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

[1] **Pritam Ghosh,**[2] **Prof. Deepa Bhattacharya,**[3] **Khushal Sontakke** [1,2,3] B.I.T. Ballarpur, Gondwana University

Email ID: [1] pritamddghosh27@gmail.com , [2] deepa.ddb@gmail.com , [3] khushalsontakke85@gmail.com

Abstract: *Reading stories or essays or any text can be onerous, however an audio reading of the text is convenient and doesn't require as much concentration as reading requires. And here comes the talking book known audiobook.* ***An audiobook allows a person to listen to a recording of the text of the book, rather than read the text of the book. Audiobooks have traditionally been used in schools by teachers of second-language learners, learning-disabled students, and struggling readers or nonreaders. In many cases, audiobooks have proven successful in providing a way for these students to access literature and enjoy books. While these have dramatically risen in popularity over the last few years alone, they're not exactly new. Audiobooks, or "talking books," as they were often referred to in the past, emerged during the 1930s. Unlike the digital versions we enjoy today, however, they often came in the physical form of a cassette tape or vinyl record. LOCUAS it's one of such power to create such taking book. LOCUAS is one of the few contributions to AI community.***

Key words:- Audiobook, OCR, PDF, Google Cloud Services.

I. INTRODUCTION

LOCUAS the audiobook or taking book which converts textual format into mp3 format. LOCUAS uses complex Deep learning algorithms to analyze the pattern of textual content from beginning to end and filters out the important part. It has the propensity to remove any diversified content from the text. It has the propensity to elect where the actual content is starting rather than exiling the PDF. When the taking book read out every punctuation between the text. LOCUAS filters out all this might and delivers the actual audio that you where expecting of. LOCUAS uses the marvelous API like Vision API, AutoML API, text-to-speech. API from Google Cloud Services. Hence LOCUAS reaches the market expectations.

formatter will need to create these components, incorporating the applicable criteria that follow.

II. LITERATURE REVIEW

1. PDF

PDF "Portable Document Format" is a file format designed to present documents often across various devices and platforms. Since 1992 it has become one of the most widely used formats for saving and exchanging documents, developed by Adobe.Maintaining the Integrity of the Specifications

Fig-1: PDF

PDF is a file format that has captured all the elements of a printed documents as an electronic image that you can view, navigate, print, or forward to someone else. PDF files are created using Adobe Acrobat, Acrobat Capture, or similar products. To view and use the files, you need the free Acrobat Reader, which you can easily download. Once

you've downloaded the Reader, it will start automatically

whenever you want to look at a PDF file. PDF files are mainly convenient for documents such as magazine articles, brochures in which you want to conserve the actual graphic outlook online. This file contains one or more page images which one can zoom in or out or can forward or backward the page. To view a PDF, you can use Adobe Reader or any program that supports PDF format. Programs that supports OCR allow you to digitally scan the text file and then update it.

2. *Machine learning (ML)*

Machine learning is shorten to "ML". It's a type of artificial intelligence (AI) that grasp or adjust over time. Instead of following fixed commands coded in a program, ML identifies input patterns and contains algorithms that progress over time. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning commence with observations or data, such as examples, direct experience, or exposure, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary goal is to allow the computers learn automatically without human intercede or hands and adjust actions correspondingly.

Machine learning algorithms are often distinguished as supervised or unsupervised algorithms.

Supervised learning, is the machine learning stint of learning a function that designed an input to an output based on example input-output pairs. Starting from the exploration of a defined training dataset, the algorithm generates an inferred function to make predictions about the output.

On the other hand, whenever the dataset are neither specified nor described *unsupervised learning.* It is an algorithm which deals in an atmosphere where the model generates a method to trace a unknown format from unlabeled data.[4]

3. *Deep Learning*

Deep learning is an artificial intelligence method that imitate the functioning of the human brain in processing data for use in detecting objects, recognizing speech, translating languages, and making decisions. It is a technique that teaches computers to do what comes naturally to humans and grasp from examples. It utilize a hierarchical level of artificial neural networks to carry out the process of machine learning. Deep learning is used across all over industries for a number of different task. Commercial apps that use image recognition such as virtual assistants, self- driving cars, money laundering and many more.[5]

4. *Optical Character Recognition (OCR)*

Optical character recognition or optical character reader (OCR) is a technology that enables users to convert different types of documents such as scanned documents and images into editable and searchable data. It recognize text within the image. OCR is mainly known for converting virtually any kind of text within the image into machine readable text data. It is often used as a invisible technology, powering many well systems and services in daily bases. Once a scanned document went through OCR processing, the text can be edited within word processor like Google Docs and etc.[6]

5. *Google Cloud Service*

The Google cloud service is a suite of public cloud computing service offered by Google. This platform include a wide range of hosted services for compute, storage and application development. Computing and hosting, storage, databases, networking, big data, machine learning are some services provided are Google. These services are designed to provide easy, affordable access to application and resources, without any need of internal software and hardware.[2]

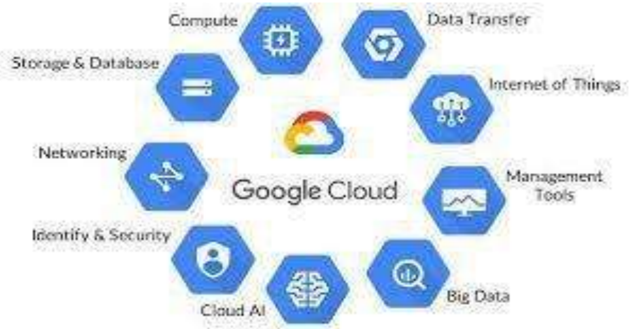

Fig-2: Google Cloud Service

6. *AutoML Table*

For any machine learning engineers, data analyst it's become important to understand the behavior of your trained data and its contribution to the final model and how the model arrived at individual predictions, which helps you to make sure your model is ideal and accurate.

Fig-3: AutoML Table

AutoML Table includes automating end to end process of applying machine learning to real time problems as per industry aspects. As per research its seems again and again that machine learning is the key to the future. This is an era where machine learning is leading from the front through various directions of research, analysis and implementations. AutoML Table provides feature importance, sometimes called feature attributions, which enables you see which feature contributed the most to model training and individual predictions.

7. *Audiobook*

The 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.

Audiobooks are usually purchased and downloaded in the same way as digital music and video. They can also be purchased from online bookstores or downloaded free from public domain sites. Audiobooks come in one of the following audio formats:[1]

. MP3

.WMA (Windows Media Audio)

.AAC (Advanced Audio Coding)

Fig-4: Audiobook

8. *API*

The An Application Programming Interface (API) contains software building tools, subroutine definitions as well as communication protocols that facilitate interaction between systems. An API may be for a database system, operating system, computer hardware or a web-based system.

An Application Programming Interface makes it simpler to use certain technologies to build applications for the programmers. API can include specifications for data structures, variables, routines, object classes, remote calls.[11]

9. *CNN (CONVOLUTIONAL NEURAL NETWORK)*

Convolutional Neural Network is a specialized type of neural network model designed for working with two-dimensional image data, although they can be used with one-dimensional and three-dimensional data. Central to the convolutional neural network is the convolutional layer that gives the network its name. This layer performs an operation called a "convolution".[9] A CNN, is a deep learning neural network designed for structured arrays of data such as images. Convolutional neural networks are widely used in computer vision and have become the state of the art for many visual applications such as image classification. Convolutional neural networks are very good at picking up on patterns in the input image, such as lines, gradients, circles, or even eyes and faces. This property that makes convolutional neural networks so powerful for computer vision. [7]

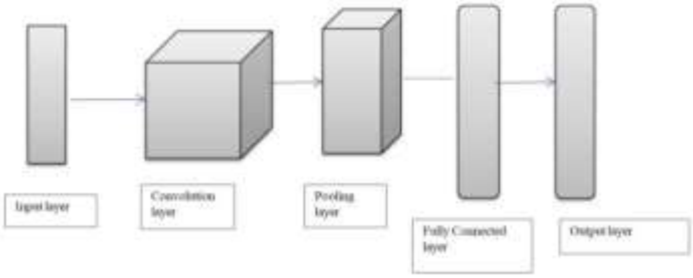

Fig-5: Convolutional Neural Network

Convolutional Neural Networks (CNNs) are traditional ANNs in that they are comprised of neurons that self-optimise through learning. Each neuron will receive an input and perform a operation.[8] It has many features such as simple structure, less training parameters and adaptability. Mostly used in voice analysis and image recognition system. It's weight shared network structure make it more similar to biological neural networks. It reduces the complexity of the network model and the number of weights.[10]

10. *Vision API*

Google have encapsulated their Machine Learning models in an API to allow developers to use their Vision technology. The Vision API can quickly classify images into thousands of categories and assign them sensible labels.[12] It detects individual objects and faces within images, and landmarks and geographical locations, and reads printed words contained within images.[13]

11. *Speech Synthesis*

Artificial production of human speech is known as speech synthesis. This machine learning-based technique is applicable in text-to-speech, music generation systems.[14] The Text-To-Speech (TTS) Synthesis consists of two main phases. The first is analyze the input or word text and then it transcribed into phonetic form or some other Linguistic representation and the second one is the generation of speech wave forms where the acoustic output is produced from this phonetic and prosodic information. These two phases are called as high and low level synthesis.

The input text might be a word processor, a mobile text-message, or scanned text from a newspaper. The character string is then pre-processed and analyzed into phonetic form, which is done with the string of phonemes. Speech sound is finally generated with the low-level synthesizer by the information from high-level one.[15] And The quality of a speech synthesizer is judged by its similarity to the human voice and by its ability to be understood.

A text-to-speech (TTS) system has two parts: front-end and a back-end. First, it converts raw text that is symbols like numbers and hand written words. This called as text normalization. The front-end assigns phonetic transcripted word and marks the text into prosodic units. This process called text-to-phoneme conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation. The back-end refered as synthesizer then converts the symbolic linguistic representation into sound.

III. PROPOSED SYSTEM

1. *Methodology*

□ Upload PDF file to Cloud.

□ Feeding PDF to Vision API to extract Text and details in JSON format.

□ Convert JSON file into CSV file.

□ Label every text in CSV file for training the model.

□ Send CSV file to AutoML Table.

□ Extract result and converting into audio file

□ Listenable audio file.

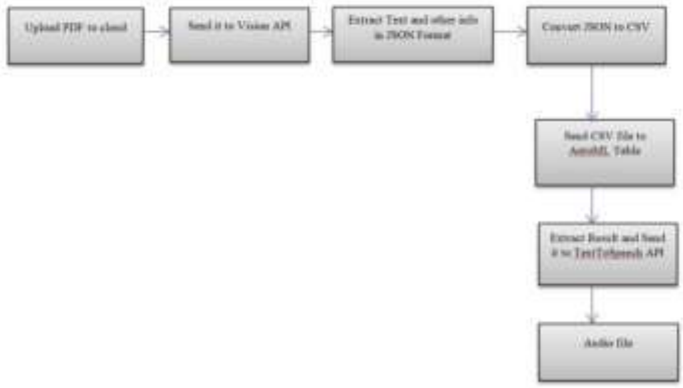
Upload PDF to cloud
Send it to Vision API
Extract Text and other info in JSON Format
Convert JSON to CSV
Send CSV file to AutoML Table
Extract Result and Send it to TextToSpeech API
Audio file

Fig-6: System Flowchart

2. *Feature of Locuaz*

□ It can extract text from any PDF document.

□ It can identify the difference between Body, Title And Index.

□ It used Vision API to extract text from document.

□ It can remove unnecessary text like hyperlinks, punctuation marks and index reference between body and text.

IV.CONCLUSION

The paper consists the OCR techniques that assists recognition of text (PDF file) into audible format. OCR an amazing technology that holds a lots of potential. We focus to develop our tool further more to extend its processing for any documented file. Our tool will be more user friendly for the mass majority of users. It will be having good knowledge to make voice snug for the users.

V. REFERENCES

[1]https://www.lifewire.com/what-are-audiobooks-243853

[2]https://searchcloudcomputing.techtarget.com/definition/Google-Cloud-Platform

[3]https://daleonai.com/pdf-to-audiobook

[4]https://techterms.com/definition/machine_learning

[5]https://www.investopedia.com/terms/d/deep-learning.asp#:~:text=DeeP%20learning%20 iS %20AN%20artificial,for%20use%20in %20decision%20making.&text=AlSO%20knowN%20aS%20 DeeP%20neural%20learning%20or%20deeP%20neural%20network.

[6]https://docparser.com/blog/what-is-ocr/

[7]https://deepai.org/machine-learning-glossary-and-terms/convolutional-neural-network

[8]https://www.researchgate.net/publication/285164623_An_I ntroduction_to_Convolutional_Neural_Networks

[9]https://machinelearningmastery.com/convolutional-layers-for-deep-learning-neural-networks/

[10]https://arxiv.org/ftp/arxiv/papers/1506/1506.01195.pdf

[11]https://www.tutorialspoint.com/application-programming-interface-api

[12]https://www.datacamp.com/community/tutorials/beginne r-guide-google-vision-api

[13]https://medium.com/@annycarolinegnr/using-google-vision-api 22d1fdb755d8#:~:text=The%20Vision%20API%20is%20 a,such%20as%20nudity%20and%20violence

[14]https://heartbeat.fritz.ai/a-2019-guide-to-speech-synthesis-with-deep-learning-630afcafb9dd

[15]https://core.ac.uk/download/pdf/83592918.pdf