# Object detection and conversion of text to speech for visually impaired

**Ankur Jyoti Sarmah[1], Kabindra Bhagawati[2], Kaustov Duwarah[3], Swetashree Dey Purkayastha[4], Antarjeeta Boro[5], Divika Muchahary[6]**

[1,2]Assistant Professor, Department of Electronics and Telecommunication Engineering,
*Assam Engineering College,*
*Jalukbari, Guwahati*
*ankur.et@aec.ac.in*

[3,4,5,6]B.Tech 8th Semester, Department of Electronics and Telecommunication Engineering,
*Assam Engineering College,*
*Jalukbari, Guwahati*
*work.kaustov@gmail.com*

***Abstract*: *Assistive technologies are being developed for visually impaired people in order to live confidently[1]. In this project work, we aim to develop a system which would help blind persons get information about objects present in their surroundings in their daily lives. The project work is framed into two stages. First, image is captured using a portable camera module, if the object is identified as cell phone, person, book or as such, then the detected object is matched with a predefined dataset. A predefined dataset is loaded in order to match the detected object with the captured image. Secondly, once it is matched the recognized text is synthesized for producing speech output. Text to speech conversion successfully converts the detected object into an audio signal using the gTTs Module with the help of the iPython audio library. The objective of the TTS is the conversion of text into a natural language. It is not only applicable for the visually impaired but also to any normal human beings who are willing to read the text as a speech as quickly as possible. The entire system is de-signed using the YOLOv4 model trained on the MS COCO dataset through a laptop. The entire system is processed by using Python as the programming language.***

***Keywords*:Visually Challenged Person, Object detection, Optical Character Recognition (OCR), text detection, Text-to- Speech (TTS), Python, dataset.**

**(Article history: Received: 8th March 2023 and accepted 11th August 2023)**

## I. Introduction

In our planet of 8 billion humans, at least 2.2 billion people have a near or distance vision impairment. In at least 1 billion or almost half of these cases, vision impairment could have been prevented or has yet to be addressed. This 1 billion people include those with moderate or severe distance vision impairment or blindness due to unaddressed refractive error (88.4 million), cataract (94 million), age-related muscular degeneration (8 million), glaucoma (7.7 million), diabetic retinopathy (3.9 million), as well as near vision impairment caused by un-addressed presbyopia (826 million). It is a fact that all over the world that the visually impaired (partially or completely blind) people face a lot of difficulties in identifying products and avoiding obstacles. The greatest inconveniences that a blind person feels in everyday life include finding information about objects and indoor mobility problems. They have difficulty recognizing simple objects, and it is not easy to distinguish objects that have similar forms[2]. Object Detection is a field of Computer Vision that detects instances of semantic objects in images or videos (by creating bounding boxes around them in our case. In this paper, we analyse object detection using a deep learning object recognition technique. Additionally, a voice assistance technology is introduced so the visually impaired people can know about the object or obstacle around them..

## II. OBJECTIVE

The main objectives of the project are as follows:

1. To help specially abled people who aren't able to see through their eyes to read and feel the surroundings around them via audio signals.

2. To devise a system that converts the text to audio signals that can be heard via a speaker.

## III. Methodology

The functional block diagram of the proposed system describes object/text to speech conversion which is shown below.

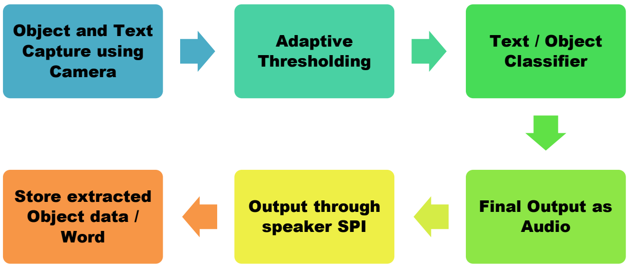
Object and Text Capture using Camera
Adaptive Thresholding
Text / Object Classifier
Final Output as Audio
Output through speaker SPI
Store extracted Object data / Word

*Figure 1Functional block diagram of Object/Text to Speech conversion*

There are six major functional blocks in the block diagram which has been explained in detailed below:

### A. Object & Text Capture using Camera:

The objects and texts to be detected are being captured using camera and converted to an OpenCV image.

### B. Adaptive thresholding:

Adaptive thresholding typically takes a grayscale or color image as input and, in the simplest implementation, out-puts a binary image representing the segmentation. For each pixel in the image, a threshold has to be calculated. If the pixel value is below the threshold it is set to the background value, otherwise it assumes the foreground value.

There are two main approaches to finding the threshold: (i) the Chow and Kaneko approach and (ii) local thresholding. The assumption behind both methods is that smaller image regions are more likely to have approximately uniform illumination, thus being more suitable for thresholding. Chow and Kaneko divide an image into an array of overlapping sub images and then find the optimum threshold for each sub image by investigating its histogram. The threshold for each single pixel is found by interpolating the results of the sub images. The drawback of this method is that it is computational expensive and, therefore, is not appropriate for real-time applications.

An alternative approach to finding the local threshold is to statistically examine the intensity values of the local neighbourhood of each pixel. The statistic which is most appropriate depends largely on the input image. Simple and fast functions include the mean of the local intensity distribution,

$$T = mean$$

The median value

$$T = median$$

Or the mean of the minimum and maximum values

$$T = \frac{\max + min}{2}$$

The size of the neighbourhood has to be large enough to cover sufficient foreground and background pixels, otherwise a poor threshold is chosen. On the other hand, choosing regions which are too large can violate the assumption of approximately uniform illumination. This method is less computationally intensive than the Chow and Kaneko approach and produces good results for some applications [3].

### C. Text/ Object classifier

Text classification is a common NLP task that assigns a label or class to text. We use various Machine Learning algorithms to classify the object/ text for the ease of processing the data and training the data to be able to extract information and convert the detected data into audio signals

### D. Output through Speaker SPI

A serial peripheral interface (SPI) is an interface that enables the serial (one bit at a time) exchange of data between two devices, one called a master and the other called a slave. An SPI operates in full duplex mode. This means that data can be transferred in both directions at the same time. We use the SPI to output the audio signals processed by the iPython library.

## IV. System Implementation

### A. Google Colaboratory

Colaboratory, or “Colab” for short, is a product from Google Research. With colab we have executed the python code through the browser. It is especially well suited to machine learning, data analysis and education that allows us to combine Python code and rich text along with charts, images, HTML, LaTeX and more into a single document stored in Google Drive. It connects to powerful Google Cloud Platform runtimes and enables you to easily share your work and collaborate with others. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.

### B. YOLO

YOLO is real-time object detection. It applies one neural network to the complete image dividing the image into regions and predicts bounding boxes and possibilities for every region.

Predicted probabilities are the basis on which these bounding boxes are weighted. A single neural network predicts bounding boxes and class possibilities directly from full pictures in one evaluation. Since the full detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

### C. Google Text to Speech (gTTS)

TTS (Text-to-speech) is the process of converting words into an audio form. The tool takes an input text from the user, and performs logical inference on the text. This processed text is passed into the next block where digital signal processing is performed on the processed text. Using many algorithms and transformations this processed text is finally converted into a speech format. This entire process involves the synthesizing of speech. There are several APIs available to convert text to speech in Python. One of such APIs is the Google Text to Speech API commonly known as the gTTS API. gTTS is a very easy to use tool which converts the text entered, into audio which can be saved as a mp3 file. The gTTS API supports several languages including English, Hindi, Tamil, French, German and many more.

## RESULT ANALYSIS

In this project, we have till now attained the required text to speech conversion to successfully convert the detected object into an audio signal using the gTTs Module with the help of the iPython audio library. The final detected output can be seen in the following figure where, using the YOLOv_4 model trained on the MS COCO dataset, we have detected the objects through the laptop camera and conversion of the same to speech is attained by using the gTTS module with the help of iPython audio library.

## CONCLUSION

By implementing this system, visually impaired can easily listen whatever they want to listen. And with the help of the translation tools, he can convert the text to the desired language and then again by using the Google speech recognition tool he can convert that changed text into voice. By that they can be independent. And it is less cost compared to other implementations. Text-to-Speech device can change the text image input into sound with a performance that is high enough and a readability tolerance of less than 2%, with the average time processing less than three minutes for A4 paper size. This portable device, doesnot require internet connection, and can be used independently by people. Through this method, we can make editing process of books or web pages easier.

## FUTURE SCOPE OF THIS WORK

The proposed model of this project is already in a feasible state of use but there are some future works for this model.

1) Recognizing process of large texts is sometimes slow, for that reason introducing of some distributed technique is reserved for future work.

2) Now this model can read only English language, but in future it should allow Assamese language as well.

3) If any image or diagram is present in between text, then recognition of that image or diagram is an important task, which is also reserved for future work.

By extending these future works, many numbers of people will get more benefits from this reader.

## REFERENCES

[1] Abhishek, R.; Kumar, K. N.; Karthik, R.; Puskhal, R.; Kumar, S. A. Smart Gadget Product Label Reading Using OCR Algorithm & TTS Engine. International Journal of New Technology and Research 4 (4), 70–72.

[2] S, B.; D, L. Image to Audio Conversion Using Portable Camera. Journal of Electrical & Electronic Systems 2018, 07 (03). https://doi.org/10.4172/2332-0796.1000268.

[3] Fisher, R.; Perkins, S.; Walker, A.; Wolfart, E. Adaptive Thresholding https://homepages.inf.ed.ac.uk/rbf/HIPR2/adpthrsh.htm (accessed 2023 -02 -06).

.