iJRASET
International Journal For Research in Applied Science and Engineering Technology



# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 9 **Issue:** VII **Month of publication:** July 2021

**DOI:** https://doi.org/10.22214/ijraset.2021.36522

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 9 Issue VII July 2021- Available at www.ijraset.com*

# Model for Converting PDF to Audio Format (Listen Your Book)

Shailendra Singh[1], Aman Sahu[2], Shubham Keshari[3]
[1]*Assistant Professor,* [2,3]*Final year students, Department of Computer Science and Engineering, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India*

***Abstract:*** ***The present paper has introduced an innovative and efficient technique that enables user to hear the contents of text images instead of reading through them. In the current world, there is a great increase in the utilization of digital technology and multiple methods are available for the people to capture images. such images may contain important textual content that the user may need to edit or store digitally. It merges the concept of Optical Character Recognition (OCR) and Text to Speech Synthesizer (TTS). This can be done using Optical Character Recognition with the use of Tesseract OCR Engine. OCR is a branch of AI that is used in applications to recognize text from scanned documents or images. The analyzed text can also be converted to audio format to help visually impaired people hear the content that they wish to know. Text-to-Speech conversion is a method that scans and reads alphabets and numbers that are in the image using OCR technique and convert it into voices. The aim is to study and compare the multiple methods used for STT conversions and to figure out the most efficient technique that can be adapted for the conversion processes. As a result, based on review study it is found that HMM is a statistical model which is most suitable for TTS conversions.***

***Keywords: Speech to Text, Text to speech, Speech recognition, communication, Hidden Markov Model (HMM)***

## I. INTRODUCTION

This paper will provide an analysis of various methods used for Text-To Speech conversion. Textual information is available in many resources likes document, newspapers, faxes, printed information, written notes, etc. Many people simply scan the document to store the data in the computer. When a document is scanned with a scanner, it is stored in the form of image. But these images are not editable, and it is very difficult to find what the user require as they will have to go through the whole image, reading each line and word to determine if it is relevant to their uses. Images also take up more space than word file in the computer. It is essential to be able to store these contents in such a way so that it becomes easier to search and edit the data. There is a growing demand for application that can recognize characters from scanned documents or captured images and make them editable and easily accessible.

In Text-To-Speech conversion the input text is analyzed and then this text is converted into its audio version to play. This functionality has an effective advantage when a person understands a language but is not fluent with reading and writing in that language and is also useful for the people who are visually impaired as they cannot read it better but understand the message by hearing it. Firstly, the text is prepared for audio conversion by performing pre-processing and texts normalization. To generate the Waveform of the text messages the linguistic analysis and prosodic prediction is done in series.

Artificial intelligence is an area of computer science where a machine is trained to think and behave like intelligent humans. Optical Character Recognition (OCR) is a branch of AI. It is used to detect and extract characters from scanned documents or image and convert them to editable form. Earlier method of OCR used convolutional neural networks, but they are complicated and usually, best suitable for single characters. These methods also had a higher error rate. Tesseract OCR Engine makes use of Long Short-Term Memory (LSTM) which is a part of Recurrent Neural Network. It is open source and is more suitable for handwritten texts. It is also suitable at analyzing larger portion of text data instead of single characters. Tesseract OCR Engine significantly reduces errors which created in the process of character recognition. Tesseract assumes that the input images are binary image and processing takes place step-by-step. The first step is to recognize connected component. Outlines are nested into blobs. These blobs are organized into text lines. Text lines are broken according to the pitching. If there is a fixed pitch between the characters then recognition of text takes place which is a two-pass process. An adaptive classifier is used here. Words that are recognized in the first pass are given to the classifier so that it can learn from the data and use that information for the second pass to recognize the words that were left out in the previous pass. Words that are joined are chopped and words that are broken are recognized with the help of A* algorithm that maintains a priority queue which contains the best suitable characters.

Then, the user can store this information in their computers by saving them in word documents or notepads that can be edited any

©IJRASET: All Rights are Reserved

time they want. It is difficult for visually impaired people to read textual information. Blind people have to make use of Braille to read. It would be easier for them to simply listen to the audio form of the data. This application can be used to convert textual data to audio format so that it is easier for people to hear the information. Google Text-To-Speech API is used to convert the text information into audio form.

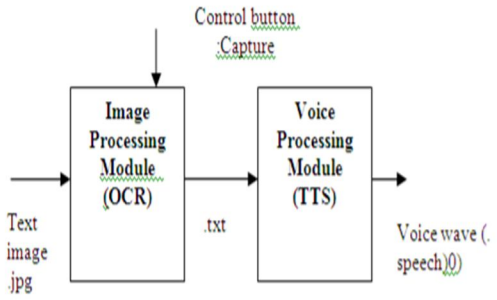
## II. METHODOLOGY

Text-to-speech device consists of two main modules, the image processing module and voice processing modules. Image processing module captures image using camera,

converting the image into text. Voice processing module changes the text into sound and processes it with specific physical characteristics so that the sound can be understood. Figure 2 shows the block diagram of Text To-Speech device, 1st block is image processing module, where OCR converts .jpg to .txt form. 2nd is voice processing module which converts .txt to speech device, 1st block is image processing module, where OCR converts .jpg to .txt form. 2nd is voice processing module which converts .txt to speech. OCR is important element in this module. OCR or Optical Character Recognition is a technology that automatically recognize the character through the optical mechanism, this technology imitates the ability of the human senses of sight, where the camera becomes a replacement for eye and image processing is done in the computer engine as a substitute for the human brain. Tesseract OCR is a type of OCR engine with matrix matching3.

The selection of Tesseract engine is because of its flexibility and extensibility of machines and the fact that many communities are active researchers to develop this OCR engine and also because Tesseract OCR can support 149 languages. In this project we are identifying English alphabets. Before feeding the image to

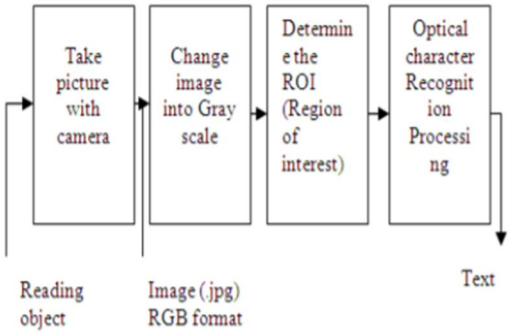
the OCR, it is converted to a binary image to increase the recognition accuracy. Image binary conversion is done by using Imagemagick software, which is another open-source tool for image manipulation. The output of OCR is the text, which is stored in a file (speech.txt). Machines still have defects such as distortion at the edges and dim light effect, so it is still difficult for most OCR engines to get high accuracy text4. It needs some supporting and condition in order to get the minimal defect.




Block diagram of text-to-speech device

### *A. Software Design*

Software processes the input image and converts into text format. The software implementation is showed in Figure below.

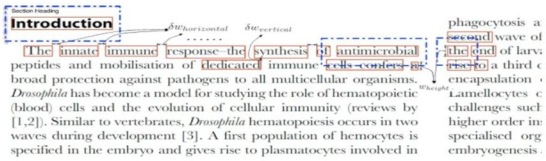



Software design of image processing module.

©IJRASET: All Rights are Reserved

### *B. The Specific Text Selection Module*

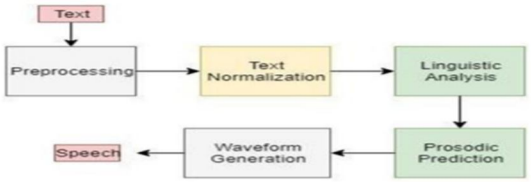
Our paper describes the construction and performance of an open source system that extracts text blocks from PDF-formatted full-text research articles and classifies them into logical units based on rules that characterize specific sections. The LA-PDFText system focuses only on the textual content of the research articles and is meant as a baseline for further experiments into more advanced extraction methods that handle multi-modal content, such as images and graphs. The system works in a three-stage process: (1) Detecting contiguous text blocks using spatial layout processing to locate and identify blocks of contiguous text, (2) Classifying text blocks into rhetorical categories using a rule-based method and (3) Stitching classified text blocks together in the correct order resulting in the extraction of text from section-wise grouped blocks.




Specific text selection

### *C. The Voice Processing Module*

In this module text is converted to speech. The output of OCR is the text, which is stored in a file (speech.txt). Here, Festival software is used to convert the text to speech. Festival is an open-source Text To Speech (TTS) system, which is available in many languages. In this project, English TTS system is used for reading the text.




## III. RESULT

Observed outcome of project:

*1)* Text is extracted from the image and converted to audio.
*2)* It recognizes both capital as well as small letters.
*3)* It recognizes numbers as well.
*4)* Range of reading distance was 38-42cm.
*5)* Character font size should be minimum 12pt.
*6)* Maximum tilt of the text line is 4-5 degree from the vertical.

## IV. CONCLUSION AND FUTURE SCOPE

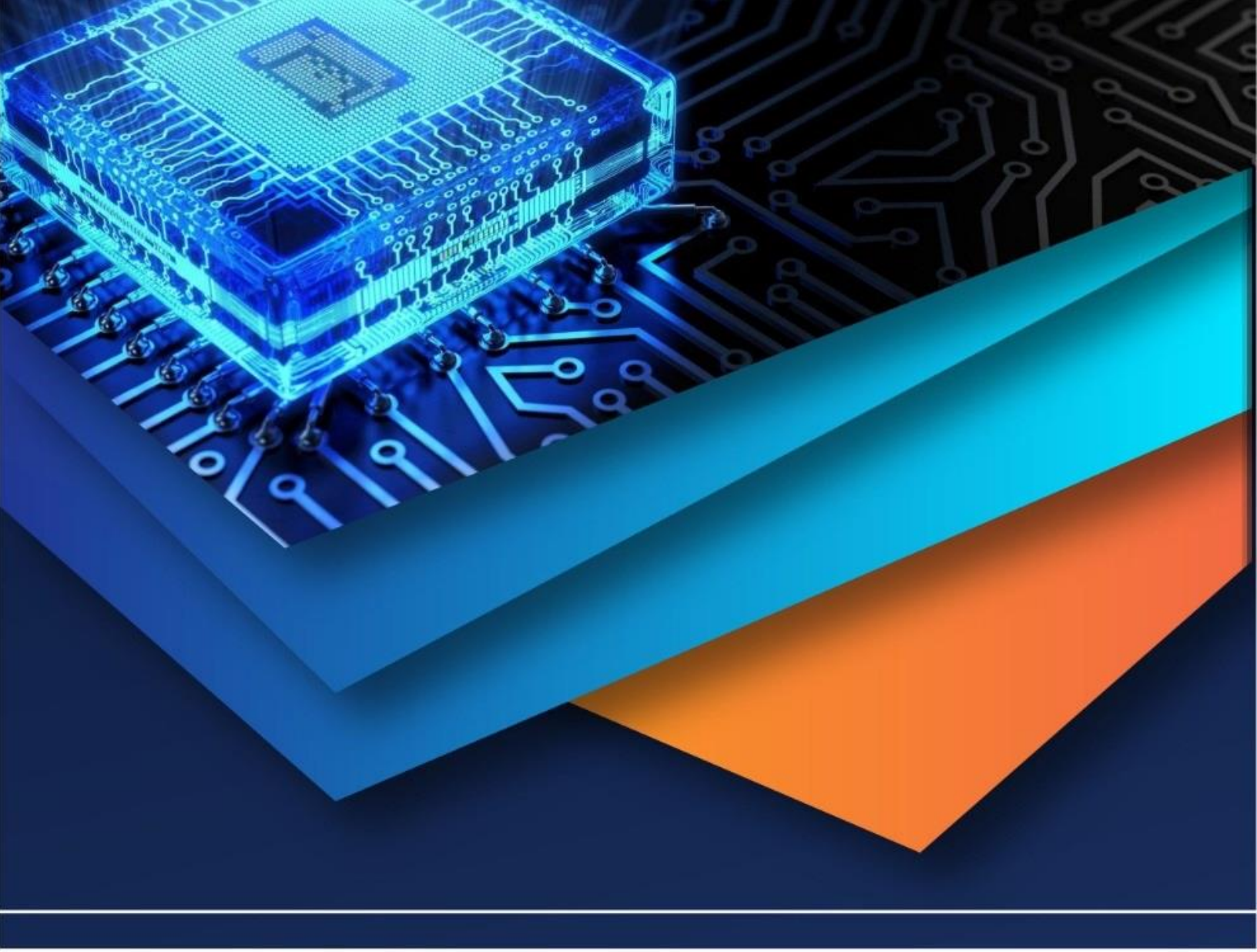
In this age of technology, there is a huge amount of data and it keeps on increasing day by day. Even though much of the data is digital, people still prefer to make use of written transcripts. However, it is necessary to store this data in digital format in computers so that it can be accessed and edited easily by the user. This system can be used for character recognition from scanned documents so that data can be digitalized. Also, the data can be converted to audio form to help visually impaired people obtain the data easily. In the future, we can expand the system to that is can recognize more languages, different fonts and handwritten notes. Various accents can also be added for audio data . In particular, the natural language processing stage of a text-to-speech (TTS) system contains the largest part of the linguistic knowledge for a given language. In this respect, one can state that building a high-quality TTS system for a new language involves many theoretical and technical challenges. Especially, extensive studies must exist (or be done) at the linguistic level, in order to endow the system with the most relevant language information; this requirement represents an essential condition to obtain a true naturalness of the synthesized speech, starting from unrestricted input text. This paper presents fundamental research and the related implementation issues in developing a complete TTS system in Romanian, emphasizing the language particularities and their influence on improving the language processing stage efficiency. The first section describes our standpoint on TTS synthesis as well as the overall architecture of our TTS system. The next sections formulate several important tasks of the natural language processing stage (input text preprocessing, letter-to-phone conversion, acoustic database preparation) and discuss the design philosophy of the corresponding modules, implementation decisions and evaluation experiments.

©IJRASET: All Rights are Reserved

## REFERENCES

[1] Archana A, Shinde D. Text pre-processing and text segmentation for OCR. International Journal of Computer Science Engineering and Technology. 2012:810–12.
[2] Mithe R, Indalkar S, Divekar N. Optical character recognition. International Journal of Recent Technology and Engineering. 2013 Mar; 2(1).
[3] Smith R. An overview of the Tesseract OCR engine, USA: Google Inc; 2007.
[4] Shah H, Shah A. Optical character recognition of Gujarati numerical. International Conference on Signals, Systems and Automation. 2009; 49–53.
[5] Text localization and extraction in images using mathematical morphology and OCR Techniques; 2013.
[6] Vanitha E, Kasarla PK, Kuamarswamy E. Implementation of text- to-speech for real time embedded system using Raspberry Pi processor. International Journal and Magazine of Engineering Technology Management and Research. 2015 Jul:1995
[7] Kumar GS, Krishna MNVLM. Low cost speech recognition system running on Raspberry Pi to support Automation applications. International Journal of Engineering Trends and Technology. 2015; 21(5).
[8] Bhargava A, Nath KV, Sachdeva P, Samel M. Reading assistant for visually Impaired. International Journal of current Engineering and Technology. 2015 Apr; 5(2).
[9] Gomes LCT, Nagle EJ, Chiquito JG. Text-to-speech conversion system for Brazilian Portuguese using a formant-based synthesis technique. LPS-DECOM-FEEC-Unicamp.
[10] Sim Liew Fong, Abdelrahman Osman Elfaki, Md Gapar bin Md Johar & Kevin Loo Tow Aik, Mobile Language Translator, 5th Malaysian Conference in Software Engineering (Misses); 2011. 12. Kamesh DBK, Nazma SK, Sastry JKR, Venkateswarlu S. Camera based text to speech conversion, obstacle and currency detection for blind persons. Indian Journal of Science and Technology. 2016 Aug; 9(30).

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)