Journal of Science, Computing and Engineering Research (JSCER)
Volume-9, Issue- 3, Mar-Apr 2026.
DOI: *https:jscer.org/doi.org/10.46379/jscer.2025.080533*

VISION TO VOICE

[1]Dr. Anusuya,[2]Vanshika Joshi,[3]Vinay Kumar,[4]Shriya Jain,[5]Mohit Saini

[1]Professor, Department of CSE, Modern Institute of Technology and Research Centre, Rajasthan, India.

[2,3,4,5]UG Student, Department of CSE, Modern Institute of Technology and Research Centre, Rajasthan, India

Article Information

Received : 27 March 2026

Revised : 28 March 2026

Accepted : 29 March 2026

Published : 31 March 2026

Corresponding Author:

Vanshika Joshi

***Abstract*—** Visually impaired individuals often face significant challenges accessing essential product information such as ingredients, usage instructions, expiry details, and safety warnings independently. This project presents Vision-to-Voice, a web-based, AI-powered assistive system that leverages computer vision, Optical Character Recognition (OCR), Natural Language Processing (NLP), and Text-to-Speech (TTS) technology within a smartphone application. When a user scans a product using the camera, a Convolutional Neural Network (CNN) model detects the object and extracts printed text from the product label. The extracted text is processed to highlight relevant information such as usage guidelines, warnings, and key ingredients, which is then delivered as natural voice output through a speech synthesis engine. By combining intelligent recognition, flexible language processing, and a highly accessible interface, Vision-to-Voice enhances independence, convenience, and safety for visually impaired individuals.

Keywords: Accessibility, Assistive Technology, Visually Impaired, Product Recognition, OCR, Natural Language Processing, Text-to-Speech, Convolutional Neural Network, FastAPI.

Citation: : Dr. Anusuya, Vanshika Joshi, Vinay Kumar, Shriya Jain, Mohit Saini, "VISION TO VOICE", Journal of Science, Computing and Engineering Research, 9(3), March2026.

I. INTRODUCTION

The **Vision-to-Voice** project represents a sophisticated intersection of Artificial Intelligence (AI) and Assistive Technology (AT). In contemporary daily life, visually impaired individuals face significant challenges accessing product information printed on labels, including ingredients, usage instructions, expiry dates, and safety warnings. These details are essential for informed, safe decision-making, yet remain inaccessible without external help.

Our project is an innovative, targeted solution explicitly designed to bridge the gap between visual information and auditory delivery. The core capability of the system lies in leveraging computer vision, OCR, NLP, and TTS within a seamless end-to-end pipeline. By acting as an intelligent reading companion, the tool provides real-time, spoken product summaries without requiring any visual input from the user.

A. The Core Mechanism

The Vision-to-Voice system accepts two types of input: a live camera feed or a voice command. The system then executes a multi-stage pipeline:

- **Object Detection:** A CNN model identifies the product in the camera frame.
- **OCR Extraction:** Tesseract OCR extracts text from the product label.
- **NLP Summarization:** Hugging Face LLMs condense the extracted text.
- **TTS Output:** gTTS converts the summary to natural voice audio.

II. PROBLEM STATEMENT

The modern product-interaction process for visually impaired individuals has evolved into a significant source of dependency and frustration. This systemic barrier is responsible for the disheartening reality that millions of independent adults require third-party assistance for routine daily tasks.

A. Lack of Independent Access

Most product labels contain dense technical text printed in small, complex fonts on curved or reflective packaging surfaces. Even assistive magnifiers fail to capture entire labels accurately, leading to incomplete or misleading understanding of the product.

B. Dependency on Others

Due to inaccessible product information, visually impaired individuals are forced to depend on family members, shopkeepers, or strangers for routine tasks, reducing privacy and personal independence..

C. Inefficiency of Existing Solutions

Key limitations of current tools include:

- Manual camera alignment requirements
- Inability to handle curved or reflective labels
- No intelligent filtering or summarization of output
- Cloud-only processing with high latency
- Lack of multilingual support for Indian regional labels

III. PROPOSED MODEL

A. System Pipeline Overview

The core logic of Vision-to-Voice is built upon a sequential pipeline that transforms raw camera frames into spoken product information. This process is divided into four primary functional modules: Detection, Extraction, Summarization, and Audio Output.

B. Key Sub-processes in Detection:

- **CNN based Object Detection:** Utilizing TensorFlow and OpenCV, the system detects the product and identifies the label region.
- **Range Feedback:** If the product is out of range, the system audibly alerts the user: "Product not in range."
- **ARCore Depth API:** Measures distance between user and product to confirm the object is within a readable range.

C. Text Extraction and Normalization:

- **Tesseract OCR:** Multi-pass OCR extracts text from the detected label region, handling small fonts and curved surfaces.
- **Text Cleaning:** Removal of non-ASCII characters, promotional text, and marketing content irrelevant to safety or usage.
- **Preprocessing:** Contrast adjustment, denoising, skew correction, and segmentation improve OCR accuracy.

D. NLP Summarization:

Following extraction, Hugging Face LLM models and a Vector Database generate concise, meaningful summaries prioritizing safety-critical information such as warnings, ingredients, and usage instructions.

E. Proposed Work Model Diagram

The diagram below illustrates the complete end-to-end pipeline from user input through to speech output:

PROPOSED WORK MODEL

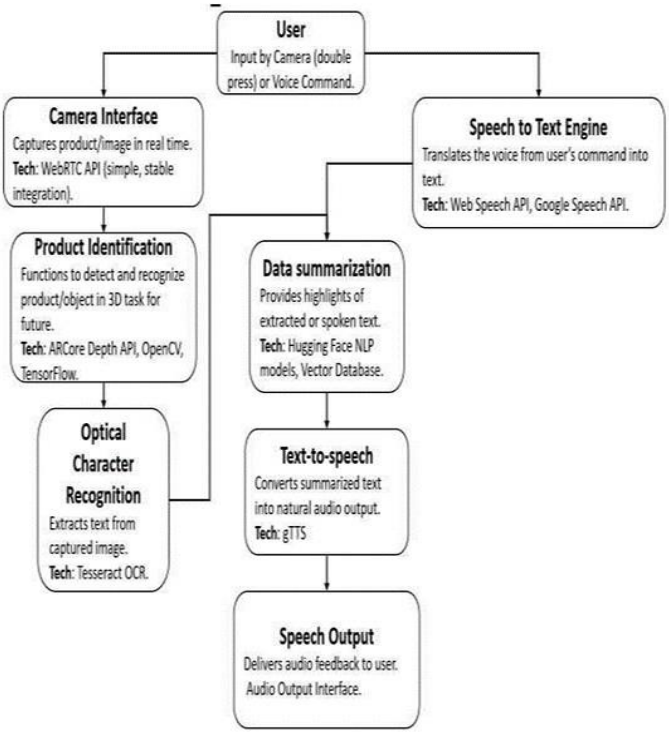

Fig. 1. Proposed Model

IV. TECH STACK

The selection of a technology stack is a critical architectural decision that determines the scalability, accessibility, and real-time performance of the system. For **VISION TO VOICE**, we adopted a hybrid stack that balances the agility of web frameworks with the computational power of modern AI libraries.

A. Frontend Technologies

The frontend is developed in **React with TypeScript**, providing a highly responsive, component-driven interface with minimal visual clutter. **WebRTC** handles live camera streaming, and **ARCore** enables depth-based range detection. The interface is intentionally accessible: a single tap activates the entire pipeline, with all feedback delivered through audio.

B. Backend Technologies

Python's FastAPI was chosen as the micro-framework for the API layer. Its lightweight, asynchronous nature is ideal for serving machine learning models in real time. The backend coordinates all processing modules: OCR, CNN, NLP summarization, and TTS.

C. AL / ML Technologies

- **Computer Vision:** TensorFlow + OpenCV for CNN-based product detection and label localization.
- **OCR Engine:** Tesseract OCR with multi-pass preprocessing for accurate text extraction from complex labels.
- **NLP Summarization:** Hugging Face LLMs with Vector Database for context-aware, concise summary.

D. Text-to-Speech

Google Text-to-Speech (gTTS) converts summarized text into natural-sounding audio. Its neural TTS engine produces smooth, human-like voice output, reducing listening fatigue during frequent daily use.

E. Database Technology

Firebase Storage is used for image upload, persistence of processing logs, and retrieval of product-specific data. Its cloud-based architecture ensures reliable access and future scalability.

V. RESULT SCREENSHOTS

The following screenshots demonstrate the complete end-to-end pipeline of the Vision-to-Voice system operating on a real product label.

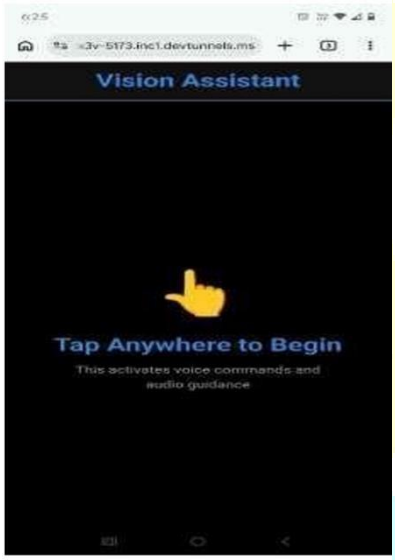

Fig. 2. Home Page

The landing page of the "Vision Assistant" web application features a dark-themed UI with a blue highlighted title and a single instruction: "Tap Anywhere to Begin."

Fig. 3. Voice Listening And Control Panel

After activation the system enters Listening mode. A microphone indicator confirms active voice recognition. Two primary buttons are displayed: Start Camera and Upload Image. Voice commands including "Open Camera," "Capture," "Retake," "Replay," and "Stop" are listed for hands-free interaction.

Fig. 4. Live Camera Capture Interface

The live camera preview shows real-time product detection. A VLCC Skin Nourishing Sunscreen Lotion SPF 40 PA+++ was detected, with an overlay instruction to tap or say "Capture." The interface supports both manual tap and voice-triggered image capture.

Fig. 5. Processing State Screen

After image capture, a loading spinner appears with "Processing…" text, indicating the backend OCR and AI summarization pipeline is executing. This screen reassures the user that the system is actively working.

Fig. 6. OCR Detection and Result Display

The processed image displays bounding boxes around detected text regions. The extracted label text is used to generate a summary identifying the product as a VLCC Natural Sciences sunscreen lotion with SPF 40 and PA+++, demonstrating successful OCR and summarization integration.

VI. CONCLUSION

A. Project Synthesis

The development of Vision-to-Voice represents a successful intersection of modern web architecture and advanced AI technologies. As we conclude this project in early 2026, the primary objective—to democratize product accessibility by providing visually impaired individuals with an independent, real-time, and reliable tool for reading product labels—has been fundamentally achieved.

The system successfully transforms live camera frames into structured voice output through a cohesive pipeline of CNN detection, Tesseract OCR, NLP summarization, and gTTS synthesis. By utilizing a modular, three-tier architecture, computationally heavy AI tasks remain decoupled from the user-facing interface, providing a seamless and responsive experience.

B. Core Achievements

The project has delivered on all its foundational promises:

- **Accessible Design:** A single-tap, voice-first interface requiring no technical knowledge, suitable for users of all age groups.
- **Intelligent Summarization:** Through Hugging Face LLM integration, the system provides concise, safety-prioritized product summaries rather than unfiltered raw text.
- **Real-time Performance:** Multi-pass OCR combined with FastAPI's asynchronous processing ensures low-latency audio responses suitable for in-store usage.
- **Inclusive Architecture:** The modular system design allows future expansion into multilingual OCR, allergen detection, and wearable device integration.

C. Concluding Remarks

In conclusion, Vision-to-Voice is not just a tool for reading labels; it is a tool for restoring independence. It helps visually impaired individuals interact confidently with everyday products, make informed safety decisions, and reduce reliance on others for routine tasks. As a B.Tech Computer Science project, this work has provided deep insights into the lifecycle of AI-driven assistive products—from data annotation and model training to API design and accessible UX deployment.

While the current version of Vision-to-Voice is robust, the rapidly evolving landscape of Large Language Models (LLMs) and on-device AI provides exciting avenues for future expansion including offline functionality, 3D product recognition, multilingual support, and wearable device integration.

REFERENCES

Research Literature and Documentation

[1] Using Computer Vision and Text-to-Speech for the Visually Impaired. Proceedings of ICT Innovations. Retrieved from https://proceedings.ictinnovations.org.

[2] Text to Speech Conversion Using Google Vision API. International Journal of Innovative Research in Technology (IJIRT), 2025.

[3] Visual Product Identification for Blind Peoples. International Journal of Creative Research Thoughts (IJCRT), 2025.

[4] Howard, A., et al. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv:1704.04861.

[5] Smith, R. (2007). An Overview of the Tesseract OCR Engine. Ninth International Conference on Document Analysis and Recognition (ICDAR).

[6] FastAPI. (2026). FastAPI Documentation (Version 0.100.x). Retrieved from https://fastapi.tiangolo.com/.

[7] Google Cloud. (2026). Text-to-Speech API Documentation. Retrieved from https://cloud.google.com/text-to-speech/docs.

[8] Hugging Face. (2026). Transformers: State-of-the-art NLP. Retrieved from https://huggingface.co/docs/transformers.

[9] Firebase. (2026). Firebase Storage Documentation. Retrieved from https://firebase.google.com/docs/storage.

[10] OpenCV. (2026). OpenCV Documentation: Computer Vision Library. Retrieved from https://docs.opencv.org/.

[11] TensorFlow. (2026). TensorFlow Object Detection API. Retrieved from https://www.tensorflow.org/hub.

[12] WebRTC. (2026). WebRTC API Reference. MDN Web Docs. Retrieved from https://developer.mozilla.org/en-US/docs/Web/API/WebRTC_API.