Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

Talkie Text: The Image Reader

Sadhana Suresh Chettiar, Bhuvana P, Harshitha P, Bhavana N M
Guided by: Dr Sahana D Gowda
Department of Computer Science
BNM Institute of Technology
(Affiliated to VTU, Karnataka)
Bangalore, India

***Abstract*— "Talkie Text" is an application based on text recognition from image, and text-to-speech conversion. It converts the text with in an image into speech format and reads it out. Image acquisition, recognition and speech conversion using Optical Character Recognition (OCR) and Text to Speech synthesizer (TTS) is an Image Processing Technology used to convert the image containing horizontal text into text documents and the extracted text is converted into speech. Text-To-Speech Synthesis is a technology that provides a means of converting written text from a descriptive form to a spoken language that is easily understandable by the end user. Text information in natural scene images serves as important clues for many image-based applications such as scene understanding, content-based image retrieval, assistive navigation, and automatic geocoding. The text present with image as a background will be detected, located, and then extracted. This text which is extracted from the complex background will be converted into speech reading out the content of the document. The input provided will be the document containing the text and images wherein only the textual content will be extracted from the complex background and the extracted text will be converted into speech.**

Keywords—Optical Character Recognition, Region Of Interest, Canny Edge, Text-to-Speech, Phonemes, Concatenative Synthesis

I. INTRODUCTION

With the present explosion of data circulating the digital space, which is mostly non structured textual data, there is a need to develop automatic text recognition tools that allow people to get insights from them easily. Locating text from a complex background with multiple colors is a challenging task. "Talkie Text" aims at detecting text in images and converting it into speech.

The first module aims in extraction of the text by separating the images depending on the contours, binarization, and color so that the text region is found by detecting mechanisms. The detected text will be further processed to identify the coordinates of the text and must be separated from the background. In order to be successfully recognizable by an OCR system, an image having text must fulfill certain requirements like a monochrome text and background where the background to-text contrast should be high. The proposed system strives toward methodologies that aid automatic detection, segmentation and recognition of visual text entities in complex natural scene images. This extracted text will be given to next module text-to-speech.

Text-to-speech (TTS) is a type of assistive technology that reads digital text aloud. It's sometimes called "read aloud" technology, since TTS can take words on a computer or other digital device and convert them into audio. The text will be analyzed, processed, and mapped to the corresponding speech units in the database. The text in the final stage will be in the form of speech which will be read out to the user. The two modules will be working by fetching the output of first module and processes it converting into speech which is obtained as the outcome from the second module. Thus, speech would be more beneficial and better for time optimization. Hence, the project can be used for extracting the text from complicated background and reading it out in a better form.

II. REVIEW OF EXISTING METHOLOGIES

A. Text to Speech Module

Text-to-Speech module aims to provide a user- friendly application to general users. This paper proposes a method at developing a complete system in which Text can be converted to Speech, Text file can be converted to Speech, Text in various Languages can be converted to Speech, Image can be converted to Text and Image can be converted to Speech using MATLAB as a programming tool. The various methods used are Preprocessing, Unicode Conversion, Segmentation, Concatenation, Prosody and Smoothing, to be then combined in an application for easy access and usability. The text-to-speech mode converts a text file or inputted text to speech which then is narrated/read using the voice database used by Microsoft SAPI. Microsoft SAPI is an application developed that allows the use of speech recognition and speech synthesis.

This processing is done using phonemes and concatenating syllables using optimal coupling algorithm. Currently the various methods used for text to speech conversion are Concatenation Synthesis which includes unit selection, diphone and domain specific synthesis. Other methods include Formant, Articulatory, HMM based and sine wave synthesis. Formant synthesis produces an output by using additive synthesis of an acoustic representation in the form of a model. Articulatory synthesis is a method of synthesizing machine speech based on simulations of the vocal tract of humans and their respective articulatory processes. Text to speech conversion can be accomplished by starting with the method of pre processing of the input text. Here the text abbreviations, acronyms and numbers are expanded. The pre-processed text will then be option between exiting the app or reusing it to operate other functions. Since it is coded using modular programming into independent functions which allows users in the future to add more functionalities to this application. However, Microsoft SAPI is confined to only windows application.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

B. Survey on Various Methods of Text to Speech Synthesis

It provides an overview of existing Text-To-Speech synthesis techniques. There are mainly three categories of Text to speech synthesis categorized into, Formant Based, Concatenative based and Articulatory. Formant based speech synthesis relies on different techniques such as cascade, parallel, klatt and PARCAS model etc. Concatenative speech synthesis can be broadly categorized into three categories, Diphone based, Corpus based and Hybrid. Concatenative method provides more natural and individual sounding speech, but the quality with some consonants may vary considerably and the controlling of pitch and duration may be in some cases difficult, especially with longer units. Articulatory synthesis involves Vocal Tract Models, Acoustic Models, Glottis Models, Noise Source Models. With concatenation methods the collecting and labeling of speech samples have usually been difficult and very time-consuming. With formant synthesis the quality of synthetic speech is more constant, but the speech sounds slightly more unnatural and individual sounding speech is more difficult to achieve. Formant synthesis is also more flexible and allows a good control of fundamental frequency. The third basic method, the Articulatory synthesis, is perhaps the most feasible in theory especially for stop consonants because it models the human articulation system directly. On the one hand, the Articulatory based methods are usually rather complex and the computational load is high. In this paper, all text to speech synthesis methods are explained with their pros and cons. A typical text to speech system has two parts – a front end and a back end. The front end is responsible for text normalization (the pre-processing part) and text to phoneme conversion. Text normalization or tokenization is the phase where numbers and abbreviations in the raw text are converted into written words. Text to phoneme conversion or grapheme-to-phoneme conversion is the process of assigning phonetic transcriptions to each word and dividing them into prosodic units such as phrases, clauses, and sentences. The output of the front-end system is the symbolic linguistic representation of the text. It is composed of the phonetic transcriptions along with the prosody information. This output is then passed on to the back-end system or the synthesizer, which converts it into sound. This paper is it gives a clear introduction about text to speech synthesis and reviews the various methods for text to speech synthesis in a comprehensible way. They emphasize on the methodologies of the three major Text to Speech conversion methods, but fail to conclude which method is the most feasible one.

C. Almost Unsupervised Text to Speech and Automatic Speech

This paper proposes an almost unsupervised learning method that only leverages few hundreds of paired data and extra unpaired data for TTS and ASR, since Text to speech (TTS) and automatic speech recognition (ASR) are two dual tasks in speech processing. The method explained in the paper consists of the following components: a de-noising auto-encoder, which reconstructs speech and text sequences respectively to develop the capability of language modeling both in speech and text domain; dual transformation, where the TTS model transforms the text into speech, and the ASR model leverages the transformed pair for training, and vice versa, to boost the accuracy of the two tasks, bidirectional sequence modeling, which addresses error propagation especially in the long speech and text sequence when training with few paired data; a unified model structure, which combines all the above components for TTS and ASR based on Transformer model. The works trying to synthesize the voice of a certain speaker with few samples leverage large amount labeled speech and text data from other speakers, which is usually regarded as a transfer learning problem but not an unsupervised learning problem. This method achieves 99.84% in terms of word level intelligible rate and 2.68 MOS for TTS, and 11.7% PER for ASR on LJSpeech dataset, by leveraging only 200 paired speech and text data (about 20 minutes audio), together with extra unpaired speech and text data. The project, along with the experimentation has been implemented using the dataset, giving us a clear insight of how the system works. This paper has proposed the almost unsupervised method for text to speech and automatic speech recognition, which leverages only few paired speech and text data and extra unpaired data.

D. An Improved Scene Text and Document Image Binarization Scheme

In this paper, a novel approach to natural scene text image binarization by tracking the text boundary based on edge and gray level variance information is proposed. Further, broken boundaries are linked to construct the complete boundary map. Here, an adaptive threshold is determined based on boundary edge information to binarize the image effectively In their proposed binarization scheme, at first, a color scene text image, is converted into a grayscale image. This grayscale image is the input of our proposed binarization scheme. The proposed method for scene text binarization is divided into three major parts: (i) Variance Computation (ii) Broken Edges (Boundary) Linking and (iii) Adaptive Thresholding. Text connected components have a few basic properties based on which they can be separated from the image - (a) they have a distinct boundary that separates it from non-text regions (b) the whole grayscale value of the text components region is almost the same and is Talkie Text dissimilar from the background non-text components. In this, work experiments are conducted on the datasets of ICDAR 2003 Robust Reading Competition, ICDAR 2011 Born Digital Dataset, Street View Text (SVT) Dataset, DIBCO dataset and our laboratory made Bangla Dataset. Text connected components scene text and document image binarization methodology. It uses both the edge and variance information of the input image. The proposed scheme is not very sensitive to image color, text font, skew and perspective variation. The proposed method is effective in terms of low contrast, non-uniform illumination and noisy text based scene images. It is not very sensitive to image colour, text font, skew and perspective variation.

E. An Improved Scene Text and Document Image Binarization Scheme

Text Recognition is to recognize the text from printed hardcopy document to desired format (like .docx). The process of Text Recognition involves several steps including pre-processing, segmentation, feature extraction, classification, post processing. Preprocessing is done for the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

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. Generally text-detection methods can be classified into three categories. The first one consists of connected component-based methods, which assume that the text regions have uniform colors and satisfy certain size, shape, and spatial alignment constraints. However, these methods are not effective when the text have similar colors with background. The second one consists of the texture based methods, which assume that the text regions have special texture. Though these methods are comparatively less sensitive to background colors, they may not differentiate the texts from the text-like backgrounds. The third one consists of the edge-based methods. The text regions are detected under the assumption that the edge of the background and the object regions are sparser than those of the text regions. An algorithm is proposed for solving problem of offline character recognition. The input is in the form of images. The algorithm was trained on the training data that was initially present in the database. Preprocessing, segmentation and detecting the line has been done. A brief survey of the applications in various fields, along with experimentation into few selected fields has been presented. The proposed method is extremely efficient to extract all kinds of bimodal images including blur and illumination. The paper will act as a good literature survey for researchers starting to work in the field of optical character recognition. This method is efficient to extract all kinds of bimodal images including blur and illumination. However, it is hard to extract the complete layout structure of very low-resolution images.

III. PROPOSED METHODOLOGY

A. Overview

1) Text Extraction from Image:

The Text Extraction module consists of the following methods: Image Pre-processing, Region of Interest, Segmentation and Optical Character Recognition.

a) Image Pre-processing: The preprocessing of image aims at selectively removing the redundancy present in captured images without affecting the details that play a key role in the overall process. The data that we collect or generate is mostly raw data, i.e. it is not fit to be used in applications directly. Therefore, we need to analyze it first, perform the necessary pre-processing, and then use it. It involves many techniques such as resizing, noise removal, RGB to grayscale, deskew, Thresholding, etc.

b) Region of Interest: The cropping box might be limited in terms of geometry. In order to extract a fiber bundle more precisely, you can define regions in the image where you want to visualize fibers that go through them. These regions are called Regions Of Interest (ROIs).

c) Image Segmentation: Image Segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze.

d) Optical Character Recognition: Optical character recognition or optical character reader is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene-photo or from subtitle text superimposed on an image.

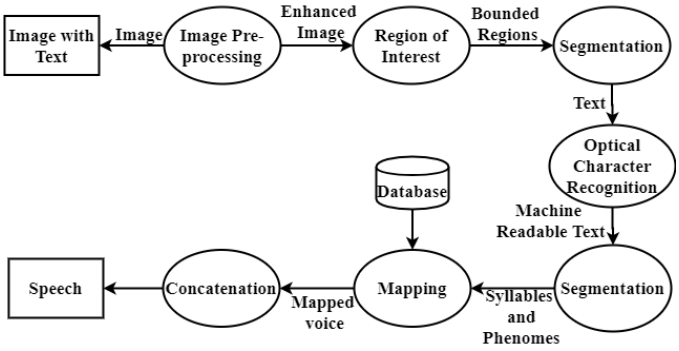

Fig. 1. Overview of Proposed System

2) Text to Speech Synthesis:

The text to speech synthesis module consists of the following methods: Pre-processing, Unicode conversion, Segmentation, concatenation and generation of the speech as output.

a) Pre-processing: It converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words which is referred as normalization. Then the text is either converted into uppercase or lowercaseformat and all the stop words like a, the, etc., are removed.

b) Segmentation: It is the process of dividing written text into meaningful units, such as words, sentences, or topics. These segments categorized as word, sentence, topic, phrase or any information unit depending on the task of the text analysis. The splitting of words into words, phrases and symbols or any meaningful units are referred to as tokens. The encoded text is the segmented into syllables and are then mapped to the pre-recorded voice database.

c) Mapping: These tokens and syllables are then mapped to a pre-recorded voice database.

d) Concatenation: It involves synthesizing sounds by concatenating short samples of recorded sound (called units). Then these concatenated modules are then output as speech.

B. Methods Used

1) Image resizing: The initial step is the resizing of the image. This is done mainly to generalize the size of the display to a standard set of dimension, irrespective of the images' original size. The Image module provides a class with the same name.

2) Image Deskewing: The process of straightening an image that has been scanned or photographed crookedly—that is an image that is slanting too far in direction or one that is misaligned. This process is done in the post-skew detectionstage where it is determined whether the image is skewed. We have implemented using the angle calculation method:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

a) We create a binary image from our original image. This step is important because that way we can detect only the important parts of the image.

b) After we have a binary image we can use morphology to detect areas where we have text. OpenCV has a function where we can detect lines on the image, using binary threshold. So with that function we can find the angle.

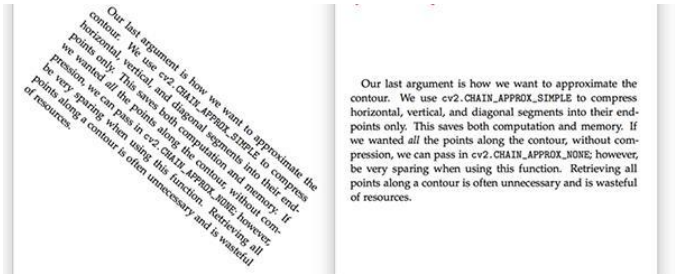

Fig. 2. Image Deskewing

c) The next step is the angle calculation, which is done using an in-built python method. After we used binary thresholding, we can calculate the start and end point of the line we detected. And from that we can calculate the angle we need.

d) The final step is the image rotation to the angle of rotation derived from the previous step. From the angle and from the image center Point we will know the rotation matrix in 2D. So we calculate a bounding box what we will need and then rotate it with our rotation matrix.

3) Grayscale Conversion: A RGB image can be viewed as three images (a redscale image, green scale image and blue scale image) stacked on top of each other. The various color spaces exist because they represent color information in ways that make certain calculations more convenient because they provide a way to identify colors that is more intuitive. For example, the RGB color space defines a color as the percentages of red, green, and blue hues mixed together. Other color models describe colors by their hue (shade of color), saturation (amount of gray or pure color), and luminance (intensity, or overall brightness). Similarly, a grayscale image can be viewed as a single layered image. It is basically M*N array whose values have been scaled to represent intensities. Conversion of a color image into a grayscale image inclusive of salient features is a complicated process. The converted grayscale image may lose contrasts, sharpness, shadow, and structure of the color image. To preserve contrasts, sharpness, shadow, and structure of the color image, Python has introduced color_bgr2gray method. To convert the color image into grayscale image, the function performs RGB approximation and reduction. The luminance of a pixel value of a grayscale image ranges from 0 to 255. The conversion of a color image into a grayscale image is converting the RGB values (24 bit) into grayscale value (8 bit). Various image processing techniques and software applications converts color image to grayscale image.

Fig. 3. Grayscale Conversion

4) Image Smoothing or Noise Removal: Filtering is a technique for modifying or enhancing an image. An image can be filtered to emphasize certain features or remove other features. Image processing operations implemented with filtering include smoothing, sharpening, and edge enhancement. Filtering is a neighborhood operation, in which the value of any given pixel in the output image is determined by applying some algorithm to the values of the pixels in the neighborhood of the corresponding input pixel. A pixel's neighborhood is some set of pixels, defined by their locations relative to that pixel. Here, the function cv2.medianBlur computes the median of all the pixels under the kernel window and the central pixel is replaced with this median value. The kernel size must be a positive odd integer. This is highly effective in removing salt-and-pepper noise.

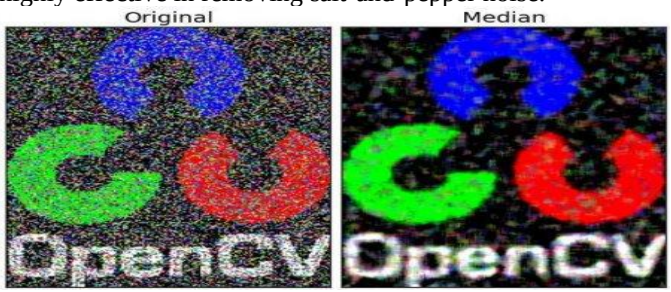

Fig. 4. Image Smoothing

5) Morphological transformations: Morphological transformations are some simple operations based on the image shape. Morphological image processing is a collection of non-linear operations related to the shape or morphology of features in an image. Morphological operations can also be applied to grayscale images such that their light transfer functions are unknown and therefore their absolute pixel values are of no or minor interest. Morphological techniques probe an image with a small shape or template called a structuring element. The structuring element is positioned at all possible locations in the image and it is compared with the corresponding neighborhood of pixels. Some operations test whether the element "fits" within the neighborhood, while others test whether it "hits" or intersects the neighborhood. It is normally performed on binary images. It needs two inputs, one is our original image, and second one is called structuring element or kernel which decides the nature of operation. Two basic morphological operators are Erosion and Dilation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

6) Canny Edge Detection: Canny edge detection is a technique to extract useful structural information from different vision objects and dramatically reduce the amount of data to be processed. It has been widely applied in various computer vision systems. Canny has found that the requirements for the application of edge detection on diverse vision systems are relatively similar. Thus, an edge detection solution to address these requirements can be implemented in a wide range of situations. The general criteria for edge detection include:

a) Detection of edge with low error rate, which means that the detection should accurately catch as many edges shown in the image as possible

b) The edge point detected from the operator should accurately localize on the center of the edge.

c) A given edge in the image should only be marked once, and where possible, image noise should not create false edges.

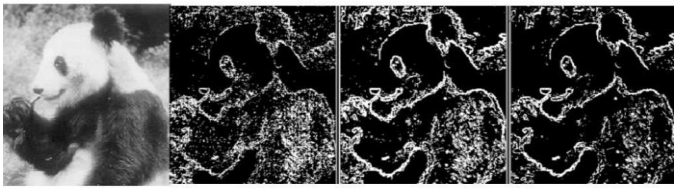

Fig. 5. Canny Edge Detection

To satisfy these requirements Canny uses the calculus of variations – a technique which finds the function which optimizes a given functional. The optimal function in among the edge detection methods developed so far, Canny edge detection algorithm is one of the most strictly defined methods that provides good and reliable detection. Owing to its optimality to meet with the three criteria for edge detection and the simplicity of process for implementation, it became one of the most popular algorithms for edge detection.

7) Optical Character Recognition: The detected text from the canny edge algorithm is then recognized and extracted using PyTesseract OCR. Python-tesseract is an optical character recognition (OCR) tool for python. That is, it will recognize and "read" the text embedded in images. Python-tesseract is a wrapper for Tesseract-OCR Engine. It is also useful as a stand-alone invocation script to tesseract, as it can read all image types supported by the Pillow and Leptonica imaging libraries, including jpeg, png, gif, bmp, tiff, and others.

8) Text-to-speech (TTS): A text-to-speech (TTS) synthesizer is a computer-based system that should be able to read any text aloud, whether it was directly introduced in the computer by an operator or scanned and submitted to an Optical Character Recognition (OCR) system. The goal of a text to speech system is to convert an arbitrary given text into a spoken waveform. Main components of text to speech system are: Text processing and Speech generation.

a) Text Processing: The text processing technique consists of Tokenisation and Normalization. Tokenization is a way of separating a piece of text into smaller units called tokens. Here, tokens can be either words, characters, or subwords. Hence, tokenization can be broadly classified into 3 types – word, character, and subword (n-gram characters) tokenization. For example, consider the sentence: "Never give up". The most common way of forming tokens is based on space. Assuming space as a delimiter, the tokenization of the sentence results in 3 tokens – Never-give-up. As each token is a word, it becomes an example of Word tokenization. Similarly, tokens can be either characters or subwords. For example, "smarter":
Character tokens: s-m-a-r-t-e-r
Subword tokens: smart-er

Text normalization is the process of transforming a text into a canonical (standard) form. For example, the word "goooood" and "gud" can be transformed to "good", its canonical form. Another example is mapping of near identical words such as "stopwords", "stop-words" and "stop words" to just "stopwords".

b) Speech Generation: The output obtained from the text normalization step is then converted into phonemes using the grapheme-to-phoneme conversion. The grapheme-to-phoneme conversion is the translation of a written text into the corresponding stream of phonemes. Phonemes are perceptually distinct units of sound in a specified language that distinguish one word from another. For example p, b, d, and t in the English words pad, pat, bad, and bat are phonemes. The pre-recorded sound for each phoneme is stored in the database in the form of individual speech units. The phonemes are mapped to their corresponding individual speech units and concatenated using concatenative synthesis. Concatenative synthesis is based on the concatenation of segments of recorded speech. It is characterized by storing, selecting, and smoothly concatenating pre-recorded human utterances (phonemes, syllables, or longer units). Concatenative synthesis produces the most natural-sounding synthesized speech. Finally, Speech synthesis block generates the speech signal, that is, converts the symbolic linguistic representation into sound.

IV. ANALYSIS AND RESULT

Before detecting text from images to improve the performance of OCR, image undergoes several preprocessing methods such as gray-scale conversion, resizing, deskewing, noise removal. Each of these methods were individually operated and tested.

Fig. 7 shows the results of each of the text extraction methods in detail. The text that is identified from this module is then sent to the text to speech module which generates a wave file (.wav).

The accuracy score percentage gives the performance of the model. This score can be calculated by

$$\text{Accuracy} = \frac{\text{Total number of correct predictions}}{\text{Total number of predictions}}$$

TABLE I. RESULT ANALYSIS

Expected Output	Obtained Output	Accuracy
STAR WARS	STAR WARS	100%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

Fig. 6. Talkie Text stepwise result

V. CONCLUSION

Talkie text is a combination of two complex systems, binding together to enable a better utility. Planning and implementing a successful text extraction and text-to-speech system is a complex enterprise— not just a single procedure but several simultaneous ones. Together, they comprise the basic strategies, systems, and procedures that are essential to an effective program. This system can be enhanced by improving the quality of the speech generation and the extraction of the text without any glitches in the output generated. This system can be extended to video text extraction process in which the text appearing in the frames of video can be detected, extracted and further converted to speech form.

REFERENCES

[1] Hussain Rangoonwala, Vishal Kaushik, P. Mohith, Dhanalakshmi Samiappan, "Text to Speech Conversion Module",2017 International Journal of Pure and Applied Mathematics, India.

[2] Desai Siddhi, Jashin M. Verghese, Desai Bhavik, "Survey on Various Methods of Text to Speech Synthesis", 2017 International Journal of Computer Applications, Laxminarayan Institute of Technology, Sarigam, India.

[3] Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, Tie-Yan Liu, "Almost Unsupervised Text to Speech and Automatic Speech Recognition", 2019. 36th International Conference on Machine Learning, Microsoft Research Asia, Zhejiang University, China.

[4] Ranjit Ghoshal, Ayan Banerjee, " An Improved Scene Text and Document Image Binarization Scheme", 4th International Conference on Recent Advances in Information Technology, RAIT-2018.

[5] Chowdhury Md Mizan, Tridib Chakraborty* and Suparna Karmaka , “Text Recognition using Image Processing", 2017 International Journal of Advanced Research in Computer Science(IJARCS).