# Bangla PDF Speaker : A Complete Computer Application to Convert Bangla PDF to Speech

**Conference Paper** · July 2021

DOI: 10.1109/ACMI53878.2021.9528221

Md Mizanur Rahaman Nayan
Bangladesh University of Engineering and Technology
**3** PUBLICATIONS **4** CITATIONS
SEE PROFILE

Mohammad Haque
Concordia University Montreal
**32** PUBLICATIONS **392** CITATIONS
SEE PROFILE

2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 8-9 July 2021, Rajshahi, Bangladesh

# Bangla PDF Speaker : A Complete Computer Application to Convert Bangla PDF to Speech

1st Md. Mizanur Rahaman Nayan
*Department of EEE*
*BUET*
Dhaka, Bangladesh
mdmizanur940@gmail.com

2nd Mohammad Ariful Haque
*Department of EEE*
*BUET*
Dhaka, Bangladesh
arifulhoque@eee.buet.ac.bd

***Abstract*—In this paper, a complete computer application is presented that can convert Bangla PDF to Bangla Speech. According to the proposed technique, images are extracted from PDF and then after processing the images, they are sent to OCR engine to extract text. Extracted text are then normalized and sent to text to speech (TTS) engine to generate speech. Image processing is a key component of the developed application as it increases the efficiency of OCR engine to a great extent. We propose a novel threshold selection method that is able to detect type of noise in the extracted image and select threshold accordingly for binary transformation. Thus it solves the problem of selecting appropriate threshold of different images and it increases the overall accuracy and efficiency of the application. Another feature that has improved the performance of introduced computer application is text normalization. Normalization of the extracted text from the OCR engine makes the text more accurate to pronounce by the TTS engine depending on the context. Finally, we present experimental results that show 80.804% accuracy on text extraction from the PDF file and 3.92 score (out of 5) on the generated speech by human evaluation.**

***Index Terms*—PDF to speech, image processing, Autonomous binarization threshold selection, text extraction, text normalization, text to speech**

## I. INTRODUCTION

Portable document format (PDF) to speech conversion tool are available in different languages. In PDF, as data is stored as content instead of text format (UNICODE3) [1], directly extracting text from the PDF file is not possible. We need to use optical character recognition (OCR) to extract text from the image, which is extracted from PDF. High performance OCR [2] can effectively extract text from an image, which is quite useful for PDF to speech conversion. Holley2009 [3] describes various techniques to obtain high accuracy from an OCR engine.

Many initiative has been taken for Bangla text extraction [4] [5] [6], conversion of text to speech [7] and even PDF to speech [8]. But there is no such complete application that can convert Bangla PDF to speech with high accuracy. In this work, a complete computer application is developed integrating all the necessary components for Bangla PDF to speech conversion that includes image extraction, image processing, Bangla text extraction, text normalization and finally text to speech conversion. We have focused on image processing part and text normalization part which has made the application more efficient. A novel method of adaptive binarization threshold selection has been used that automatically detects the binarzation threshold depending on the image noise quality. Other basic image processing method like grey scaling, contrast scaling etc has been used. For text normalization, insertion error has been reduced and other basic normalization method has been used. We have used Google's OCR engine pytesseract [9] in this application, which has been tuned for Bangla language with Bengali dataset to extract Bangla text. For text to speech conversion, gTTS engine [10] has been used.

Finally, a graphical user interface (GUI) has been developed to make the application more user friendly. The application will be very helpful to listen any PDF instead of reading it manually. Accuracy measuring of PDF to text conversion has been defined according to the method referred in Morris and Andrew's paper [11]. Human evaluation has also been done to evaluate quality of the generated speech.

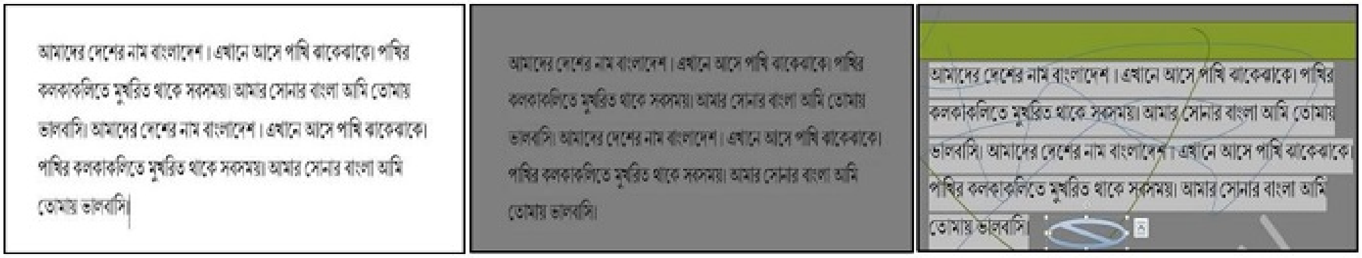
## II. METHODOLOGY

The developed computer application for Bangla PDF to speech conversion consists of five major processing blocks as shown in Fig. 2. In what follows we describe each of these components in more detail.

### A. Image Extraction

Images of the selected pages that a user want to listen are extracted first from the PDF file. Inputs are taken through a GUI.Then the images are extracted from the PDF file as follows:

978-1-6654-3843-8/21/ $31.00 ©2021 IEEE



আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির
কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায়
ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে।
পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি
তোমায় ভালবাসি।

আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির
কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায়
ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে।
পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি
তোমায় ভালবাসি।

আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির
কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায়
ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে।
পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি
তোমায় ভালবাসি।

Fig. 1. Images with different noises: low noise (leftmost), moderate noise (middle), high noise (rightmost)

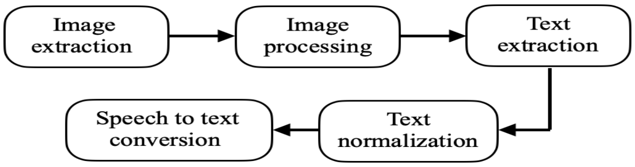


Fig. 2. Block diagram of the complete system

*1) Taking Input:* Fig. 3 shows the GUI where three inputs are taken. They are: PDF file address, starting and ending page number.

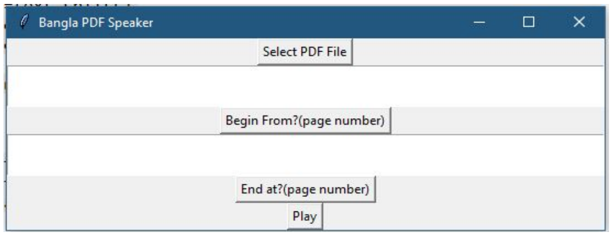


Fig. 3. GUI of the Bangla PDF Speaker

*2) Extracting images of selected pages:* Images of specified pages are extracted using a python library named PDF2image [12] and subsequently sent to the OCR engine after image processing.

## *B. Image Processing*

Image processing is the most important part of the program since efficiency of the OCR engine depends largely on the quality of processed image. Moreover, properly filtered and noise cancelled image improves the accuracy of image to text conversion. The image processing step consists of four operations as shown in Fig.4.

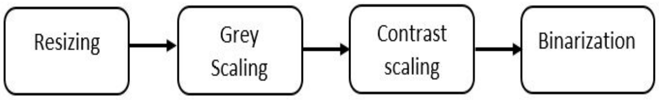


Fig. 4. Image Processing Steps

*1) Resizing:* The extracted images of the specified pages are resized to (900,1300) pixels. Image of this size provides better accuracy than others.

*2) Grey Scaling:* Grey scaling of image is necessary to feed image to the OCR engine. Normally images are in RGB format which has three frames per image. Grey scaling increases the efficiency of the OCR engine to a large scale [13].

*3) Contrast Scaling:* Contrast Scaling is a very helpful tool. Image with poor contrast produces very low OCR performance [14]. Contrast scaling also improves the performance of subsequent binarization operation.

*4) Binarization:* Binarization is a image processing tool where pixel values lower than a certain threshold is converted to 0 and pixel values higher than the threshold is converted to 255. As a result image noise can be overcome that improves the OCR performance. Thresholding is a common image processing operation applied to gray-scale images to obtain binary or multilevel images. O'Gorman, Lawrence [15] has shown how thresholding helps to increase image readability. But different image requires different threshold depending on the image noise. A popular method for adaptive threshold selection for a wide variety of applications is Otsu's method [16]. The method exhibits relatively good performance if the histogram of the image has bimodal distribution and contains a deep and sharp valley between the two peaks [17]. But if the object area is small compared to the background area, the histogram no longer exhibits bimodality [18] and the performance deteriorates. Therefore, we have proposed a novel threshold selection technique for our application. Fig. 5 shows the block diagram of the proposed technique.

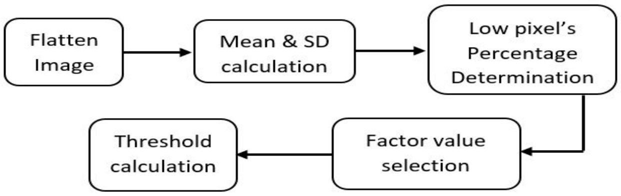


Fig. 5. Block diagram of the threshold selection algorithm

As shown in the figure, the image is first flattened, and standard deviation ($SD$) and $mean$ of pixel values are measured. $SD$ and $mean$ are good indicators of noise present in the image. For example, an ideal image with very low noise containing texts only have pixels with value less than $(mean-SD)$ and pixels with value greater than $(mean + SD)$. So,

TABLE I
RESULTS

| Dataset PDF No | Without image Processing & without text normalization | | | Without text normalization and with image processing | | | Without image Processing & with text normalization | | | Final Test (with both Image processing & Text normalization) | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | identically detected words(%) | detected minored changed(%) | overall accuracy(%) | identically detected words(%) | detected minored changed(%) | overall accuracy(%) | identically detected words(%) | detected minored changed(%) | overall accuracy(%) | identically detected words(%) | detected minored changed(%) | overall accuracy(%) |
| 1 | 54.84 | 9.09 | 63.93 | 56.04 | 11.78 | 67.82 | 56.89 | 10.85 | 67.74 | 61.33 | 19.06 | 80.39 |
| 2 | 58.95 | 8.8 | 67.75 | 60.88 | 11.95 | 72.83 | 63.08 | 11.63 | 74.71 | 63.08 | 16.81 | 79.89 |
| 3 | 47.9 | 9.95 | 57.85 | 53.47 | 10.44 | 63.91 | 49.83 | 10.48 | 60.31 | 63.49 | 18.28 | 81.77 |
| 4 | 56.92 | 8.79 | 65.71 | 55.83 | 12.31 | 68.14 | 59.82 | 12.07 | 71.89 | 61.76 | 15.95 | 77.71 |
| 5 | 60.6 | 10.49 | 71.09 | 59.91 | 10.49 | 70.4 | 64.5 | 13.12 | 77.62 | 64.5 | 19.76 | 84.26 |
| Average | 55.842 | 9.424 | 65.266 | 57.226 | 11.394 | 68.62 | 58.824 | 11.63 | 70.454 | 62.832 | 17.972 | 80.804 |

pixels which have values in the range $(mean - SD) < Pixel\ value < (mean + SD)$ can be considered as non-significant. From this observation, a threshold ($Th$) can be calculated as,

$$Th = mean - f \times SD$$

where, $f$ is a factor and it is selected depending on low pixels percentage (LPP) of the image as

$$\begin{aligned} f &= \frac{LPP}{20} \\ LPP &= \frac{\#\ pixels\ with\ values < (mean - SD)}{\#\ total\ pixels} \times 100 \end{aligned}$$

Therefore, if $LPP$ is greater than 20%, $f$ is selected higher than 1 considering the image is very noisy. Again for low LPP, $f$ is selected lower than 1 considering low noise. As a result, the threshold for low noisy image is higher than threshold for high noisy image. In Fig. 1, three images of different noise has been shown. According to the proposed method, the calculated threshold values are 219.89, 112.48, 88.89 respectively. The lower threshold for a more noisy image helps better noise removal as shown in the experiment section. After obtaining the threshold, image is binarized and thus image get ready for OCR engine.

### C. Text Extraction

Pytesseract [9] engine have been used to extract text from image. Pytesseract or Python-tesseract is a wrapper for Google's Tesseract-OCR engine. A Bengali language image data set has been used to adapt the OCR engine for Bengali text extraction. The engine can extract text from a noise-free image with very high accuracy. But for a noisy image, the accuracy falls tremendously.

### D. Text Normalization

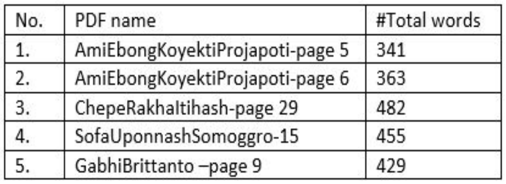
Text normalization is a basic text processing that transforms the written form of text into appropriate spoken form. This is a necessary step before sending the text to a TTS engine. Without text normalization the generated speech will be unnatural (e.g if there is a date write on as ”12/12/12 then text to speech engine will say ”twelve twelve twelve” instead ”twelve December two thousand twelve”) [19]. We have used some rule based text normalization to overcome this challenge. Again, extracted text has some noise like unwanted character. They are defined as insertion error. Insertion error produce unwanted speech which decrease the quality of speech. Although Google text to speech engine has some internal text normalization, some additional text normalization has been done to increase the accuracy. Specially, insertion error has been removed by our text normalization.

### E. Text to Speech Conversion

There are several speech synthesis systems such as Festival [20], Google text to speech engine (GTTS) [10] for text to speech conversion. We have used GTTS has in our application due to its efficiency and better Bengali pronunciation. The engine takes text input and produces an mp3 file containing the generated speech of the input text.

## III. EXPERIMENT AND RESULTS

In this section, we present experimental results to evaluate the overall performance and accuracy of the developed application. Five PDF books have been used as test data for these experiments as shown in Table 1. Overall accuracy is measured using word error rate (WER) [21] as described in Andrew & Morris's work [11] and minor changes of words are detected as in [22]



| No. | PDF name | #Total words |
|---|---|---|
| 1. | AmiEbongKoyektiProjapoti-page 5 | 341 |
| 2. | AmiEbongKoyektiProjapoti-page 6 | 363 |
| 3. | ChepeRakhaltihash-page 29 | 482 |
| 4. | SofaUponnashSomoggro-15 | 455 |
| 5. | GabhiBrittanto –page 9 | 429 |

Fig. 6. Data-set details

Before presenting the results, we show a sample generated speech waveform along with its spectrogram in Fig. 7. It is seen that a realistic speech spectogram is obtained from the developed computer application. In what follows we present experimental results along with some illustrative examples.

### A. The necessity of image processing

Fig. 8 shows two binarized images based on Otsu's method and our proposed adaptive method. Original image is shown in Fig. 1 (rightmost). It is clear that Otsu's thresholding method produce a noisy image, whereas the proposed method is able to reduce the noise to a large extent.

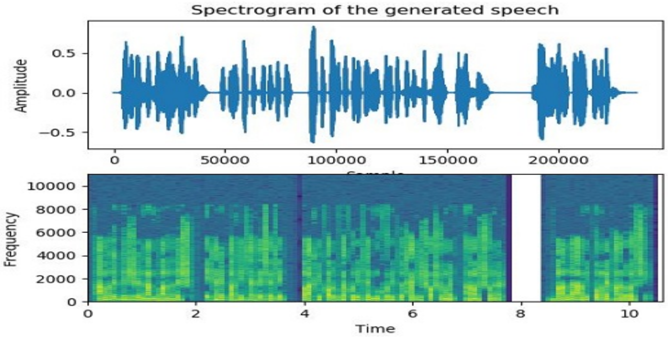
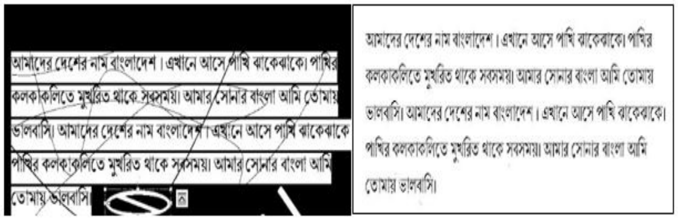


Fig. 7. Waveform and spectrogram of generated speech



আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির
কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায়
ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে
পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি
তোমায় ভালবাসি।

আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির
কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায়
ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে।
পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি
তোমায় ভালবাসি।

Fig. 8. Binarized image using Otsu's method (left) vs proposed threshold selection method (right)

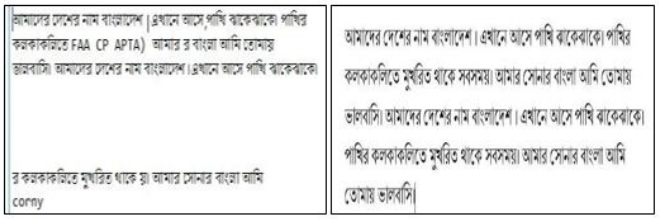
The impact of image processing on the performance of the OCR engine has been shown In Fig. 9, the left image is of extracted text without any image processing and the right one is the image of extracted text after image processing. The source image is the same as Fig. 8. It is clear that without image processing, the indentation has broken down and the text contains some gibberish characters, all of which are removed due image processing operations.

### B. The necessity of text normalization

Fig. 10 compares the raw OCR output and the OCR output after text normalization. It is seen that the raw text contains insertion error, which has been removed through text normalization.

### C. Accuracy of the application

The accuracy of text extraction has been evaluated while keeping both image processing and text normalization function included and excluded. The results are presented in Table 1. We see that an average of 65.71% accuracy is obtained if no image processing and text normalization are performed. Inclusion of image processing step improves the average accuracy to 68.62%. If we perform text normalization but no image processing, the average accuracy is 70.45%. Finally, using both text normalization and image processing, a higher average accuracy of 80.804% has been obtained.

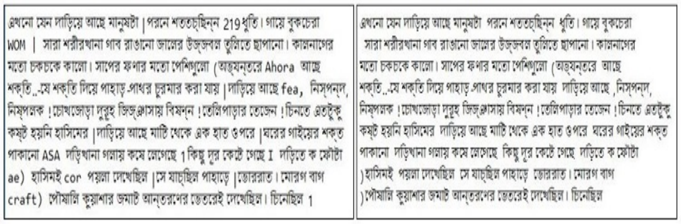


আমাসের সেশের নাম বাংলাসেশ |এখানে আসে,পাখি ঝাকেঝাকে। পাখির
কলকাকলিতে FAA CP APTA) আমার র বাংলা আমি তোমায়
ভালবাসি। আমাসের সেশের নাম বাংলাসেশ।এখানে আসে পাখি ঝাকেঝাকে।

র কলকাকলিতে মুখরিত থাকে য়। আমার সোনার বাংলা আমি
corny

আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির
কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায়
ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে।
পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি
তোমায় ভালবাসি।

Fig. 9. Visualization of image processing effect on OCR



এখনো যেন দাড়িয়ে আছে মানুষটা |পরনে শততচ্ছিন্ন 219 ধুতি। গায়ে বুকচেরা
WOM | সারা শরীরখানা গাব রাঙানো জালের উজ্জ্বল তুলিতে ছাপানো। কালনাগের
মতো চকচকে কালো। সাপের ফণার মতো পেশিগুলো (অভ্যন্তরে Ahora আছে
শক্তি..যে শক্তি দিয়ে পাহাড়-পাথর চুরমার করা যায় |দাড়িয়ে আছে fea, নিস্পন্দ,
নিষ্পলক !চোখজোড়া দুরূহ জিজ্ঞাসায় বিষণ্ন !তেলিপাড়ার তেজেন !চিনতে এতটুকু
কষ্ট হয়নি হাসিমের |দাড়িয়ে আছে মাটি থেকে এক হাত ওপরে |ঘরের গাইয়ের শক্ত
পাকানো ASA দড়িখানা গলায় কসে লেগেছে 1 কিছু দূর কেটে গেছে I দড়িতে ক স্ফোটা
ae) হাসিমই cor পয়লা দেখেছিল |সে যাচ্ছিল পাহাড়ে |ভোররাত। মোরগ বাগ
craft) পৌষালি কুয়াশার জমাট আন্তরণের ভেতরেই দেখেছিল। চিনেছিল 1

এখনো যেন দাড়িয়ে আছে মানুষটা পরনে শততচ্ছিন্ন ধুতি। গায়ে বুকচেরা
সারা শরীরখানা গাব রাঙানো জালের উজ্জ্বল তুলিতে ছাপানো। কালনাগের
মতো চকচকে কালো। সাপের ফণার মতো পেশিগুলো (অভ্যন্তরে আছে
শক্তি..যে শক্তি দিয়ে পাহাড়-পাথর চুরমার করা যায় দাড়িয়ে আছে ,নিস্পন্দ,
নিষ্পলক !চোখজোড়া দুরূহ জিজ্ঞাসায় বিষণ্ন !তেলিপাড়ার তেজেন !চিনতে এতটুকু
কষ্ট হয়নি হাসিমের দাড়িয়ে আছে মাটি থেকে এক হাত ওপরে ঘরের গাইয়ের শক্ত
পাকানো দড়িখানা গলায় কসে লেগেছে কিছু দূর কেটে গেছে দড়িতে ক স্ফোটা
)হাসিমই পয়লা দেখেছিল সে যাচ্ছিল পাহাড়ে ভোররাত। মোরগ বাগ
)পৌষালি কুয়াশার জমাট আন্তরণের ভেতরেই দেখেছিল। চিনেছিল

Fig. 10. Text normalization :left image shows extracted text without normalization and right one is after normalization

### D. Human evaluation

A survey has been conducted, where three individual generated speech and human generated ground-truth speech have been attached. The users are asked to compare the quality of the generated speech with respect to human generated speech and evaluate them out of 5. A total of 21 people responded and an average score of 3.92 has been obtained.

## IV. Conclusion

In this project, an effort to develop a computer application has been made, which can help to read a Bangla PDF file. To increase the accuracy, attempts like image processing, automated threshold selection and text normalization are made, and we obtained an overall accuracy of 80%. The application can be made more efficient by training a Bangla OCR engine from scratch. Text normalization can also be improved with the help of neural network based techniques. Development of TTS for Bangla voice will also help to improve speech quality.

## References

[1] Deepak Massand. System and method for reflowing content in a structured portable document format (pdf) file, September 30 2010. US Patent App. 12/413,486.

[2] Thomas M Breuel, Adnan Ul-Hasan, Mayce Ali Al-Azawi, and Faisal Shafait. High-performance ocr for printed english and fraktur using lstm networks. In *2013 12th International Conference on Document Analysis and Recognition*, pages 683–687. IEEE, 2013.

[3] Rose Holley. How good can it get? analysing and improving ocr accuracy in large scale historic newspaper digitisation programs. *D-Lib Magazine*, 15(3/4), 2009.

[4] Umapada Pal, Partha Pratim Roy, Nilamadhaba Tripathy, and Josep Lladós. Multi-oriented bangla and devnagari text recognition. *Pattern Recognition*, 43(12):4124–4136, 2010.

[5] Supriya Kurlekar. Reading device for blind people using python, ocr and gtts.

[6] Umapada Pal and BB Chaudhuri. Ocr in bangla: an indo-bangladeshi language. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, volume 2, pages 269–273. IEEE, 1994.

[7] Shaheena Sultana, MAH Akhand, Prodip Kumer Das, and MM Hafizur Rahman. Bangla speech-to-text conversion using sapi. In *2012 International Conference on Computer and Communication Engineering (ICCCE)*, pages 385–390. IEEE, 2012.

[8] Md Rafiqul Islam, Ram Shanker Saha, and Ashif Rubayat Hossain. Automatic reading from bangla pdf document using rule based concatenative synthesis. In *2009 International Conference on Signal Processing Systems*, pages 521–525. IEEE, 2009.

[9] Pytesseract , ocr engine. https://pypi.org/project/pytesseract/, 2020.

[10] Google text to speech engine. https://pypi.org/project/gTTS/, 2020.

[11] Andrew Cameron Morris, Viktoria Maier, and Phil Green. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*, 2004.

[12] Pdf2image. https://pypi.org/project/pdf2image/, 2020.

[13] Tracy Powell and Gordon Paynter. Going grey? comparing the ocr accuracy levels of bitonal and greyscale images. *D-Lib Magazine*, 15(3/4), 2009.

[14] Mande Shen and Hansheng Lei. Improving ocr performance with background image elimination. In *2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 1566–1570. IEEE, 2015.

[15] Lawrence O'Gorman. Binarization and multithresholding of document images using connectivity. *CVGIP: Graphical Models and Image Processing*, 56(6):494–506, 1994.

[16] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[17] Otsu's method limitation,wikipedia. https://en.wikipedia.org/wiki/Otsu\%27s_method, 2020.

[18] Josef Kittler and John Illingworth. On threshold selection using clustering criteria. *IEEE transactions on systems, man, and cybernetics*, (5):652–655, 1985.

[19] Firoj Alam, SM Habib, and Mumit Khan. Text normalization system for bangla. Technical report, BRAC University, 2008.

[20] Firoj Alam, Promila Kanti Nath, and Mumit Khan. Text to speech for bangla language using festival. Technical report, BRAC University, 2007.

[21] Word error rate. https://en.m.wikipedia.org/wiki/Word_error_rate, 2020.

[22] Minor changed word detector. https://copyleaks.com/text-compare, 2020.