreting words and letters. We wish to enable these people to be independent and self-reliant as they will no longer need assistance to understand printed text. Such people will always have access to information hence they will never feel at a disadvantage.

The traditional way of reading a book is through a printed copy. It’s not possible to carry a physical copy of a book everywhere, that’s when e-books came in handy. Printed books as well as e-books affect the eyes and hence having an audio versions helps a lot.

Several research(s) say that listening improves the power of imagination, so the end users can simulate the scenes occurring in book faster than reading.

Literature Survey

This paper[1] proposed by H. Wang,D. Mohamad, N. A. Ismail attempts to discuss the evolution of the retrieval techniques focusing on development, challenges and trends of the image retrieval. It highlights both the already addressed and outstanding issues. However, Image retrieval researches are moving from keyword, to low level features and to semantic features.

Over past decade many researchers form computer vision and Content Based Image Retrieval (CBIR) domain have been actively investigating possible ways of retrieving images and videos based on features such as color, shape and objects

Paper [7] introduced by Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang presents a novel domain adaptation approach for solving cross domain pattern recognition problem where data and features to be processed and recognized are collected for different domains.

Mizan, C. M., Chakraborty, T., & Karmakar, S.[4] proposed that the process of Text Recognition involves several steps including preprocessing, segmentation, feature extraction, classification, post processing. Preprocessing is for done the basic operation on input image like binarization which convert gray Scale image into Binary Image, noise reduction which remove the noisy signal from image. Segmentation stage for segment the given image into line by line and segment each character from segmented line. Future extraction calculates the characteristics of character. A classification contains the database and does the comparison. Nowadays it plays an important role in office, colleges etc.

Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu [3] proposed an image parsing to text description that generates text for images and video content. Image parsing and text description are the two major tasks of his framework. It computes a graph of most probable interpretations of an input image. This parse graph includes a tree structured decomposition contents of scene, pictures or parts that cover all pixels of image.

In their Paper[5] Allen, John, Hunnicutt, Sharon, and Dennis Klatt stated that progress in this area has been made possible by advances in linguistic theory, acoustic–phonetic characterization of English perceptual mathematical of production, structured programming, and computer hardware design.

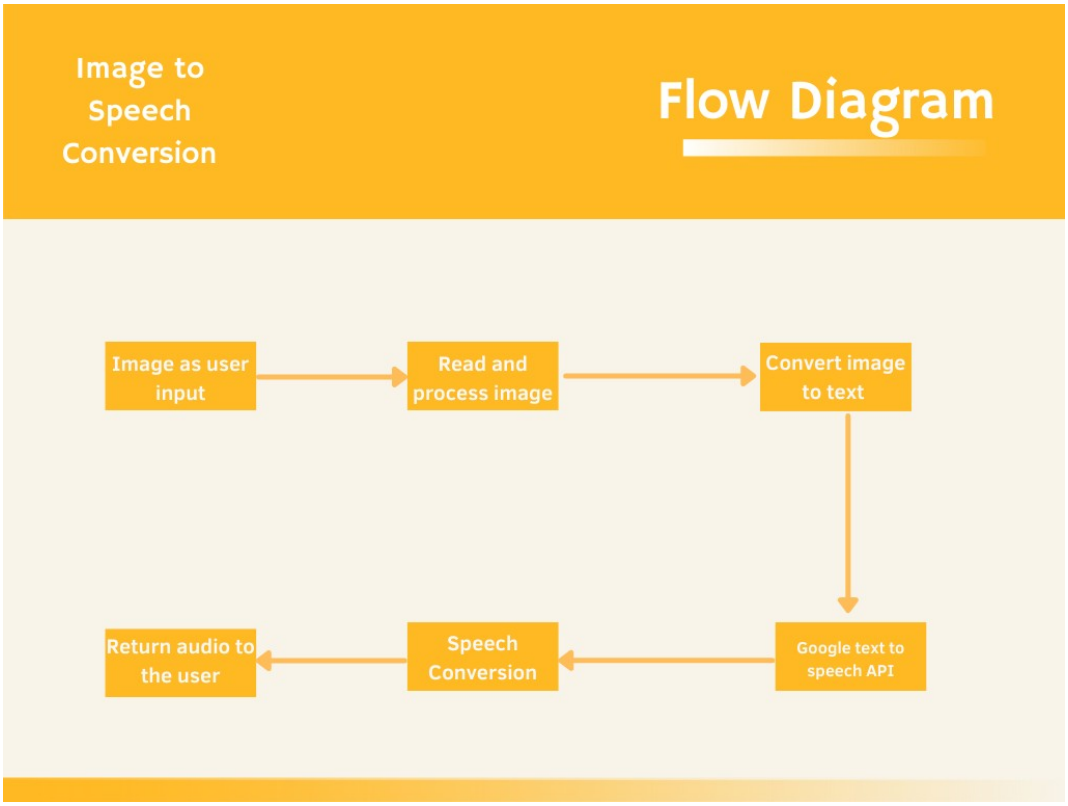
In this Paper[6] , Mattingly I. G described phonetics as traditionally concerned with the ways in which the sounds of speech are produced, but the resulting descriptions normally mix auditory factors with articulatory ones, thus depending ultimately upon percepts of the phonetician

In his paper[7] A. Coates et al stated that reading text from photographs is a challenging problem that has received a significant amount of attention. Two key components of most systems are (i) text detection from images and (ii) character recognition, and many recent methods have been proposed to design better feature representations and models for both. In this paper, we apply methods recently developed in machine learning -- specifically, large-scale algorithms for learning the features automatically from unlabeled data -- and show that they allow us to construct highly effective classifiers for both detection and recognition to be used in a high accuracy end-to-end system.

Research Gaps

- There are Text-to-phoneme challenges like determining the tone and pronunciation of a word/phrase based on the spelling in context.
- There are many words in the English vocabulary which contextually require a different pronunciation.
- Image Segmentation is done to the noiseless grayscale image. If the image is noisy, then image to text conversion would be difficult.
- Evaluation challenges occur while processing speech. Hence, there is always a compromise between the production proficiency and replay prerequisite required in speech synthesis.

Methodology/ Planning of Work



Steps Involved:

1. **Input image**: Image is taken from the user, on our website

2. **Reading and Processing**: In this step, the image is first read, using python's openCV library. In the next step, noise removal and filtering is done and then the images are converted to an array of 0's and 1's. The image is then processed and converted into gray scale in order to pass it into the OCR test

3. **Converting image to text**: Images are converted into text by using the Tesseract OCR Engine.The OCR engine converts the typed data in the image to string format. The string's accuracy depends on the clarity of the image.

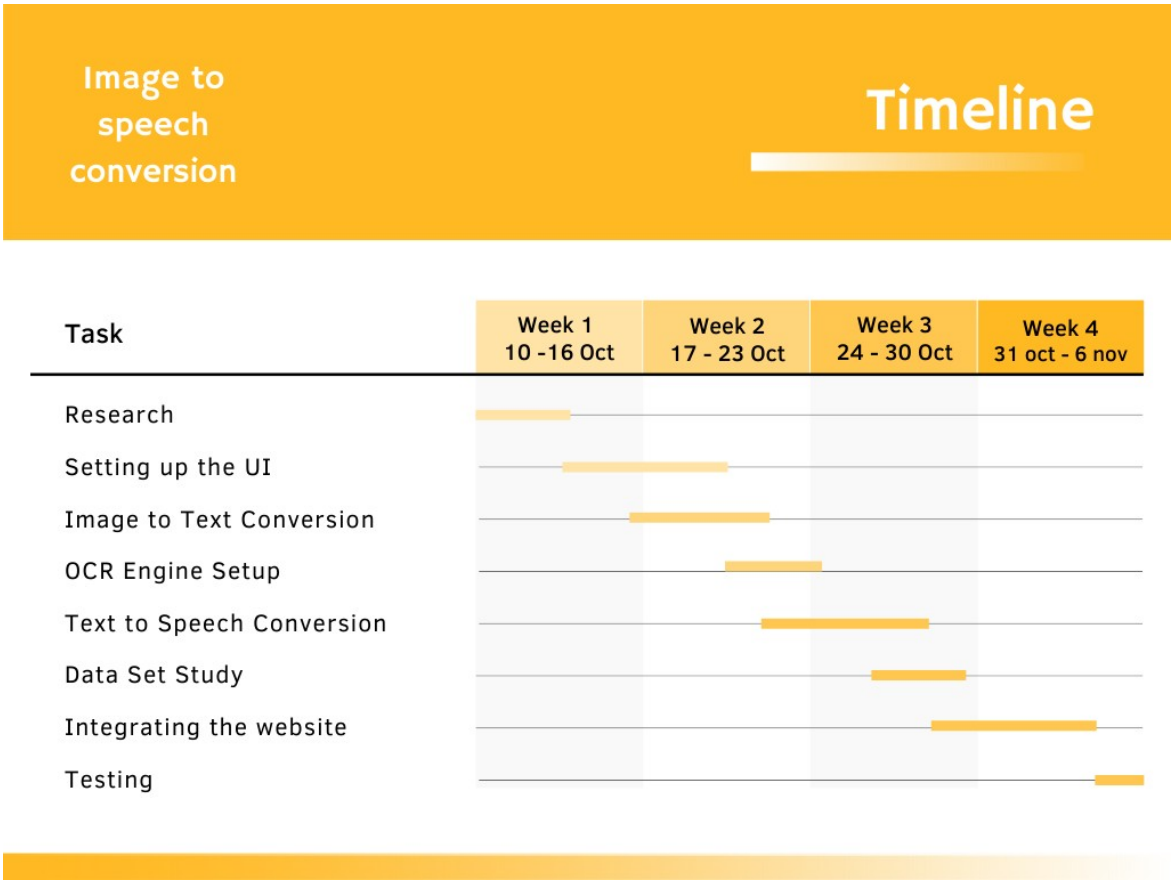
4. **Speech Conversion**: Google Text to Speech API gTTS is used to convert the generated string from previous step to audio format. The input text is first analyzed, normalized and transcribed into a phonetic or some other linguistic representation.

5. The generated mp3 audio file is then returned to the user on the website, using flask.

Facilities Required

- DATASETS
- PYTHON LIBRARIES
- TESSARACT OCR
- GOOGLE TTS

Timeline



Bibliography

[1] H. Wang, D. Mohamad, N. A. Ismail, “Image Retrieval: Techniques, Challenge, and Trend” Computer Science World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 2009

[2] Y. Rui, T. S. Huang, and S. F. Chang, “Image retrieval: Current techniques, promising directions, and open issues,” Journal of Visual Communication and Image Representation, vol. 10,1999.

[3] Benjamin Z. Yao, Xiong Yang, Liang Lin, Mun Wai Lee and Song-Chun Zhu, “I2T: Image Parsing to Text Description” IEEE Conference on Image Processing, 2008 .

[4] Mizan, C. M., Chakraborty, T., & Karmakar, S. (2017). Text Recognition using Image Processing. International Journal of Advanced Research in Computer Science, 8(5)

[5] Allen, John, Hunnicutt, Sharon, and Dennis Klatt, Text To Speech, The MITTALK System (Cambridge: Cambridge University Press, 1987).

[6] Mattingly I. G., Speech Synthesis for Phonetic and Phonological Models, T.A. Sebeok (Ed.) Current Trends in Linguistics, Vol. 12, (1974) p. 2451-2487

[7] A. Coates et al., “Text Detection and Character Recognition in Scene Images with Unsupervised Feature Learning.”

[8] Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang FrankWang, “Heterogeneous Domain Adaptation and Classification by Exploiting the Correlation Subspace,” IEEE Transactions on Image Processing, vol. 23, no. 5, May 2014