Convert PDF to Audiobook using OCR (Optical Character Recognition) and Machine Learning

[1] Mohammad Adil Sheikh,[2] Prof. Deepa Bhattacharya,[3] Sharda Dharawat

[1,2,3] B.I.T. Ballarpur, Gondwana University

Email ID: [1] adil.sheikh97@gmail.com , [2] deepa.ddb@gmail.com, [3] shardadharawat@gmail.com

*****

Abstract: Audiobooks are Favorable for everyone who is always zealous. It is just not possible to buy store them in a bookshelf in your house. Audiobooks can also be a great way to ease your eyes, a rest from the constant charge of digital screens. Yet others can use them to save time. For instance, you can keep up with books and multitask. It can not only solve the problems for millennials but can also be very beneficial tool for visually impaired person. The power to convert any documented file to audiobook is nothing but a pure gift to society. Our technology can be put on work to create such tools. LORO is one of the few contributions we make to our generation.

***Keywords:* -Audiobook, OCR, PDF, Google Cloud Services**

I. INTRODUCTION

LORO can convert any documented PDF to mp3 audio. It uses complex Deep learning algorithms to analyze the pattern of Title to body in a page and presents only the Important content. It has the ability to remove any miscellaneous content from the page, it can remove the iterative author name or Indexes of a file. It has the ability to select where the actual content is starting rather than reading every single bit of the PDF. It is annoying when the audiobook read out every punctuation in between of a very import part of the document, rather more disturbing is to listen to every bit of website mentioned. **LORO** filters out all those stuff for you so you can listen to what you actually intended for. **LORO** uses the very best API like Vision API, Auto ML API, TextToSpeach API from Google Cloud services. So, it delivers product which is best in class. **LORO** can not only converted the written documents into audio files but can also recognize text from a human handwriting, further extending its usability.

II. LITERATURE REVIEW

1) PDF (Portable Document Format)

Stands for "Portable Document Format." PDF is a file format designed to present documents consistently across multiple devices and platforms. A PDF file can store a wide variety of data, including formatted text, vector graphics, and raster images. It also contains page layout information, which defines the location of each item on the page, as well as the size and shape of the pages in the document.

Fig-1: PDF

This information is all saved in a standard format, so the document looks the same, no matter what device or program is used to open it. For example, if you save a PDF on a Mac, it will appear the same way in Windows, Android, and iOS. Adobe Acrobat Export PDF supports optical character recognition, or OCR, when you convert a PDF file to Word (.doc and .docx), Excel (.xlsx), or RTF (rich text format). OCR is the conversion of images of text (scanned text) into editable characters, so that you can search, correct, and copy the text.

2) Machine learning (ML)

[8] ML is a subset of AI, which includes all the approaches that allow machines to learn from data without being explicitly programmed. The intention of ML is to train machines based on the provided data and algorithms.Using the processed data and information, the machines learn how to make decisions.ML is dynamic, meaning that it has the ability to modify itself when exposed to more data. The 'learning' aspect of ML means that the ML algorithms attempt to minimize the errors and maximize the likelihood of their predictions being true. In short, ML is simply a technique to realize AI.

3) OCR (Optical Character Recognition)

[1] Literally, OCR stands for Optical Character Recognition. It is a widespread technology to recognise text inside images, such as scanned documents and photos. OCR technology is used to convert virtually any kind of images containing written text (typed, handwritten or printed) into machine-readable text data. Probably the most well-known use case for OCR is converting printed paper documents into machine-readable text documents. Once a scanned paper document went through OCR processing, the text of the document can be edited with word processors like Microsoft Word or Google Docs.

4) Deep Learning

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans: learn by example. It is the key to voice control in consumer devices like phones, tablets, TVs, and hands-free speakers. Deep learning is getting lots of attention lately and for good reason. It's achieving results that were not possible before.

5) Google Cloud Service

[2] The term "cloud services" refers to a wide range of services delivered on demand to companies and customers over the internet. These services are designed to provide easy, affordable access to applications and resources, without the need for internal infrastructure or hardware. From checking email to collaborating on documents, most employees use cloud services throughout the workday, whether they're aware of it or not and that's true for google cloud services as well. The services offer by Google cloud include compute, storage, big data/analytics, artificial intelligence, and other networking, developer, and management services.

6) Auto ML Table

[7] Automated machine learning (Auto ML) basically involves automating the end-to-end process of applying machine learning to real-world problems that are actually relevant in the industry. In recent years, it has been noticed as well as proven time and time again that ML or machine learning is the key to the future.

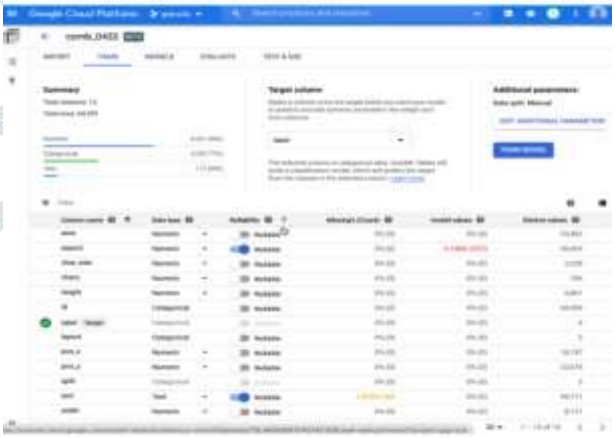

Fig-2: Auto ML Table

Auto ML tools are the need of the hour for data scientists to reduce their workloads in the world where the generation of data is only increasing exponentially. Readily available Auto ML tools make the data science practitioner's work more comfortable and covers necessary foundations needed to create automated machine learning modules. Train models from labelled images and evaluate their performance. Leverage a human labelling service for datasets with unlabelled images. Register trained models for serving through the Auto ML API.

7) Audiobook

Audiobooks are voice recordings of the text of a book that you listen to rather than read. Audiobooks can be exact word-for-word versions of books or abridged versions. You can listen to audiobooks on any smartphone, tablet, computer, home speaker system, or in-car entertainment system. When you purchase or download audiobooks from the internet, they usually come in one of the following audio formats:

- MP3
- WMA (Windows Media Audio)
- AAC (Advanced Audio Coding)

8) API

[3] API stands for Application Programming Interface. In basic terms, APIs are a set of functions and procedures that allow for the creation of applications that access data and features of other applications, services, or operating systems. An API usually is related to a software library. The API describes and prescribes the "expected behaviour" (a specification) while the library is an "actual implementation" of this set of rules.A single API can have multiple implementations (or none, being abstract) in the form of different libraries that share the same programming interface.

9) CNN (CONVOLUTIONAL NEURAL NETWORK)

[5] The term Deep Learning or Deep Neural Network refers to Artificial Neural Networks (ANN) with multi layers. Over the last few decades, it has been considered to be one of the most powerful tools, and has become very popular in the literature as it is able to handle a huge amount of data. The interest in having deeper hidden layers has recently begun to surpass classical methods performance in different fields; especially in pattern recognition. One of the most popular deep neural networks is the Convolutional Neural Network (CNN). It takes this name from mathematical linear operation between matrixes called convolution. CNN have multiple layers; including convolutional layer, non-linearity layer, pooling layer and fully-connected layer. The convolutional and fully-connected layers have parameters but pooling and non-linearity layers don't have parameters.

Fig-3: Convolutional Neural Network

The CNN has an excellent performance in machine learning problems. Specially the applications that deal with image data, such as largest image classification data set (Image Net), computer vision, and in natural language processing (NLP) and the results achieved were very amazing. In this paper we will explain and define all the elements and important issues related to CNN, and how these elements work. In addition, we will also state the parameters that effect CNN efficiency. This paper assumes that the readers have adequate knowledge about both machine learning and artificial neural network.

10) Vision API

Google Cloud's Vision API offers powerful pre-trained machine learning models through REST and RPC APIs. Assign labels to images and quickly classify them into millions of predefined categories. Detect objects and faces, read printed and handwritten text, and build valuable metadata into your image catalog. Google Vision can detect whether you're a cat or a human, as well as the parts of your face. It tries to detect whether you're posed or doing something that wouldn't be okay for Google Safe Search—or not. It even tries to detect if you're happy or sad.

11) Speech Synthesis

[6] Speech synthesis is a process of automatic generation of speech by machines/computers. The goal of speech synthesis is to develop a machine having an intelligible, natural sounding voice for conveying information to a user in a desired accent,

language, and voice.Research in T-T-S is a multi-disciplinary field: from acoustic phonetics (speech production and perception) over morphology (pronunciation) and syntax (parts of speech, grammar), to speech signal processing (synthesis). There are several processing stages in T-T-S system: the text front end analyses and normalizes the incoming text, creates possible pronunciations for each word in context, and generates prosody (emotions, melody, rhythm, intonation) of the sentence to be spoken. For evaluation of T-T-S systems three parameters need to be evaluated: accuracy, intelligibility and naturalness.

III. PROPOSED SYSTEM

1) Methodology

1. Feeding PDF to Vision API to extract Text and details in Json format
2. Converting the Json File to CSV file
3. Labeling every text in CSV file (for training the model)
4. Passing CSV to Auto ML Table
5. Extracting Result and Converting it to audio using TextToSpeech API
6. Downloading the MP3 file.

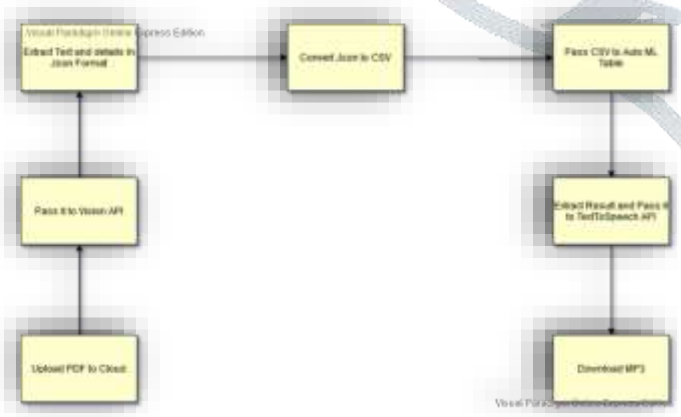
Visual Paradigm Online Express Edition
Extract Text and details in Json Format
Convert Json to CSV
Pass CSV to Auto ML Table
Pass it to Vision API
Extract Result and Pass it to TextToSpeech API
Upload PDF to Cloud
Download MP3
Visual Paradigm Online Express Edition

Fig-5: System Flowchart

2) Features of LORO

1. It can extract text from any documented PDF
2. It can Identify the difference between Body, Title, Index
3. It uses Vision API to extract text from documents
4. It can remove any miscellaneous text like website links, any index reference between the body of text, any punctuations
5. It can identify human written text.
6. It uses Auto ML Table so we can configure the model by our needs
7. It uses Googles TextToSpeech API to convert text into audio
8. It can speak in 4 different accents of English.

IV. CONCLUSION

The paper consists the OCR methodology that assists recognition of text (PDF file) into mp3 audio. OCR is very remarkable technology that hold a lot potential. We aim to develop our model further more to extend its processing for any documented file. We aim to give even picture its voice so it's easy for the children and visually impaired people to under the context easily. Our model will be more user friendly for the mass majority of users. It will be having good knowledge to make voice comfier for the users.

V. REFERENCES

[1]https://techterms.com/definition/ocr-(2018).OCR(optical character recognition).

[2]Gilad,D.M.,(2020).Deep Leaning.https://heartbeat.fritz.ai/optical-character-recognition-using-deep-learning-techniques-1376605b022a

[3]https://cloud.google.com/terms/services-(2020).Google cloud client service.

[4]Jonathan.(2019).API(Application programminginterface). https://www.infoworld.com.

[5] Reference: Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. 2017 International Conference on Engineering and Technology (ICET) PP-01-02.

[6]https://www.ijser.org/researchpaper/A-Comparative-Study-of-Different-Text-to-Speech-Synthesis-Techniques.pdf PP-01-02.

[7] Ted H., Steven P., (2018).Google Cloud Platform for Developers: Google cloud service.

[8] Jakhar, D., & Kaur, I. (2019). Artificial intelligence, machine learning & deep learning: Definitions and differences. Clinical and Experimental Dermatology.