International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

From Vision To voice: A Multi -Model Assistive Framework

Annapoorna S
Student Of Computer Science Department
Vidya Academy of Science and Technology (VAST)
Thiruvananthapuram, India

Anjali B R
Student Of Computer Science Department
Vidya Academy of Science and Technology (VAST)
Thiruvananthapuram, India

Feba Jayasing
Student Of Computer Science Department
Vidya Academy of Science and Technology (VAST)
Thiruvananthapuram, India

Devika V S
Student Of Computer Science Department
Vidya Academy of Science and Technology (VAST)
Thiruvananthapuram, India

Ms. Jisha Raj T
Assistant Professor of Computer Science Department
Vidya Academy of Science and Technology (VAST)
Thiruvananthapuram, India

Dr. C. Brijilal Ruben
Head of Computer Science Department
Vidya Academy of Science and Technology (VAST)
Thiruvananthapuram, India

***Abstract*—This project titled "From Vision to Voice: A Multi-modal Assistive Framework" presents a Python-based assistive solution that converts visual content into audible speech. The system processes both digitally generated and scanned PDF documents by integrating Optical Character Recognition (OCR) and Text-to-Speech (TTS) technologies. It automatically detects page types, extracts text accurately, and converts it into multilingual audio output. The framework enhances accessibility for visually impaired users, supports auditory learning, and improves content usability. Parallel processing and intelligent text chunking are employed to ensure efficiency and scalability. This project demonstrates how multimodal processing can effectively bridge accessibility gaps using opensource tools.**

***Keywords*—Optical Character Recognition; Text-to-Speech; Assistive Technology; Image Processing; Accessibility; Tesseract; gTTS; Visually Impaired Users; Multi-Modal Systems.**

"The implementation and evaluation results of the proposed system have been submitted to the institution for academic purposes. This paper primarily focuses on the system design, architecture, and conceptual framework."

I. INTRODUCTION

The rapid advancement of digital technology has significantly transformed the way information is generated, stored, and accessed. A large share of everyday information is presented in visual formats such as images, scanned documents, handwritten notes, and Portable Document Format (PDF) files. While convenient for sighted individuals, these formats create serious accessibility barriers for visually impaired users, who rely heavily on assistive technologies such as screen readers to access digital content.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Screen readers convert digital text into speech but are limited to machine-readable text; they cannot extract or interpret text embedded in images, scanned documents, or non-editable PDFs. As a result, visually impaired individuals face difficulty accessing textbooks, printed documents, medicine labels, signboards, and handwritten notes, forcing dependence on others and reducing independence.

To address this gap, "From Vision to Voice: A Multi-Modal Assistive Framework" integrates Optical Character Recognition (OCR) and Text-to-Speech (TTS) technologies to convert visual text into spoken audio. The term "multi-modal" refers to the use of multiple modes of input and output: the input is visual (images or documents) and the output is auditory (speech), allowing users to access information through a different sensory channel and improving overall usability and accessibility.

II. LITERATURE SURVEY

The development of assistive technologies has become increasingly important with the growing dependence on digital information. This section reviews the evolution of the core technologies used in the proposed system and identifies the research gap that motivates it.

2.1 Evolution of Optical Character Recognition

Early OCR systems relied on template matching, which struggled with variations in font style, size, and orientation. Later systems incorporated statistical models such as Hidden Markov Models (HMMs), improving accuracy but still facing limitations under complex image conditions. Modern OCR engines such as Tesseract use deep learning techniques, particularly Long ShortTerm Memory (LSTM) networks, which learn sequential character patterns and significantly improve recognition accuracy, though performance remains sensitive to resolution, noise, skew, and lighting.

2.2 Image Preprocessing Techniques

Preprocessing plays a crucial role in OCR accuracy. Common techniques include grayscale conversion, which reduces computational complexity; noise reduction, which removes unwanted variation; and thresholding, which separates text from background. Additional operations such as edge detection, dilation, and erosion further improve text visibility, making preprocessing an essential component of any OCR pipeline.

2.3 Text-to-Speech Technology

TTS technology has evolved from rule-based concatenative synthesis, which produced unnatural, robotic speech, to modern parametric and neural network-based approaches capable of natural, expressive output. Google Text-to-Speech (gTTS) is one such cloud-based system. The TTS pipeline typically involves text normalization, phonetic transcription, and waveform generation.

2.4 Review of Existing Systems

Screen readers remain the most widely used assistive tool but cannot process image-based content. Mobile text-recognition apps combine device cameras with OCR and TTS but typically depend on internet connectivity. Wearable devices and smart glasses have also been explored for real-time assistance but remain expensive and less accessible.

2.5 Limitations and Research Gap

Existing systems are constrained by dependency on internet connectivity, sensitivity of OCR to image quality, and a lack of integration between OCR and TTS components, which forces users to rely on multiple disconnected tools. These limitations motivate the need for a single, integrated, accessibility-first framework, which this project addresses.

III. PROPOSED SYSTEM

The proposed system, "From Vision to Voice," is designed to address the limitations of existing solutions by providing an integrated pipeline for converting visual content into audible speech. The system accepts input in the form of images or PDF documents and processes it through image preprocessing, OCRbased text extraction, text cleaning, and Text-to-Speech synthesis.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The system is implemented in Python, integrating OpenCV for image processing, Tesseract OCR for text extraction, and gTTS for speech synthesis, with a graphical user interface built using Tkinter for user interaction. By combining all components into a single framework, the system eliminates the need for multiple applications and provides a seamless, accessible experience focused on simplicity and efficiency.

IV.PROBLEM STATEMENT

Visually impaired individuals continue to face significant challenges accessing textual information embedded in visual formats. Traditional assistive tools cannot interpret visual data, many OCR-based systems are complex to use or depend on internet connectivity, and extraction accuracy degrades under poor lighting, low resolution, or complex backgrounds. A strong need therefore exists for a system that can efficiently extract text from visual content and convert it into speech in a simple, reliable, and accessible manner.

V. SYSTEM ARCHITECTURE

The system architecture accepts an input image through an API request, which is passed through a pre-processing stage before being analysed by the Tesseract OCR engine. The OCR engine draws on the Leptonica image-processing library and a trained dataset to extract text, which is then refined by a post-processing stage before being returned as the final text output.

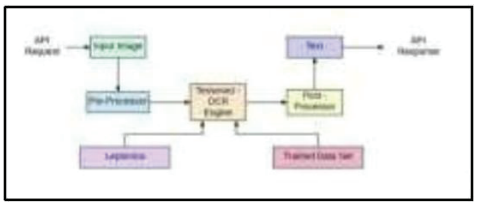

Fig-1: System Architecture of the Proposed System

VI. SYSTEM FLOWCHART

The overall workflow accepts images with text, PDF files, and scanned, printed, or handwritten documents as input. Each input type is routed through the OCR module, which extracts the underlying text before it is handed off to the Text-to-Speech stage for audio conversion, giving the system a single unified entry points regardless of the source format.

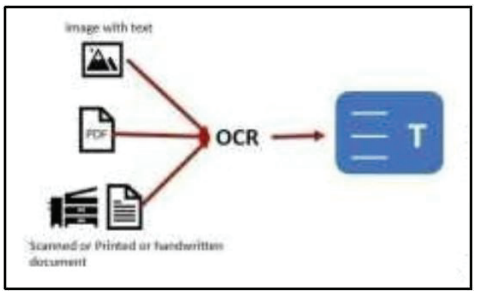

Fig-2: Flowchart of the Proposed System

VII. SOFTWARE REQUIREMENTS

- Programming Language: Python
- Image Processing: OpenCV
- OCR Engine: Tesseract OCR (with Leptonica)
- Speech Synthesis: Google Text-to-Speech (gTTS)
- Graphical User Interface: Tkinter
- Supporting Libraries: standard Python text-processing and file handling libraries for PDF and image input

VIII. HARDWARE REQUIREMENTS

- A standard desktop or laptop computer capable of running Python and the associated libraries
- A camera or scanner (optional), for capturing images of printed or handwritten text
- A speaker or headphone output for delivering the generated audio
- A stable internet connection, required for cloud-based speech synthesis via gTTS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

IX. SYSTEM WORKFLOW

Step 1 – Input Module: The user provides an image or PDF document as input.

Step 2 – Image Preprocessing: The input is enhanced through grayscale conversion, noise reduction, and thresholding to improve text visibility.

Step 3 – OCR Module: The pre-processed image is analysed by the Tesseract OCR engine, which extracts the embedded text and stores the input file.

Step 4 – Text Processing: The raw extracted text is cleaned, refined, and prepared for speech conversion, with the extracted text stored for reference.

Step 5 – TTS Module: The processed text is converted into an audio waveform using the Text-to-Speech engine and delivered to the user as spoken output.

X. DATA FLOW DIAGRAM

A Data Flow Diagram (DFD) describes how data moves through the system, the data stores involved, and the transformations applied at each stage. At the context level, the system as a whole receives visual input from the user and returns audio output. At Level 0, this is broken down into distinct processes with their own data stores, and Level 1 further decomposes the disease-relevant sub-processes into finer steps.

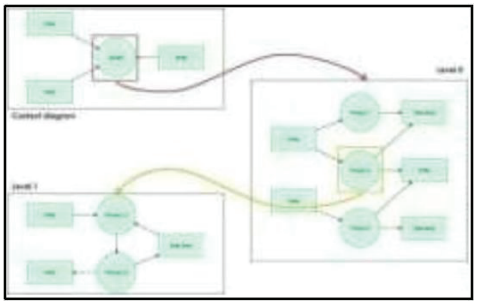

Fig-3: Context-Level Data Flow Diagram

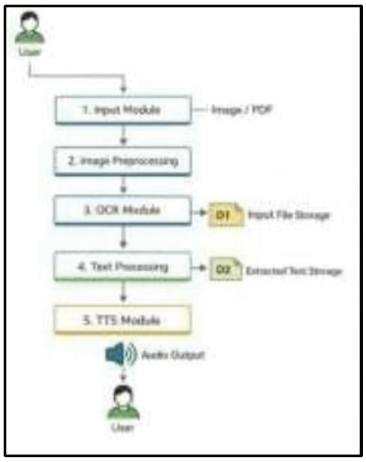

Fig-4: Data Flow Diagram – Input Module through TTS Module

XI. USE CASE DIAGRAM

The Use Case Diagram represents the interaction between the visually impaired user and the assistive system. The primary use cases include uploading an image or PDF document, capturing an image using a camera, extracting text through OCR, and converting the extracted text into speech through the TTS module before the audio output is played back to the user. This interaction highlights the simplicity of the system, which requires minimal user input to produce accessible audio output.

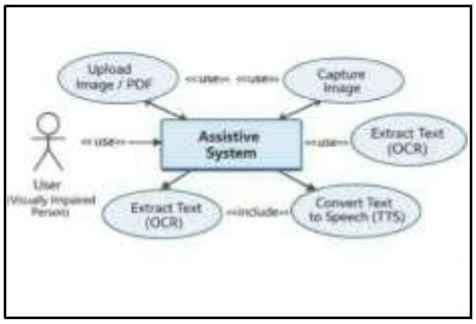

Fig-5: Use Case Diagram

XII. CONCLUSION

"From Vision to Voice: A Multi-Modal Assistive Framework" demonstrates the development of a system that converts visual content into audible speech using Optical Character Recognition and Text-to-Speech technologies, enabling visually impaired individuals to access textual information independently. By integrating image preprocessing, text extraction, and speech synthesis into a single framework using Python and open-source libraries, the system provides a cost-effective and easy to implement solution.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Testing indicates that the system performs well for clear and structured inputs, providing accurate text extraction and understandable speech output, while remaining sensitive to image quality and internet connectivity for cloud-based speech synthesis. Planned improvements include an offline TTS module, deep learning-based OCR for low-quality and handwritten text, additional language support, and real-time camera capture, extending the system toward a more portable and fully accessible assistive tool.

XIII. REFERENCES

[1] Ademi, V., and Ademi, L., "Natural language processing and text-to-speech technology," Journal of Natural Sciences and Mathematics, vol. 8, nos. 15-16, pp. 299-306, 2023.

[2] Agrawal, S., and Agrawal, N., "Recognition and speech conversion of Devanagari script using CNN," Proc. 2nd Int. Conf. Innovative Technology (INOCON), Mar. 2023, pp. 1-4.

[3] Amin, M., "Writing to speech conversion application using an Android-based camera," Proc. Int. Conf. Science Development Technology, vol. 3, no. 1, 2023, pp. 91-95.

[4] Anjaneyulu, P., Deekshitha, M. V. S., Sridevi, P., Srilekha, P., and Reddy, K. P., "A novel OCR system developed to synthesize speech from text using Raspberry Pi," AIP Conference Proceedings, vol. 2808, no. 1, 2023, Art. no. 030050.

[5] Babu, M. P., and Anitha, G., "OCR-based image text-to-speech conversion using KNN and comparison with fuzzy K-means clustering," Proc. Int. Conf. Advanced Computing, May 2023, pp. 1-5.

[6] Bhat, S., Bhat, P., and Kolekar, S. V., "From Vision to Voice: A System for the Physically Impaired," IEEE Access, 2025.

[7] Bhasin, K., Goel, A., Gupta, G., Singh, S. K., and Bhowmick, A., "A secure mobile application for speech-to-text conversion using AI techniques," Proc. WINS/CVMLH, 2023, pp. 44-53.

[8] Gopi, S., Palanivasan, S., Padmanaban, M., and Gowtham, C. V., "Virtual learning environment for visually impaired people using OCR and TTS," Proc. Int. Conf. RMKMATE, Nov. 2023, pp. 1-5.

[9] Kumar, S., Prabhu, P. S., Bhat, M. I., Kumar, S., and Shubha, B., "Text detection and recognition using machine learning," Proc. National Conf. Control Instrumentation Systems, Springer, Jan. 2024, pp. 389-407.

[10] Kunekar, P., et al., "Camera detection for blind people using OCR," Proc. 5th Biennial Int. Conf. Nascent Technologies in Engineering (ICNTE), Jan. 2023, pp. 1-6.

[11] Nakano, Y., Saeki, T., Takamichi, S., Sudoh, K., and Saruwatari, H., "VTTS: Visual-text to speech," Proc. IEEE Spoken Language Technology Workshop (SLT), Jan. 2023, pp. 936-942.

[12] Padmavathi, P., et al., "OCR and text-to-speech generation system using machine learning," Proc. 2nd Int. Conf. Applied AI and Computing (ICAAIC), May 2023, pp. 1-6.

[13] Prajapati, N. K., et al., "OCR-based assistive system for blind people," Soft Computing and Signal Processing, Springer, Jul. 2021, pp. 71-79.

[14] Raja, M., and Chary, B. P., "Development and deployment of a mobile application assistive device focused on neuro-OCR with speech production," AIP Conference Proceedings, vol. 2758, no. 1, 2023, Art. no. 030029.

[15] Sathana, V., Sneka, S., Sruthika, I., Sujitha, S., and Yogaasri, T., "A soundbite-based framework for text and object detection using OCR and YOLO technique to assist blind and deaf," Proc. 7th Int. Conf. Trends in Electronics and Informatics (ICOEI), Apr. 2023, pp. 1596-1602.

[16] Singh, A. R., Bhardwaj, D., Dixit, M., and Kumar, L., "An integrated model for text-to-text, image-to-text, and audio-to-text conversion using machine learning," Proc. 6th Int. Conf. Information Systems and Computer Networks (ISCON), Mar. 2023, pp. 1-7.

[17] Sivasubramanian, A., Shah, S., Narayanaswamy, A., Rindhya, C., and Ganesh, H. B. B., "Performing text segmentation to improve OCR on multi-scene text," Proc. Int. Conf. Artificial Intelligence and Speech Technology, Nov. 2024, pp. 66-77.

[18] Thanneru, S. H., Kumari, K., Kunta, N., and Manchalla, P. K., "Image to audio, text to audio, text to speech, video to text conversion using NLP techniques," E3S Web of Conferences, vol. 391, Jan. 2023, p. 01092.

[19] Umatia, S., Varma, A., Syed, A., Tiwari, K., and Shah, M. F., "Text recognition from images," International Journal of Research in Applied Science and Engineering Technology, vol. 10, no. 11, pp. 1003-1009, 2022.

[20] Venkatesh, M., et al., "Application of multilingual OCR algorithm for converting text from images and PDFs," Proc. Int. Conf. Data Science, Oct. 2024, pp. 1025-1031.