
Speak My Scan: An Assistive Image-to-Speech System for Visually Impaired Users

Supriya Jagtap¹, R. Bhutada², P. Bangar³, R. Bhusare⁴, K. Akolkar⁵, R. Biradar⁶

^{1,2,3,4,5,6}Assistant Professor, Information Technology Department, Vishwakarma Institute of Technology, Pune, India
Information Technology Department, Vishwakarma Institute of Technology, Pune, India

¹supriya.jagtap@vit.edu, ²riddhi.bhutada24@vit.edu, ³pranali.bangar24@vit.edu,
⁴rohan.bhusare24@vit.edu, ⁵kartik.akolkar24@vit.edu, ⁶ritesh.biradar24@vit.edu

Abstract.

Visually impaired users have not yet access to printed information, as the conventional screen readers are not able to directly interpret the content of images or of a scanned document. To solve this issue, this paper introduces Speak My Scan, an offline assistive system based on Optical Character Recognition (OCR) and Text-to-Speech (TTS) technologies which converts image and document text to speech. The system proposed is based on OpenCV for image preprocessing, Tesseract for the extraction of information in the form of text, and pyttsx3 for offline speech synthesis. It accepts image files, multi page PDF and live camera input via a GUI. A test set of 50 images of printed documents, book pages, notices and signage were used in the experimental evaluation under various lighting and image conditions. For printed documents in good lighting, the system achieved a top scorer of over 94% OCR accuracy and adaptive preprocessing improved OCR accuracy by about 11 percentage points over the direct processing of raw images. The speed of the processing time of a single A4 page was 2.3 seconds on an average laptop. The suggested system shows that an offline OCR-TTS system can be practical and accessible for providing reading assistance, even without special equipment or Internet access.

Keywords. Optical Character Recognition, Text-to-Speech, Assistive Technology, Image Processing, Computer Vision, Visually Impaired

1. INTRODUCTION

Printed information is a difficult resource to access for the visually impaired. Traditional screen readers work well with digital text on computer screens, but can't access information in pictures, books, notices or scanned documents. Consequently, many of the print materials that are readily available are inaccessible to visually impaired people. OCR can be used to convert an image to text, and once the text is extracted from an image, it can be converted into speech using TTS. These technologies can be used together in an OCR-to-speech pipeline to enable access to printed information by speech output. The Speak My Scan system consists of adaptive image preprocessing, Tesseract OCR and pyttsx3 speech synthesis by a single desktop application. The proposed solution is not cloud dependent, nor requires any special hardware to be installed. It also can accept various input types such as

static images, PDF files and live camera input. The goal of the system is to give visually impaired users an easy, low-cost and convenient alternative access to the printed information..

2. RELATED WORK

There are a number of researches on OCR and TTS for assistive applications. Smith developed one of the popular open-source OCR systems, Tesseract OCR, and OpenCV offers image-processing methods to enhance the image quality prior to OCR. Ben Atitallah built an on-the-fly TTS system which required special hardware, but it worked with Raspberry Pi. Ariyoshi and Chun suggested a cloud-based processing scheme based on machine learning for OCR-to-audio recognition, which has a high recognition accuracy, but is not applicable in areas with weak Internet connectivity. Al-Mansoori studied on OCR and generation of braille using IoT, and Raj et al. implemented OCR based reading system using LabVIEW. Other research works explored the design of the multilingual image-to-speech systems and the newspaper reading application with the aid of OCR and layout analysis. The Speak My Scan proposed is a lightweight open source desktop application that integrates preprocessing, OCR, pdf processing, camera input and offline TTS.

3. PROPOSED INTEGRATED SYSTEM AND METHODOLOGY

The Speak My Scan proposed system is composed of three modules, namely Input Acquisition, Image Preprocessing, OCR-Based Text Extraction, Text Post-Processing and Speech Synthesis, which take the visual input and convert it into audible speech. The stages are separately upgradeable as well as replaceable.

3.1 Input Acquisition

Here are three types of input supported by the system: static images (JPEG, PNG, BMP), multi-page PDF documents converted to images by pdf2image and real-time camera input (OpenCV). This makes a system that can read documents, books, notices and signs.

3.2 Image Preprocessing

The image is enhanced with OpenCV prior to OCR. The pipeline consists of several operations: grayscale conversion, 5×5 Gaussian blur, adaptive Gaussian thresholding with a block size of 11, and morphological dilation with a kernel of size 2×2. These operations dampen noise, deal with uneven lighting and increase the visibility of characters. 50 document images were used to fine-tune the parameters.

3.3 OCR-Based Text Extraction

Tesseract OCR v5.x is fed with the preprocessed image via pytesseract. The system adopts the LSTM based OCR engine with PSM 6, which is appropriate for document-like text. English is default language, other languages can be added via language data files.

3.4 Text Post-Processing

Non printable characters, normalizes whitespace and removes isolated character artefacts using regular-expression rules are cleaned when creating raw OCR output. The text is next cleaned and then presented in the GUI for speaking.

3.5 Speech Synthesis

The cleaned text is then converted to speech with pyttsx3 using the OS's built-in speech synthesizer and without requiring any internet connection. Speech rate, voice and volume can be set, and TTS works in a separate thread to maintain responsiveness of the GUI.

4. RESULTS AND DISCUSSION

50 images of various types of printed documents, book pages, notices and signage were used in various conditions to test Speak My Scan, including good lighting, dim lighting and slight skew.

4.1 OCR Performance

It obtained a higher OCR accuracy of over 94% at the character level for the printed documents under good lighting conditions. Adaptive thresholding was used to enhance the accuracy of the OCR by about 11 percentage points over the raw images processed straight to Tesseract.

4.2 PDF and Speech Processing

The preprocessing pipeline was able to extract high-fidelity text from the PDFs generated from the digital files and the scanned PDFs. Understandable speech was produced by the pyttsx3 engine, and preliminary testing with five visually impaired volunteers resulted in acceptable speech understanding. A laptop computer using an Intel Core i5 processor with 8 GB RAM was used to measure the processing time for a single page of A4: 2.3 seconds, adequate for assistive use.

4.3 Limitations

Low-quality images -- especially those below about 150 DPI -- and handwritten or ornamental fonts suffer in the quality of the recognition performance. Documents with multiple columns can also cause reading order issues if no further layout analysis is performed.

The default mode of the current system is English; other languages will need to use the appropriate language packages of the Tesseract software.

5. CONCLUSION

Speak My Scan is an offline and open-source assistive system presented in this paper to translate text from images and documents to audible speech for visually impaired users. The system features pre-processing using OpenCV, OCR using Tesseract, and then post-processing the text and using pyttsx3 for the TTS through a Tkinter GUI. Works with images, PDFs, and camera input—no Internet connection or special equipment. The system gained OCR accuracy of over 94% from printed documents under good illumination, and adaptive preprocessing increased the accuracy of OCR by about 11 percentage points. The average processing time per A4 page was 2.3 seconds and preliminary testing indicated good speech

comprehension. Plans involve handwritten and low-quality text recognition, multilingual support, deep learning-based OCR, document layout analysis, processing of video streams, deployment on low-cost hardware like Raspberry Pi, and more.

6. REFERENCES

- [1] Ben Atitallah, "Development of Real-Time TTS System Using Raspberry Pi for Visually Impaired," in Proc. Int. Conf. on Innovative Trends in Electronics and Communication Engineering (ICITECE), Bangalore, India, 2024, pp. 1–6.
- [2] J. Ariyoshi and J. Chun, "Machine Learning-Based OCR for Text-to-Audio Conversion," IEEE Access, vol. 13, pp. 1–12, 2025, doi: 10.1109/ACCESS.2025.0000001.
- [3] D. Raj, H. F. Tazeen, V. Ravish, A. H. N. Anupama, and R. N. Nagashree, "An Assistive Reading System for Visually Impaired and Blinds Using OCR and TTS Techniques on LabVIEW," Int. J. Eng. Res. Technol. (IJERT), vol. 7, no. 10, pp. 1–5, Jun. 2019.
- [4] Al-Mansoori, "IoT-Powered OCR-Braille System for Blind and Deaf-Blind Users," IEEE Internet Things J., vol. 11, no. 8, pp. 1–10, 2024.
- [5] R. Smith, "An Overview of the Tesseract OCR Engine," in Proc. 9th Int. Conf. Document Analysis and Recognition (ICDAR), Parana, Brazil, 2007, pp. 629–633, doi: 10.1109/ICDAR.2007.4376991.
- [6] Sundar, O. Abinaya, and V. Ajithkumar, "An Optical Character Recognition Framework Based Newspaper Reader Application for Blind," in Proc. 2021 Int. Conf. Computing, Communication, and Intelligent Systems (ICCAP), Greater Noida, India, Dec. 2021, pp. 1–6, doi: 10.4108/eai.7-12-2021.2314975.
- [7] R. Rosalina, H. Fahmi, and G. Sahuri, "Artificial Intelligence Multilingual Image-to-Speech for Accessibility and Text Recognition," Int. J. Artif. Intell., vol. 14, no. 1, pp. 1–10, 2025.
- [8] S. Chaturvedi, M. Chechani, K. Gautami, and K. Joshi, "System to Provide Reading Aid to Visually Impaired People," Asian J. Comput. Technol., vol. 9, no. 2, pp. 1–8, 2021.
- [9] Sharma, A. Srivastava, and A. Vashishth, "An Assistive Reading System for Visually Impaired Using OCR and TTS," Int. J. Comput. Appl., vol. 95, no. 2, pp. 13–18, Jun. 2014, doi: 10.5120/16566-6231.
- [10] P. Langde, S. Patil, and P. Langde, "Automating Document Narration: A Deep Learning-Based Speech Captioning System for Visually Impaired Person," Int. J. Intell. Syst. Appl. Eng. (IJISAE), vol. 12, no. 2, pp. 1130–1140, 2024, doi: 10.18201/ijisae.2024.7469.
- [11] G. Bradski, "The OpenCV Library," Dr. Dobb's J. Softw. Tools, vol. 25, no. 11, pp. 120–125, Nov. 2000.

Biographies



Prof. Supriya Jagtap is currently working as an Assistant Professor in the Information Technology Department at Vishwakarma Institute of Technology, Pune, India. Her professional interests include teaching, student mentoring, and academic activities in the field of Information Technology.



Riddhi Bhutada is a student in the Information Technology Department at Vishwakarma Institute of Technology, Pune, India. Her interests include Artificial Intelligence, Machine Learning, Data Science, and software development.



Pranali Bangar is affiliated with Vishwakarma Institute of Technology, Pune, India, as a student in the Information Technology Department. Her interests include Artificial Intelligence, Machine Learning, Data Science, and software development.



Rohan Bhusare is affiliated with Vishwakarma Institute of Technology, Pune, India, as a student in the Information Technology Department. His interests include Artificial Intelligence, Machine Learning, Data Science, and software development.



Kartik Akolkar is affiliated with Vishwakarma Institute of Technology, Pune, India, as a student in the Information Technology Department. His interests include Artificial Intelligence, Machine Learning, Data Science, and software development.



Ritesh Biradar is affiliated with Vishwakarma Institute of Technology, Pune, India, as a student in the Information Technology Department. His interests include Artificial Intelligence, Machine Learning, Data Science, and software development.