Bangla PDF Speaker : A Complete Computer Application to convert Bangla PDF to Voice

1st Md. Mizanur Rahaman Nayan
EEE department
BUET
Dhaka, Bangladesh
mdmizanur940@gmail.com
s

***Abstract*—In pdf data is stored as content so we lost our text form which is in Unicode3. [1] .that is why extracting text from pdf directly is not possible. So, for extracting text from a PDting threshold for binarization which increases efficiency highly [2] So, after taking image from pdf page we do some image processing and and make it more useful as an input for pytesseract engine. Then pytesseract engine extract text from the image. We store this text in a tuple .At the time of tuning the program, We have measured the accuracy for the test images and improved the parameter value that we used in image processing . When we found the accuracy is high enough then we used the parameter for general images which are automatically generated from pdf file. After extracting text using OCR ,we normalize the texts. Normalization increases words readability and spoken sentence quality [3].After normalization we send the text to the text to speech engine ,gTTS. This engine convert text to speech that we hear.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Bangla text extraction and text to to voice conversion has become very popular now a days. Many initiative has been taken for Bangla text extraction [4] [5] [6] ,conversion of text to speech [7] and even pdf to voice [8]. But there is no such complete application that can convert Bangla PDF to voice with high accuracy. Extracting feature from different types of pdf file get's error prone if we don't process and normalize the extracted text from the pdf files properly. In this paper , We are presenting Bangla PDF speaker .It is a computer application that can read through a pdf containing a Bangla Writings . Basically, this application is a cascaded version of four big area of research. They are as follows: Image Processing, Image to Bangla Text , Text normalization and finally text to speech .We have focused on image processing part and text normalization part which has made the application more efficient. This application will be very helpful to listen any PDF instead of readings. English pdf speaker is available, but there is no such application that can convert bangla pdf to voice with more than mean 94% accuracy

II. Methodology

A. Taking Input

Our program will take a pdf file as input and range of page number that the user want to hear. I have used input() function

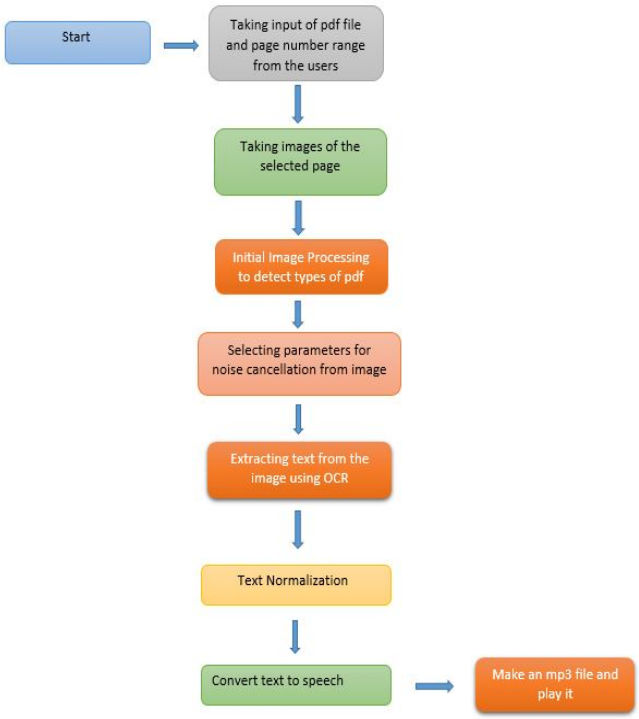


Fig. 1. Flow-Chart of the project

for taking input of the pdf file location as string and range of page number as string. Then I converted page number into integer from string for feeding it in next step.

B. Converting PDF's selected page to images:

I have used a python lib pdf2image that convert pdf pages to images of desired size.

C. Image Processing

Image processing is the most important part of our program since efficiency of the program depends largely on the quality of processed image. If we don't give properly filtered and noise cancelled image we won't get good accuracy. We will explain how accuracy falls if we don't use image processing properly .We will complete image processing in two steps.

First step is to detect the quality of image which measures mean of standard deviation of each pixel value and mean value of pixels . In second step, depending on the value of mean standard deviation and mean pixel value we will set parameters for next level image processing where we will grey-scale and binarize our image finally and will send to pytesseract engine to extract text from the image .We will set a threshold to binarize . Thresholding is a common image processing operation applied to gray-scale images to obtain binary or multilevel images.O'Gorman, Lawrence show how thresholding helps to increase image readability [9]

1) Image Pre-processing to detect type of noise in image: We convert images in grey-scale. After that we calculate amount of noise by calculating standard deviation and mean of pixel .We found if mean standard deviation of pixel values is very high then we can identify the image of low noise since we expect image only contain bangla text. So, by measuring mean standard deviation we can measure the quality of noise of an image . We classify different type of pdf file depending on the mean standard deviation value and mean of pixel value. Where high standard deviation with high mean means high quality of image hence good pdf and with low mean standard deviation means low quality of image hence low quality of pdf .We will now explain how by measuring mean and standard deviation of pixel value we can measure image noise quantity.

Fig. 2. Image Pre-processing

To do this at first we take image as an array and then we flatten it to give it input to the numpy.std () function. Then we use numpy.std() and numpy.mean() function to measure standard deviation and mean of pixel values.We observed for several picture but here we are summarizing by presenting few image and their mean and standard deviation showing the noise measurement.(Here noise is considered as very close pixel value to the pixel value of writings)

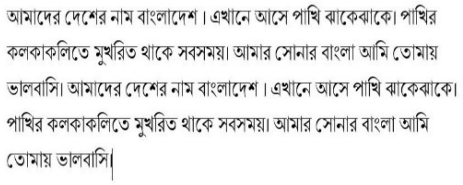

আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায় ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায় ভালবাসি।

Fig. 3. Image with very low noise

average	float64	1	240.2959423517041
im	Image	(951, 340)	<JpegImageFile @ 0x29AADBE9FD0> …
im_arr	uint8	(340, 951, 3)	[[[255 255 255] [255 255 255]
im_flatten	uint8	(970020,)	[255 255 255 ... 255 255 255]
standard_deviation	float64	1	53.15424899218377

Fig. 4. Standard deviation and mean of Image in figure 2

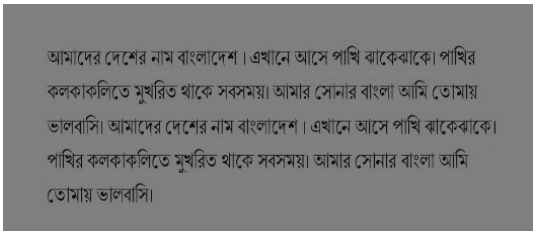

আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায় ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায় ভালবাসি।

Fig. 5. Image with medium noise

average	float64	1	121.2318348580141
im	Image	(954, 352)	<JpegImageFile @ 0x29AADBE9630> …
im_arr	uint8	(352, 954, 3)	[[[127 127 127] [127 127 127]
im_flatten	uint8	(1007424,)	[127 127 127 ... 127 127 127]
standard_deviation	float64	1	23.825888805954975

Fig. 6. Standard deviation and mean of Image in figure 4

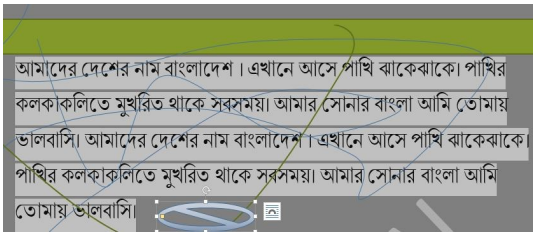

আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায় ভালবাসি। আমাদের দেশের নাম বাংলাদেশ। এখানে আসে পাখি ঝাকেঝাকে। পাখির কলকাকলিতে মুখরিত থাকে সবসময়। আমার সোনার বাংলা আমি তোমায় ভালবাসি।

Fig. 7. Image with high noise

average	float64	1	149.56073189509723
im	Image	(833, 342)	<JpegImageFile @ 0x29AA96574E0> …
im_arr	uint8	(342, 833, 3)	[[[127 127 127] [127 127 127]
im_flatten	uint8	(854658,)	[127 127 127 ... 127 127 127]
standard_deviation	float64	1	51.09475578476383

Fig. 8. Standard deviation and mean of Image in figure 6

From our observation on more than 50 different types of noisy and clean image we classified them as follows on basis of of standard deviation and mean value of pixels:

Standard Deviation	Mean	Type
High(> 40)	High(> 200)	Very small noise(Standard pdf)
High(> 40)	Low(< 170)	Very noisy
Low(< 40)	High(> 200)	Noisy
Low(< 40)	Low(< 170)	Medium Noisy

TABLE I
CLASSIFICATION OF IMAGE WITH RESPECT TO NOISE

Now, depending on the type described above program will set binarization threshold value automatically. It's very important to select threshold value properly. Otherwise, accuracy will be very poor [9] [10] Threshold value differ from image to image which depends on type of noise. We have found optimum threshold value for different types of image e.g. One value for very small noisy image and another value for other noisy image.so that our program can extract text from different types of pdf files with very high accuracy .If we use same threshold value for different images of different type accuracy will fall . We'll discuss it in result and analysis section again.

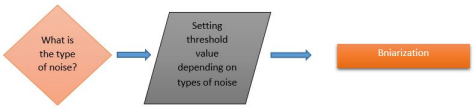


Fig. 9. Image post processing

2) Extracting text from image: We have used Pytesseract engine to perform this task. Pytesseract Engine: Pytesseract or Python-tesseract is a wrapper for Google's Tesseract-OCR engine. It extract text from an image. We use data for Bengali language to extract Bengali text from the image. Pytesseract engine can extract data from a noise free image with 100% accuracy. But if it's given noisy image then it's accuracy falls tremendously. Let's test for an image: Following image is directly given to the engine and after processing , again the processed image has sent to the engine. We observed the following result:

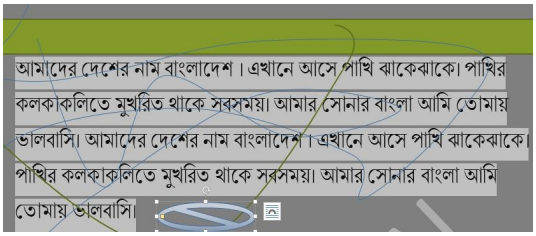


Fig. 10. Input Image

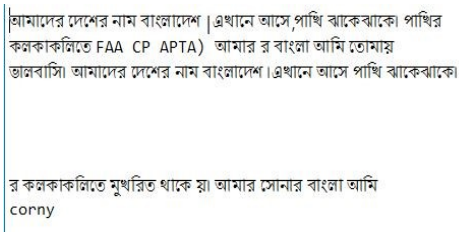


Fig. 11. Output of pytesseract without processing

Observation: Indentation has break down and failure of text extractions is 9 words.

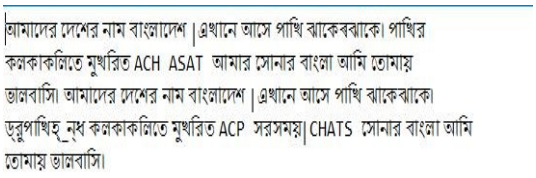


Fig. 12. Output of pytesseract with processing

Observation: By applying , preprocessing indentation has been preserved and number of failure of text extraction is 4 words. That is error words amount has reduced by less than half.

3) Text Normalization: Text Normalization is basic text processing before sending text to a text to speech engine. Without text normalization speech form will hamper a lot (e.g if there is a date write on as 12/12/12 then text to speech engine will say 'twelve twelve twelve ' in 'Twelve December Two thousand twelve') [3]. So make speeches more realistic we have done text normalization to the text that we get from pytesseract engine.To make paper independent, normalization process has described as follows:

4) Converting Normalized text to Voice: We used Google text to speech engine for converting text to voice. It take input as text format and then produce an mp3 file.There are several speech synthesis system like festival synthesis system [11] and google. But we have used google text to speech synthesis system due to its efficiency and pronunciation is much better than others.

III. EXPERIMETAL RESULT & DISCUSSION

A. Checking Image Quality generated from PDF

We have used pdf2image python library's pdf2image() function to generate image from pdf of specific pages.We will see some of the image that we have generated from pdf in this section .

Following two images are of randomly selected two pdf page that we converted to image from pdf format.We generated images of size (800,1200) of each.

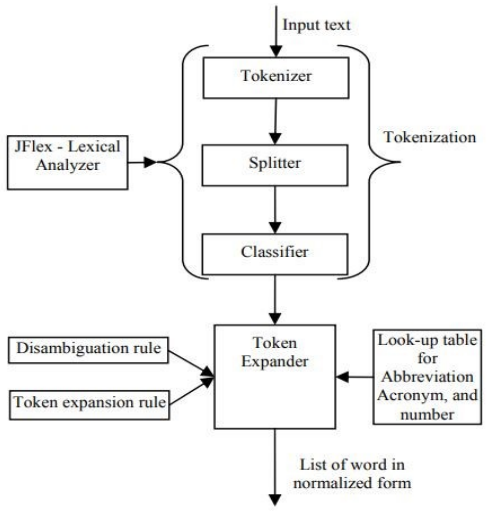
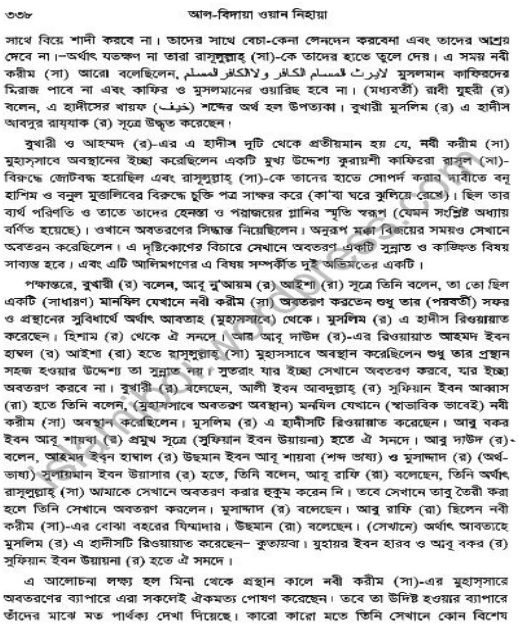


Fig. 13. Text Normalization process [3]



৩৩৮ আল-বিদায়া ওয়ান নিহায়া

সাথে বিয়ে শাদী করবে না। তাদের সাথে বেচা-কেনা লেনদেন করবেনা এবং তাদের আশ্রয় দেবে না।–অর্থাৎ যতক্ষণ না তারা রাসূলুল্লাহ্ (সা)-কে তাদের হতে ভুলে দেয়। এ সময় নবী করীম (সা) আরো বলেছিলেন, لايرث المسلم الكافر ولاالكافر المسلم মুসলমান কাফিরদের মিরাজ পাবে না এবং কাফির ও মুসলমানের ওয়ারিছ হবে না। (মধ্যবর্তী) রাবী যুহরী (র) বলেন, এ হাদীসের খায়ফ (خيف) শব্দের অর্থ হল উপত্যকা। বুখারী মুসলিম (র) এ হাদীস আবদুর রায্যাক (র) সূত্রে উদ্ধৃত করেছেন।

বুখারী ও আহমদ (র)-এর এ হাদীস দুটি থেকে প্রতীয়মান হয় যে, নবী করীম (সা) মুহাস্সাবে অবস্থানের ইচ্ছা করেছিলেন একটি মুখ্য উদ্দেশ্য কুরায়শী কাফিররা রাসূল (সা)-বিরুদ্ধে জোটবদ্ধ হয়েছিল এবং রাসূলুল্লাহ্ (সা)-কে তাদের হাতে সোপর্দ করার দাবীতে বনূ হাশিম ও বনূল মুত্তালিবের বিরুদ্ধে চুক্তি পত্র সাক্ষর করে (কা'বা ঘরে ঝুলিয়ে রেখে)। ছিল তার ব্যর্থ পরিণতি ও তাতে তাদের হেল্তা ও পরাজয়ের গ্লানির স্মৃতি স্বরূপ (যেমন সংশ্লিষ্ট অধ্যায় বর্ণিত হয়েছে)। ওখানে অবতরণের সিদ্ধান্ত নিয়েছিলেন। অনুরূপ মক্কা বিজয়ের সময়ও সেখানে অবতরন করেছিলেন। এ দৃষ্টিকোণের বিচারে সেখানে অবতরণ একটি সুন্নাত ও কাঙ্ক্ষিত বিষয় সাব্যস্ত হবে। এবং এটি আলিমগণের এ বিষয় সম্পর্কীত দুই অভিমতের একটি।

পক্ষান্তরে, বুখারী (র) বলেন, আবূ মু'আয়ম (র) আইশা (রা) সূত্রে তিনি বলেন, তা তো ছিল একটি (সাধারণ) মানযিল যেখানে নবী করীম (সা) অবতরণ করতেন শুধু তার (পরবর্তী) সফর ও প্রস্থানের সুবিধার্থে অর্থাৎ আবতাহ (মুহাস্সাবে) থেকে। মুসলিম (র) এ হাদীস রিওয়ায়াত করেছেন। হিশাম (র) থেকে ঐ সনদে। আর আবূ দাউদ (র)-এর রিওয়ায়াত আহমদ ইবন হাম্বল (র) আইশা (রা) হতে রাসূলুল্লাহ্ (সা) মুহাস্সাবে অবস্থান করেছিলেন শুধু তার প্রস্থান সহজ হওয়ার উদ্দেশ্য তা সুন্নাত নয়। সুতরাং যার ইচ্ছা সেখানে অবতরণ করবে, যার ইচ্ছা অবতরণ করবে না। বুখারী (র) বলেছেন, আলী ইবন আবদুল্লাহ্ (র) সুফিয়ান ইবন আব্বাস (রা) হতে তিনি বলেন, (মুহাস্সাবে অবতরণ অবস্থান) মনযিল যেখানে (স্বাভাবিক ভাবেই) নবী করীম (সা) অবস্থান করেছিলেন। মুসলিম (র) এ হাদীসটি রিওয়ায়াত করেছেন। আবূ বকর ইবন আবূ শায়বা (র) প্রমুখ সূত্রে (সুফিয়ান ইবন উয়ায়না) হতে ঐ সনদে। আবূ দাউদ (র)-বলেন, আহমদ ইবন হাম্বাল (র) উছমান ইবন আবূ শায়বা (শব্দ ভাষ্য) ও মুসাদ্দাদ (র) (অর্থ-ভাষ্য) সুলায়মান ইবন উয়াসার (র) হতে, তিনি বলেন, আবূ রাফি (রা) বলেছেন, তিনি অর্থাৎ রাসূলুল্লাহ্ (সা) আমাকে সেখানে অবতরণ করার হুকুম করেন নি। তবে সেখানে তাবূ তৈরী করা হলে তিনি সেখানে অবতরণ করলেন। মুসাদ্দাদ (র) বলেছেন। আবূ রাফি (রা) ছিলেন নবী করীম (সা)-এর বোঝা বহরের যিম্মাদার। উছমান (রা) বলেছেন। (সেখানে) অর্থাৎ আবতাহে মুসলিম (র) এ হাদীসটি রিওয়ায়াত করেছেন– কুতায়বা। যুহায়র ইবন হারব ও আবূ বকর (র) সুফিয়ান ইবন উয়ায়না (র) হতে ঐ সনদে।

এ আলোচনা লক্ষ্য হল মিনা থেকে প্রস্থান কালে নবী করীম (সা)-এর মুহাস্সারে অবতরণের ব্যাপারে এরা সকলেই ঐকমত্য পোষণ করেছেন। তবে তা উদিষ্ট হওয়ার ব্যাপারে তাঁদের মাঝে মত পার্থক্য দেখা দিয়েছে। কারো কারো মতে তিনি সেখানে কোন বিশেষ

Fig. 14. image of pdf page no 336

B. Efficiency Measurement

[h!] We have shown in Methodology section, how image processing increased overall efficiency of our computer application. Now we will measure overall accuracy of our application.We will use following formula for measuring accuracy of our application.

$$= \frac{numberofaccuratelydetectedcharacter}{Totalnumberofcharacters} * 100\%$$

let's test for last two images that we have generated in last subsection.

for first page:

আল-বিদায়া ওয়ান নিহায়া ৩৩৭

ইমাম আহমদ (র) বলেন, নূহ ইবন মায়মূন (র) ইবন উমর (রা) হতে এমর্মে বর্ণনা করেন যে, রাসূলুল্লাহ্ (সা) এবং আবূ বকর, উমার ও উসমান (রা) মুহাস্সাবে অবতরণ করেছেন। ইমাম আহম্দ (র)-এর মুসনাদে আবদুল্লাহ আল আমরী নাফি (র) সনদের হাদীস রূপে আমি এভাবে অধ্যয়ন করেছি। তিরমিযী (র) এ হাদীসটি রিওয়ায়ত করেছেন, ইসহাক ইবন মনসুর (র) হতে। ইবন মাজা (র) রিওয়ায়াত করেছেন, মুহাম্মদ ইবন ইয়াহয়া (র) হতে– (উভয় সনদ....) নাফি–ইবন উমর (রা) হতে, তিনি বলেন, রাসূলুল্লাহ (সা), আবূ বকর, উমর ও উছমান (রা) আবতাহে অবতরণ করতেন। তিরমিযী (র) বলেছেন, এ প্রসংগে আইশা, আবূ রাফি ও ইবন আব্বাস (রা) হতে রিওয়ায়াত রয়েছে। ইবন উমার (রা) হতে আগত হাদীসটি হাসান-গারীব একক সূত্রে উত্তম।

কেননা, শুধু আবদুর রায্যাক– উবায়দুল্লাহ ইবন উমার (রা) সনদের হাদীস রূপে এটির সাথে আমাদের পরিচিতি। মুসলিম (র) এ হাদীসটি রিওয়ায়াত করেছেন। মুহাম্মদ ইবন সিহ্রান আল রাযী (র) ইবন উমর (রা) সূত্রে এমর্মে যে, রাসূলুল্লাহ্ (সা) এবং আবূ বকর ও উমর (রা) আবতাহে অবতরণ করতেন। মুসলিম (র) সাখ্র ইবন জুওয়ায়রিয়া নাফি ইবন উমর (রা) হতেও রিওয়ায়াত করেছেন যে, তিনি ইবন উমার মুহাস্সাবে অবতরণ করতেন এবং নাফ্র দিবসের যুহর সালাত হাসবায় (মুহাস্সাবে) আদায় করতেন। নাফি (র) বলেন, রাসূলুল্লাহ (সা) এবং তাঁর পরবর্তী খলীফাগণ মুহাস্সাবে অবতরণ করেছেন। ইমাম আহম্দ (র) আরো বলেন, ইউনুস (র)....ইবন উমার (র) হতে এ মর্মে যে, রাসূলুল্লাহ (সা) যুহর, আসর, মাগরিব ও ইশার সালাত বাতাহায় (আবতাহে) আদায় করলেন, তারপর কিছুক্ষণ শুয়ে থাকলেন। তারপর মক্কায় প্রবেশ করলেন এবং বাযতুল্লাহ তাওয়াফ করলেন। আহম্দ (র) আফ্ফান (র)....ইবন উমর (রা) হতেও রিওয়ায়াত করেছেন। এ রিওয়ায়াতের শেষ অংশে অধিক রয়েছে ইবন উমর (রা) ও তা করতেন। আবূ দাউদ (র) ও আহম্দ ইবন হাম্বল (র) হতে অনুরূপ রিওয়ায়াত করেছেন।

বুখারী (র) বলেন, হুমায়দী (র)....(যুহরী)....আবূ হুরায়রা (রা) হতে তিনি বলেন, মিনায় নাহর দিবসের পর তারিখ-এর পরের দিন রাসূলুল্লাহ্ (সা) বললেন–

نحن نازلون غدا بخيف بنى كنانة حيث تقاسموا على الكفر -

আমরা আগামী দিন বনূ কিনানার উপত্যকায় অবতরণ করব, যেখানে তারা কুফরীতে অটল থাকার চুক্তিতে অংগীকারাবদ্ধ হয়েছিল। তিনি এ গিরিপথ বলতে মুহাস্সাবকে বুঝিয়েছিলাম। (পূর্ণ হাদীস) মুসলিম (র) হাদীসটি অনুরূপ রিওয়ায়াত করেছেন, মুহায়র ইবন হারব (র) (আওয়াই) সূত্রে ইমাম আহম্দ (র) বলেন, আবদুর রায্যাক (র) উসামা ইবন যায়দ (রা) হতে। তিনি বলেন, আমি বললাম, ইয়া রাসূলুল্লাহ্ (সা) আগামী কাল কোথায় অবস্থান নেবেন? এটি তার হজ্জের সময়ের কথা। তিনি বললেন, وهل ترك لنا عقيل منزلا আকীল (ইবন আবূ তালিব) কি আর আমাদের কোন ঘরবাড়ি রেখেছে? তারপর বললেন, আগামী কাল আমরা ইনশাআল্লাহ্ বনূ কিনানার উপত্যকায় অর্থাৎ মুহাস্সাবে–অবস্থান নেব যেখানে তারা কুরায়শীদের সাথে কুফরীতে ও অংগীকারাবদ্ধ হয়েছিল। ঘটনাটি হল যে, বনূ কিনানা বনূ হাশেমীর বিরুদ্ধে অন্যান্য কুরায়শীদের সাথে এ ব্যাপারে শপথ যুক্ত আঁতাত ও চুক্তি করল যে তারা হাশিমীদের

Fig. 15. image of pdf page no 337

$$= \frac{1132}{1189} * 100\% \ = 95.21\%$$

for second page:

$$= \frac{1378}{1451} * 100\% \ = 94.902\%$$

from above result we observed that accuracy is almost 95%

IV. Conclusion

In this project , we tried to develop a computer application that can read pdf file. We tried to increase the accuracy of the system by implementing processing method where image taken from the pdf is classified depending on the value of standard deviation and mean of the pixels . The method automatically select parameters(e.g threshold for binarization) depending on the value of SD and Mean. We have also focused on normalization process to increase text readability.Our application is ready and it Is very efficient on reading any bangla pdf .

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] B Fadiora, F Wada, and OB Longe. Combining optical character recognition (ocr) and edge detection techniques to filter image-based spam. *African Journal of Computing & ICT January*, 5(1):59–68, 2012.

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

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

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

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