Advancements in OCR and TTS Systems: A Review of Technologies for Accessibility

Lalana[1], Prerana Hebsur[2], Maruthi Reddy B M[3], Mahammed Kaif H[4], Mithun B N[5], Krushitha Shetty[6]
[1,2,3,4,5,6]*Dept of Information Science and Engineering, APSCE, Bengaluru, Karnataka, India*

***Abstract*— Optical Character Recognition (OCR) and Text-To-Speech (TTS) are vital technologies that allow machines to read and articulate written content. OCR retrieves text from images, scanned papers, and photographs, whereas TTS translates this text into natural, human-like voice. When combined, these technologies form powerful OCR-TTS systems that improve accessibility for individuals with visual disabilities by delivering real-time audio output from both printed and digital materials. In addition to accessibility, their integration benefits various sectors. In education, OCR-TTS tools convert textbooks and educational materials into audio format, promoting inclusivity for students with reading difficulties. In the logistics and supply chain sectors, they optimize processes by reading labels, invoices, and shipping papers. For language learners, these systems provide auditory feedback that improves pronunciation and understanding. In smart cities, they can vocalize signs, timetables, and announcements for the public. This review emphasizes developments in segmentation, language modeling, Transformer-based OCR, and neural TTS, highlighting their potential for automation, inclusivity, and improved user interaction.**

***Index Terms*— Embedded system, OCR, TTS, pytesseract, gtts, trOCR.**

I. INTRODUCTION

Optical Character Recognition (OCR) is a technology that transforms handwritten or printed text from images, scanned documents, or photographs into a format that machines can read. By leveraging AI and pattern recognition, it detects characters, words, and paragraphs, making it extensively useful in book digitization, automated data entry, and text extraction for editing or translation. Text-to-Speech (TTS) transforms written text into spoken words by dissecting and synthesizing text into smaller components, such as words and sentences, and then generating audible speech with digital voices. It enhances accessibility for individuals with visual impairments or reading challenges and facilitates communication between written and spoken format.

The integration of Optical Character Recognition (OCR) and Text-to-Speech (TTS) technology has led to the development of an innovative system that transforms written text found in images or printed materials into spoken language, thereby enhancing accessibility and fostering automation in diverse areas. By connecting visual information with auditory forms, OCR-TTS systems open up avenues for both inclusivity and efficiency. A significant use case is the enhancement of accessibility for individuals who are visually impaired, as they frequently encounter difficulties in obtaining printed texts, signs, books, and handwritten materials. With a smartphone or an OCR-enabled device, users can take a picture of the text, which the OCR technology then processes to extract the information, while TTS vocalizes it nearly instantly, offering timely support and increased autonomy (see Fig 1).

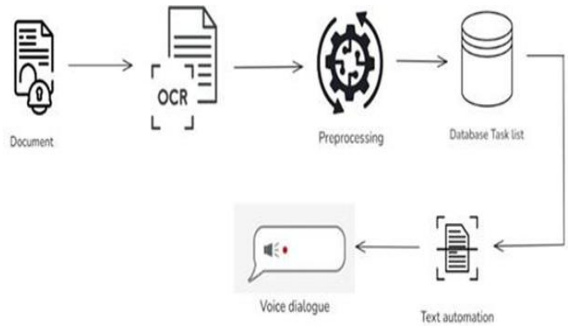

Figure 1 OCR with TTS Engine

It improves access for individuals with visual disabilities, aids students with learning differences or dyslexia, and helps those learning new languages. Different sectors utilize it in logistics, healthcare,

corporate offices, and smart city initiatives to enhance automation and effectiveness. By merging visual and auditory data, OCR-TTS promotes inclusivity, educational opportunities, productivity, and real-time translation.

II. REVIEW OF LITERATURE

The OCR with TTS Converter initiative evaluates essential studies, datasets, benchmarks, significant projects, and methods, while contrasting them with past research. Although machines have reached a point where they can emulate human skills such as reading and writing, people with visual impairments continue to encounter difficulties in obtaining information. These obstacles hinder movement in situations where text is not readily accessible. The project's goal is to overcome these challenges, improving accessibility and fostering independence.

The [1] paper introduces a text-to-speech system that integrates OCR and TTS, utilizing Tesseract for text extraction and pyttsx3 for speech synthesis, designed to aid visually impaired individuals. Core elements encompass image enhancement, text detection, and audio output, with LSTM processing handwritten text and accommodating multiple languages. Its applications include assistive devices, document digitization, and video content analysis. Benefits consist of offline functionality and support for various languages, while drawbacks include challenges with low-quality images, cursive writing, and real-time processing limitations. The evaluation focuses on accuracy, efficiency, and user experience, with future enhancements aimed at improving handwriting recognition, broadening language support, and increasing processing speed.

The [2] research introduces TrOCR, an OCR model that leverages the Transformer architecture and utilizes pre-trained Transformers for both images and text, incorporating DeiT and BEiT for image encoding. It surpasses conventional CNN-RNN models in the recognition of printed, handwritten, and scene text. Potential uses include digitizing documents, processing receipts, enhancing assistive technologies, automating data entry, and extracting multilingual text. Challenges faced include the requirement for high-quality images, difficulties with distorted text, and significant computational requirements. Nevertheless, TrOCR's effectiveness and straightforwardness render it a valuable OCR solution, though further enhancements are necessary for real-time applications.

The [3] paper presents a method for extracting text from images using Tesseract-OCR, combined with Python libraries such as OpenCV and PIL for image processing. The primary steps involve segmenting lines, identifying text locations, and recognizing individual characters. This solution is provided as an executable file (.exe) that is compatible with formats like .jpg and .png, facilitating the digitization of documents and assisting individuals with visual impairments. Its uses include transforming text into editable formats, processing historical documents, and supporting multilingual text. The challenges faced include managing curved text, busy backgrounds, and images with poor resolution, highlighting the need for further enhancements to address complex text and diverse languages.

The [4] paper introduces a Text-to-Speech (TTS) synthesis system capable of accommodating phonetic and prosodic differences with a limited amount of training data. It incorporates an accented front-end for converting graphemes to phonemes (G2P) and an acoustic model that estimates pitch and duration. The evaluation criteria encompass phoneme error rate (PER), word error rate (WER), Kullback-Leibler divergence (KLD), and root mean square error (RMSE) for fundamental frequency (F0) and duration. Potential applications include tailored voice assistants, audiobooks, and customer service, offering natural, accented speech to improve user engagement.

The [5] research introduces a Handwritten Text Recognition system that utilizes Pytesseract, which is a Python interface for Tesseract-OCR. Essential steps involveimage preprocessing such as converting to grayscale, applying Gaussian blur, and performing thresholding, followed by OCR processing using tailored configurations and post-processing to address spelling, grammar, and formatting issues. Potential uses encompass digitizing old documents, assisting individuals with visual impairments, and enhancing workflows in business, research, and legal document mana gement. Challenges include diverse handwriting styles, subpar image quality, and the requirement for

significant preprocessing; however, Pytesseract provides a versatile and accessible OCR solution.

The [6] paper introduces a portable text-to-speech (TTS) converter designed for users with visual impairments, capable of converting both printed and handwritten materials into spoken words. A handheld page scanner captures the text, which is then processed using Tesseract OCR, followed by audio conversion through a TTS API. It supports English, Hindi, and Bengali, and sends scanned images to an Android device via Bluetooth. Potential applications include assistive technology, multilingual translation, and educational resources. However, it has some limitations, such as an 88.6% accuracy rate for handwriting recognition, limited language options, and processing times of 7 seconds for the scanner and 12 seconds for the phone camera. Nevertheless, it remains an affordable, accessible, and user-friendly option.

The [7] research details Image Processing-Based Optical Character Recognition (OCR) with Text-To-Speech (TTS) System is designed to aid those with visual impairments by transforming printed text into spoken words using a webcam coupled with OCR technology. The system captures images, utilizes Tesseract OCR to extract text, and then vocalizes the information through Google Text-to-Speech (gTTS). Key components involve image processing (OpenCV, LSTM, RNN), text segmentation, feature extraction, and speech synthesis. Its applications extend to reading documents, postal services, and various assistive technologies. While the system enhances accessibility, it does face challenges such as challenges with recognizing handwritten text, dependency on the quality of the webcam, and delays in processing

The [8] paper reviews Voice Transformer Network (VTN) a sequence-to-sequence voice conversion model that utilizes Transformer architecture along with TTS pretraining, surpassing conventional RNN/CNN-based systems in converting prosody and duration. It incorporates a two-phase pretraining approach: first, it trains a Transformer-based TTS model, and then it influences the encoder pretraining with the decoder through an autoencoder framework. This method facilitates the effective acquisition of detailed speech representations. VTN achieves better intelligibility, naturalness, and similarity, even when data is scarce, requiring reduced training time and performing well with as few as 80 utterances, thus improving efficiency and quality in voice conversion.

The [9] paper describes a multistage system that converts images from OCR to TTS, which includes steps like geometric correction, image preprocessing, text detection, and speech generation. Geometric correction employs homography to rectify perspective distortions, while noise is eliminated using both median and Gaussian filters. Text recognition is carried out with Tesseract OCR utilizing LSTM, and the identified text is transformed into speech through Google Text-to-Speech (gTTS). The system generates audio narrations from input images, achieving approximately 85% accuracy and offering support for multiple languages. The difficulties faced include various fonts, images of low quality, and segmentation mistakes, which could be mitigated by improving OCR training and filtering techniques.

The [10] paper introduces an OCR system that utilizes Tesseract along with Python, OpenCV, and PHP, improved through preprocessing techniques such as binarization, thresholding, conversion to grayscale, and noise reduction. Before Tesseract extracts the text, adaptive thresholding is employed to segment the images. The system is capable of processing various types of images, including pages from books, handwritten notes, and digital text, resulting in well-structured text files. It achieved a training accuracy of 87% and 91% on test images, performing optimally with high-quality images but facing difficulties with blurred or poorly lit photographs.

The [11] paper describes a MATLAB-based system that integrates Optical Character Recognition (OCR) and Text-to-Speech (TTS) technologies. It focuses on recognizing uppercase English letters (A–Z) and digits (0–9) from images. The OCR module processes images by converting them to grayscale, then binary, extracts features, and uses a trained neural network for classification. Recognized characters are saved as text files. The TTS module uses Microsoft's Win32 Speech API to convert this text into speech. It also supports direct e-text input for audio conversion. Simulation results confirm successful recognition and

speech output for various fonts. This cost-effective system aids visually impaired users and lays the groundwork for future expansion to word and sentence-level recognition.

The [12] research introduces a detailed image preprocessing procedure for Image-to-Text-to-Speech (ITTTS), which combines Optical Character Recognition (OCR) for text extraction with Google Text-to-Speech (gTTS) for audio assistance to visually impaired individuals. The procedure utilizes techniques such as rescaling, converting to grayscale, applying morphological transformations, and implementing thresholding to minimize noise and enhance recognition accuracy. Experimental results indicate a decrease in Character Error Rate, Word Error Rate, and Levenshtein Distance compared to approaches that do not include preprocessing. Future research plans involve broadening language compatibility, optimizing the system for various image qualities, and adapting the application for a range of uses.

The [13] paper reviews key OCR datasets for offline handwriting, online handwriting, and machine-printed text.CEDAR, CENPARMI, and NIST offer large samples of handwritten characters, digits, and words under realistic conditions. UNIPEN compiles diverse online handwriting data in a unified format for benchmarking. University of Washington datasets provide annotated English and Japanese documents with ground truth and degradation models.Design challenges include ensuring representative data, consistent segmentation, and robust training/testing splits.These datasets have greatly advanced OCR research, evaluation, and multilingual capabilities.

The [14] paper analyzes OCR performance, highlighting common error sources: imaging defects, similar symbols, punctuation, and typography.OCR accuracy has improved but still lags far behind human reading ability, often requiring manual correction. Large-scale ISRI tests reveal limitations and guide improvements through better image processing, font adaptation, and multi-character recognition. Linguistic context use can further enhance accuracy by leveraging word patterns, syntax, and semantics. The authors conclude that combining multiple strategies is the most promising path for advancing OCR systems.

Models for OCR based on Transformers, such as TrOCR, paired with sophisticated neural TTS that manages prosody and supports multiple languages, provide enhanced accuracy and more natural speech. Although Tesseract is straightforward and economical, the best strategy combines Transformers with neural TTS to achieve the highest levels of inclusivity, scalability, and real-time functionality.

III. INFERENCES CAPTIVATED

The evolution of OCR systems has been remarkable, presenting options with diverse capabilities and weaknesses. OCR-TTS systems utilize PyTesseract and pyttsx3 to deliver real-time accessibility for individuals with visual impairments by transforming images and videos into spoken words, although they have difficulties with handwritten or multilingual text and lack standardized evaluation criteria. Traditional tools like Tesseract are proficient with printed, organized documents but encounter obstacles with noisy, intricate, or handwritten formats. Contemporary Transformer-based models, such as TrOCR, surpass CNN-RNN methods in recognizing printed, handwritten, and scene text, attaining cutting-edge accuracy (96.6% F1 on SROIE, 2.89% CER on IAM). Transformers improve accuracy, inclusivity, and scalability, while OCR-TTS is essential for providing prompt accessibility.

IV. PROPOSED SYSTEM

The proposed solution improves accessibility by transforming text from images into spoken language, utilizing OCR for text extraction and TTS for audio production. It accommodates inputs such as images or scanned files (JPEG, PNG, PDF) and enhances readability with image enhancement methods, including noise reduction, skew correction, and contrast optimization. The OCR system detects text areas, separates characters, and retrieves raw text, which is subsequently cleaned, formatted, and organized into coherent sentences or paragraphs for additional applications, such as NLP tasks like summarization or automation (see Fig 2).

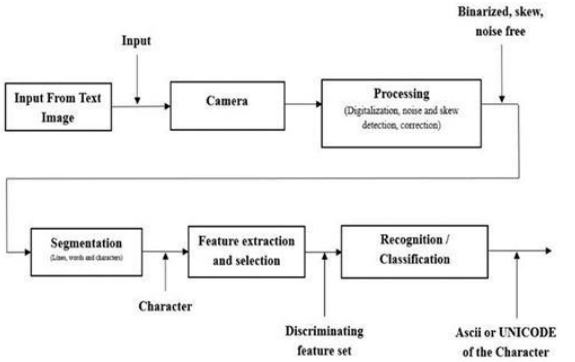

Figure 2 Architecture of the Project

The polished text is converted into natural-sounding audio using TTS technologies like GTTS, generating clear and understandable speech for users. An interactive dialogue interface facilitates real-time text or voice exchanges, allowing users to navigate the system, request particular actions, and receive voice responses, thus ensuring effortless accessibility, especially for those with visual impairments.

V. CONCLUSION AND FUTURE ENHANCEMENTS

This paper introduces an OCR-driven Text-to-Speech (TTS) system that aims to enhance accessibility for users with visual impairments by translating text extracted from images and videos into audible speech. The system utilizes Tesseract for extracting text and pyttsx3 for generating speech, consisting of modular components that handle input processing, image enhancement, text recognition, and audio output, thereby accommodating both printed text and video content for immediate reading. Future developments are planned to incorporate multilingual capabilities, recognition of handwritten text using Transformer-based OCR, options for voice customization, and enhancements in real-time performance, which will further enrich usability, inclusivity, and overall user experience.

REFERENCES

[1] M. R. Anusha, N. N. Bharadwaja, R. Indraja, and T. Raju, "OCR based image text to speech conversion," TKR College of Engineering and Technology, Telangana, India, 2024.

[2] M. Li, T. Chen, J. Chen, and Z. Li, "TrOCR: Transformer-based optical character recognition with pre-trained models," Beihang University and Microsoft Corporation, 2023.

[3] S. K. Garai, O. Paul, and U. Dey, "A novel method for image to text extraction using Tesseract-OCR," Dept. of CSE, University of Engineering & Management, Kolkata, India, 2024.

[4] X. Zhou, M. Zhang, Y. Zhou, Z. Wu, and H. Li, "Accented text-to-speech synthesis with limited data," *IEEE*, 2024.

[5] K. Soumya, D. Bhavana, P. Priyanka, and P. Roshini, "Handwritten text recognition using Pytesseract," *Double-Blind Peer Reviewed Refereed Open Access International Journal*, 2024.

[6] S. Chithra and N. Bhalaji, "Enhanced portable text to speech converter for visually impaired," Dept. of IT, SSN College of Engineering, Chennai, India, 2023.

A. Vijayanarayanan, R. Savithiri, P. Lekha, and R. S. Abbirami, "Image processing based on optical character recognition with text to speech for visually impaired," Dept. of CSE, PERI Institute of Technology, Chennai, India, 2023.

[7] W. C. Huang, T. Hayashi, Y. C. Wu, H. Kameoka, and T. Toda, "Voice Transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," Nagoya University, Japan, 2020.

[8] H. Dave, A. Gobse, A. Goel, and S. Bairagi, "OCR text detector and audio convertor," Dept. of EXTC, MPSTME, NMIMS University, 2020.

[9] N. Kotwal, A. Sheth, G. Unnithan, and N. Kadaganchi, "Optical character recognition using Tesseract engine," Dept. of Instrumentation Engineering, Vishwakarma Institute of Technology, Pune, India, 2021.

[10] C. S. Thu Thu and T. Zin, "Implementation of text to speech conversion," *International Journal of Engineering Research and Technology*, 2022.

[11] S. Reddy, R. R. Das, and A. Mohapatra, "An integrated pipeline with internal image processing for efficient image to text to speech conversion," 2023.

[12] Guyon, R. M. Haralock, J. J. Hull, and I. T. Phillips, "Datasets for OCR and document image understanding research," Seattle University, 2020.

[13] G. Nagy, T. A. Nartker, and Rice, *Optical Character Recognition: An Illustrated Guide to the Frontier*. New York, USA: Springer, 2019.