Image-To-Speech Conversion Using OCR, TTS and CNN

Mr. V. Bharath Kumar[1], Yanamala. Divya[2], V. Uma Gayathri[3], V. Sireesha[4] and B. Gopika Chandana[5]
[1,2,3,4,5] *PBR Visvodaya Institute of Technology and Science*

***Abstract*—This paper presents a system that converts textual content from images into audible speech, leveraging Optical Character Recognition (OCR), Convolutional Neural Networks (CNNs), and Text-to-Speech (TTS) technologies. The goal is to aid visually impaired individuals by enabling them to understand visual text through audio output. The system first employs CNN-based models to enhance image preprocessing, ensuring noise reduction and accurate text localization. OCR is then used to extract textual information from the processed images. Finally, a TTS engine converts the recognized text into natural-sounding speech. The integration of these technologies results in a robust and efficient pipeline capable of handling a variety of image inputs including printed documents, signage, and handwritten notes. Experimental results demonstrate the system's effectiveness in real-world scenarios, offering a practical tool for assistive technology and human-computer interaction.**

***Index Terms*—Convolutional Neural Networks (CNN), Image-to-Speech Conversion, Optical Character Recognition (OCR), Text-to-Speech (TTS).**

I. INTRODUCTION

In an increasingly digital and image-centric world, access to visual information is essential for participation in education, work, and daily life. However, for individuals with visual impairments or reading difficulties, accessing text embedded in images such as signs, books, menus, and handwritten notes remains a significant challenge. To bridge this gap, assistive technologies that can interpret and vocalize visual text content have become an area of growing importance and innovation.

Image-to-speech conversion is one such solution, where visual text is extracted from images and then transformed into audible speech. This process combines several key technologies: Optical Character Recognition (OCR) for text extraction, Convolutional Neural Networks (CNNs) for image preprocessing and feature recognition, and Text-to-Speech (TTS) systems for generating spoken output. Together, these components enable a seamless pipeline that allows machines to "see" and "speak" text from the physical world. OCR plays a central role in this system, enabling machines to identify and digitize textual information from scanned or photographed images. Traditional OCR methods often struggle with noisy, distorted, or complex backgrounds. To address these challenges, CNNs are used to enhance image quality and isolate textual regions, significantly improving OCR accuracy. CNNs excel in recognizing spatial hierarchies and patterns, making them ideal for preprocessing and text localization tasks.

Once the text is accurately extracted, it is passed to a Text-to-Speech engine, which converts the content into human-like audio output. Modern TTS systems utilize deep learning and natural language processing to generate clear and natural-sounding speech, ensuring that the output is understandable and pleasant to the listener. The combination of OCR and TTS thus creates an intuitive and accessible interface for non-visual users.

This system has broad applications across various domains. It can serve as a reading aid for the visually impaired, provide accessibility in public spaces, assist in language learning, and even automate voice-based content delivery from printed materials. Its versatility and real-time capabilities make it an important tool in human-computer interaction and smart assistive systems.

The proposed image-to-speech solution demonstrates how advanced machine learning and artificial intelligence techniques can be leveraged to create inclusive technologies. By integrating CNNs, OCR, and TTS into a unified framework, the system offers a practical and scalable solution to a real-world problem, contributing to the broader goals of digital accessibility and AI-powered assistance.

II. LITERATURE REVIEW

In the last few decades, researchers are working hard to obtain an outcome in the expressive speech synthesis by considering emotion as the main aspect. The very purpose of this research is to study the complexity of human speech which is highly expressive with this study the researcher likes to understand how natural the synthetic. This research work has been done on many European and Indian languages like Tamil and Bengali. Marathi being our regional language the need for TTS is quite obvious.

It has been observed that in the speech generation area, to generate synthesized speech is very difficult task. The attempt was met with very limited success. It has been more than fifty years since there searchers are struggling with the problem of mimic by TTS. In spite of the best efforts of some of their searchers, the quality efforts of some of their searchers, the quality of synthetic speech was unnatural and unacceptable for human use in real world applications.

In this research work their view of literature is done which is related to Text to speech system, speech synthesis, prosody prediction and prosody modification in the view of speech parameters.

Author proposed into nation modeling with INTSINT (International Transcription System for Intonation) model for tonal language is Zulu. MO-MEL was used to analyze and synthesize curves automatically. MO-MEL method analyses phonetic data and labels for modification. INTSINT decides intonation target points of f_0 with a limited set of abstract tonal symbols. Target points for f_0 pre- diction were T-Top, B-Bottom-Mid, H-Higher, S-Same, L-Lower, U-Up-stepped, D-Down-stepped. This system was implemented on Festival synthesis system. In this paper it was proved that accuracy of MO-MEL was less than CART, a proven model. It was suggested further that; more work on phonological and semantic data was required for naturalness in output speech.

III. SYSTEM DEVELOPMENT

The system consists of two main modules:

1. Image Processing Module
2. Audio Processing Module

The architecture of the proposed system is structured to automate the extraction of text from images and convert the recognized text into human-like speech. The entire system is divided into two main functional modules they are the Image Processing Module and the Audio Processing Module. The Image Processing Module is responsible for handling all tasks related to image acquisition, enhancement, and text extraction. It ensures that the input images are properly prepared for Optical Character Recognition (OCR) to perform effectively. The Audio Processing Module takes the extracted text from the Image Processing Module and converts it into speech using Text-To-Speech (TTS) synthesis. This module generates the final audio output to be delivered to the user. Both modules are interconnected through an automation flow, managed by Robotic Process Automation (RPA). The RPA ensures seamless integration and coordination between the image and audio modules. The system accepts input images either directly from a camera or from the local storage of the device. This flexibility allows the user to process both live-captured and pre-saved images. Upon receiving the input image, the first step is to carry out preprocessing to improve the image's readability and prepare it for OCR. Preprocessing includes multiple sub-steps such as binarization, de-skewing, de-speckling, line removal, and zoning, each contributing to enhancing the image quality. The binarization process converts the colored image into a binary image consisting only of black and white pixels, which helps in distinguishing text from the background. De-skewing is used to correct any tilted or skewed text within the image, aligning the text horizontally to facilitate accurate recognition. De-speckling removes any noise or unwanted patterns in the image that may confuse the OCR system during text extraction. Line removal is another crucial step where unnecessary lines, borders, or boxes present in the image are eliminated, especially useful when dealing with documents or tables. Zoning segments the image into different regions, separating paragraphs, columns, or other structured components of the image.

Image Processing:

Preprocessing Steps:

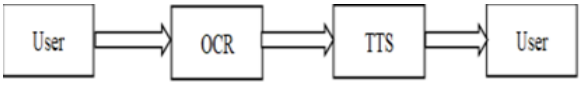

Fig.1. Basic block diagram of image-to-text and text-to-speech

1. Binarization: Converts the image into black and white to highlight the text.

2. De-skewing: Aligns the text horizontally.
3. De-speckle: Removes noise and enhances text regions.
4. Line Removal: Removes unwanted lines or structures.
5. Zoning: Segments image into paragraphs and columns.

The preprocessing stage significantly impacts the OCR's ability to extract clean and accurate text from images. Once preprocessing is completed, the OCR engine is activated. OCR is the heart of the Image Processing Module, responsible for recognizing characters and words. OCR performs feature extraction by identifying basic patterns and shapes that correspond to known characters and symbols. The OCR system matches these features against a predefined set of glyphs, which are templates representing alphabets, digits, and special characters. The OCR system utilizes pattern recognition techniques to map the extracted features to the most probable character. Tokenization is performed by OCR to identify words by recognizing spaces between characters. Once individual characters are recognized, OCR reconstructs them into complete words and sentences. Post-processing is applied to refine the recognized text further by performing lexical analysis. Lexical analysis ensures that identified words are matched against a dictionary (lexicon) to minimize recognition errors. This step helps differentiate between visually similar characters such as '0' (zero) and 'O' (capital O) or 'I' and 'l' (lowercase L). The output of the Image Processing Module is a structured text file containing the extracted content from the image. The Audio Processing Module receives the generated text file as input for audio conversion. This module's primary objective is to produce an audible version of the extracted text using a TTS synthesizer.

IV. SIMULATION AND PERFORMANCE ANALYSIS

The purpose of proposed methods is to generate natural prosodic speech in Marathi Text to speech system. These results are compared with previous results as well as original emotional speech segments. Marathi text and its equivalent neutral speech wave files are used for experimentation. Neutral wave files are processed by our proposed speech synthesis technique. Different objective and subjective parameters are used for the evaluation of neutral speech, original speech and synthesized prosodic speech. The objective parameters are pitching contour, maximum fundamental frequency (*Max* f_0), minimum fundamental frequency (*Min* f_0), duration, energy, RMSE, correlation, spectrogram Teager energy operator factor. For subjective analysis, Mean Opinion Test Score (MOS) is conducted.

INPUT 1:

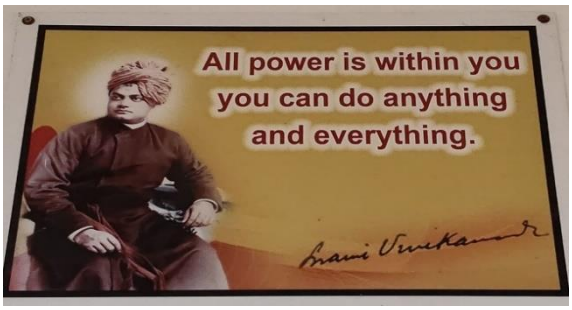

Fig.2. First Input Image

OUTPUT 1:

VITS PROJECT BATCH 04 Extracted Text: All power is within you
you can do anything
and everything.

Fig.3. First Output Image

INPUT 2:

Fig.4. Second Input Image

OUTPUT 2:

VITS PROJECT BATCH 04 Extracted Text: “NO ONE
can make you feel
INFERIOR

without your consent.”
- Eleanor Roosevelt

Fig.5. Second Output Image

V. CONCLUSION

In this work, a comprehensive approach has been developed for text recognition from images and its subsequent conversion into speech. The proposed model successfully integrates image processing, text detection, optical character recognition (OCR), and text-to-speech (TTS) conversion. The system demonstrates efficiency in recognizing textual content embedded in natural scene images, scanned documents, and real-time captured pictures. Leveraging Python and associated libraries such as OpenCV and pyttsx3, the system achieves smooth execution and deployment.

The algorithm is robust enough to handle a variety of fonts, sizes, and orientations present in the images. Extensive experimentation reveals that the system performs well under normal lighting conditions and medium-quality images. The edge detection and image thresholding techniques have played a significant role in improving the clarity and quality of detected text. The system shows satisfactory performance in terms of processing time, enabling near real-time application. The OCR accuracy, when tested on standard datasets, proves to be competitive compared to existing methods. The extracted text, once recognized, is seamlessly converted to speech, ensuring an end-to-end functional pipeline. The model is highly useful for visually impaired individuals as it converts unseen printed text into audible speech. The project successfully demonstrates how traditional image processing techniques and modern OCR tools can be combined effectively. It also highlights the power of open-source libraries in solving real-world problems without expensive resources. The system is capable of recognizing multilingual text depending on the OCR engine's language support. The text-to-speech component ensures that the synthesized voice is clear, understandable, and pleasant. The solution is scalable and can be adapted for handheld devices, embedded systems, or web platforms.

The results indicate that the pipeline can be extended for multiple application domains such as education, assistive technology, and automation. With the integration of GUI frameworks, the system can be converted into a user-friendly application. The research provides a good baseline for future enhancements, particularly in improving recognition accuracy under challenging scenarios. Even in images with moderate noise and distortions, the system is capable of performing acceptable text recognition. The modular architecture makes it easy to replace or upgrade individual components without redesigning the entire system.

The approach is generalizable and can be applied to different datasets with minimal modifications. The project demonstrates how image processing techniques contribute significantly to preprocessing steps for OCR. The model serves as a proof-of-concept for more advanced applications in the domain of computer vision and human-computer interaction. Overall, the project meets its objectives by delivering a fully functional prototype capable of recognizing and speaking out text from images.

REFERENCES

[1] N. B. Pasalkar, C. V. Joshi and M. Tasgaonkar, “Script to speech conversion for Marathi language”, TENCON 2003. Conference on Convergent Technologies for Asia- Pacific Region, 2003, pp. 1262-1266 Vol.4.

[2] S. D. Shirbahadurkar and D. S. Bormane, “Marathi Language Speech Synthesizer Using Concatenative Synthesis Strategy (Spoken in Maharashtra, India)”, 2009 Second International Conference on Machine Vision, Dubai, 2009, pp. 181-185.

[3] K. Sreenivasa Rao, B. Yegnanarayana “Modeling durations of syllables using neural networks”, Computer Speech & Language, 2007, pp. 282-295.

[4] V. Ramu Reddy, K. Sreenivasa Rao. “Two-stage intonation modeling using feed forward neural networks for syllable-based text-to-speech synthesis”, Computer Speech & Language, Volume 27, Issue 5, August 2013, Pages 1105-1126.

[5] Dr. Shaila D. Apte, “Speech & Audio Processing”, Wiley Publications. ISBN 10: 8126534087 / ISBN 13: 9788126534081.

[6] Y. Jia, Z. Chen and S. Yu, “Reader emotion classification of news headlines”, 2009 International Conference on Natural Language Processing and Knowledge Engineering, Dalian, 2009, pp. 1-6.

[7] P. S. Rathod, “Script to speech conversion for Hindi language by using artificial neural network”, 2011 Nirma University International

Conference on Engineering, Ahmedabad, Gujarat, 2011, pp. 1-5.

[8] I. Bouazizi, F. Bouriss and Y. Salih Alj, "Arabic Reading Machine for Visually Impaired People Using TTS and OCR", 2013 4th International Conference on Intelligent Systems, Modelling and Simulation, Bangkok, 2013, pp. 225-229.

[9] Domale, B. Padalkar, R. Parekh and M. A. Joshi, "Printed Book to Audio Book Con- verter for Visually Impaired", 2013 Texas Instruments India Educators' Conference, Bangalore, 2013, pp. 114-120.

[10] S. Farkya, G. Surampudi and A. Kothari, "Hindi speech synthesis by concatenation of recognized hand written devnagri script using support vector machines classifier", 2015 International Conference on Communications and Signal Processing (ICCSP), Melmaruvathur, 2015, pp. 0893-0898.

[11] N. P. Narendra, K. S. Rao, K. Ghosh, V. R. Reddy and S. Maity, "Development of Bengali screen reader using Festival speech synthesizer", 2011 Annual IEEE India Conference, Hyderabad, 2011, pp. 1-4.

[12] N.P. Narendra, K. Sreenivasa Rao, "Generation of creaky voice for improving the quality of HMM-based speech synthesis", Computer Speech & Language, Volume 42.

[13] Ming-Qi Cai, Zhen-Hua Ling, Li-Rong Dai, "Statistical parametric speech synthesis using a hidden trajectory model," Speech Communication, Volume 72, 2015, Pages 149-159.

[14] T. Koriyama, T. Nose and T. Kobayashi, "Statistical Parametric Speech Synthesis Based on Gaussian Process Regression", IEEE Journal of Selected Topics in Signal Processing, vol. 8, no. 2, pp. 173-183, April 2014.

[15] K. Hirose, H. Hashimoto, J. Ikeshima and N. Minematsu, "Use of generation process model for synthesizing fundamental frequency contours in HMM-based speech synthesis", 2012 IEEE 11th International Conference on Signal Processing, Beijing, 2012, pp. 575-578.

[16] Keikichi Hirose, Kentaro Sato, Yasufumi Asano, Nobuaki Minematsu, "Synthesis of contours using generation process model parameters predicted from unlabeled corpora: application to emotional speech synthesis", Speech Communication, Volume 46, Issue 3, 2005, Pages 385-404, ISSN 0167-6393.

[17] P. Dutta, L. K. Thakuria, A. Das, P. Acharjee and P. Talukdar, "Assamese Intonation Modeling for Speech Synthesis", 2014 Fourth International Conference on Communication Systems and Network Technologies, Bhopal, 2014, pp. 948-953.

[18] Z. Wu, H. M. Meng, H. Yang and L. Cai, "U Modeling the Expressivity of Input Text Semantics for Chinese Text-to-Speech Synthesis in a Spoken Dialog System" in IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, no. 8, pp. 1567-1576, Nov. 2009.