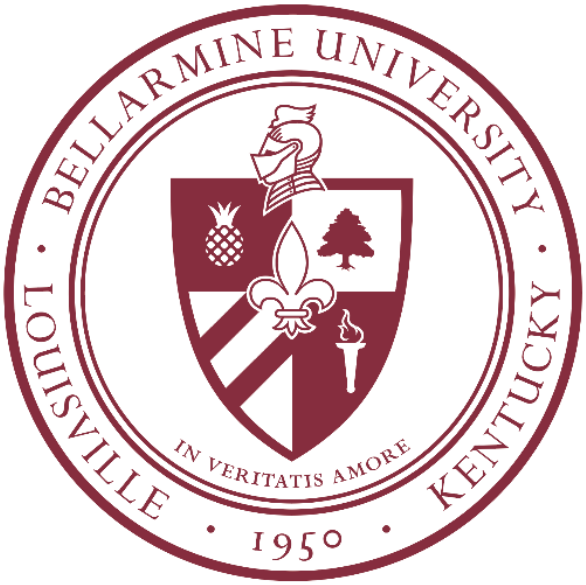
Automated Robotic Text and Speech Assistant



Jerrin C. Redmon

April 29, 2025

Bellarmine University: 2001 Newburg Rd, Louisville, KY 40205

ABSTRACT: ARTASA, a handheld robotic vision system, is designed to assist individuals who have difficulty reading due to blindness, dyslexia, or other vision impairments. Many people with these conditions struggle to determine whether an object contains text. ARTASA addresses these challenges by processing the environment, detecting and interpreting text, and then converting it into speech using text-to-speech software. The system utilizes a dedicated Serial Peripheral Interface (SPI) camera module designed to capture images at an optimal resolution and color balance, ensuring efficient processing by the Optical Character Recognition (OCR) system. A Raspberry Pi is employed as a dedicated computer to manage the overall software, enhancing both the speed and accuracy of the process. The system operates through three primary stages: image acquisition, image and text processing, and audio output. For optimal performance, it identifies text with a confidence level of 75% or higher before reading it aloud. The simple handheld device makes it easy to automatically scan objects, identify and verbalize text. This allows for a compact and adaptable design that can be deployed in various environments.

1. Introduction

Access to printed text remains a significant barrier for people with visual or cognitive impairments. Traditional assistive tools can be expensive or overly complex, limiting accessibility. ARTASA addresses this by providing an intuitive, embedded system capable of capturing text in real-time and converting it into clear speech output. It simplifies the user interaction to a single button press and automates all backend processes.

2. System Overview

The ARTASA system is composed of four primary components that work together to provide real-time text recognition and speech output. At the core of image acquisition is an ESP32 microcontroller paired with an Arducam Mega 5MP camera, which captures images upon button activation. These images are transmitted via serial communication to a Raspberry Pi 5, which handles the processing pipeline. The Raspberry Pi utilizes PaddleOCR, a deep learning-based optical character recognition engine, to extract text from the received images. Once the text is recognized, it is vocalized using Piper TTS, a lightweight, offline-capable text-to-speech engine, enabling immediate auditory feedback to the user.

3. Design and Hardware

The hardware configuration of the ARTASA system includes several integrated components designed for efficient image capture and user interaction. At the forefront is the ESP32 module equipped with an Arducam Mega 5MP camera, which serves as the primary imaging device. A tactile button, wired directly to the Raspberry Pi 5 running Pi OS, allows users to manually trigger image capture. Audio output is facilitated by an 8-ohm speaker connected to the Pi, enabling clear verbal feedback. All components are housed within a custom 3D-printed enclosure (see *Figure 1*), providing a compact and ergonomic form factor suitable for handheld operation.

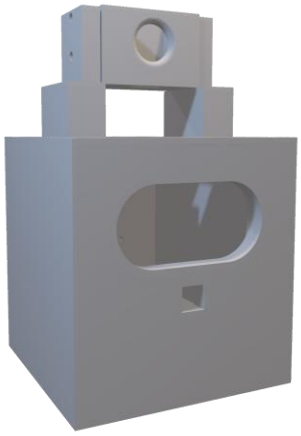


Figure 1: 3D Model of casing

3.1 Software Workflow

The software workflow of the ARTASA system is designed to efficiently process and vocalize textual information from captured images. The sequence begins with a user-initiated trigger event, where a button press activates the system. Upon activation, the ESP32 captures a JPEG image using the Arducam Mega 5MP camera. This image is then transmitted to Raspberry Pi via USB serial communication. Once received, the Raspberry Pi processes the image using PaddleOCR, a deep learning-based engine that extracts text. To ensure reliability, only text with a confidence score of 75% or higher is retained. Finally, the filtered text is vocalized through Piper TTS, which converts the recognized content into speech and outputs it through the connected speaker. A systemd .service file (artasa.service) was configured to run the Python OCR-to-speech script on startup. This ensures ARTASA operates autonomously after power-up.

3.2 Electrical Schematic

This schematic, designed in KiCad, demonstrates the physical wiring and logic levels for interfacing the Arducam module with the ESP32 UART pins. It includes decoupling, buzzer output for feedback, and clearly labeled GPIO usage, ensuring repeatable set up.

The full wiring schematic of ARTASA includes:

- An ESP32 wired to an Arducam module and Buzzer
- USB serial communication lines connecting to the Raspberry Pi
- Button connection between Raspberry Pi GPIO pins

4. Testing and Results

Testing was conducted using a vertically arranged sentence printed on paper, viewed at increasing distances. Confidence scores from PaddleOCR were recorded, (see *Figure 3).* OCR Confidence over Distance illustrates the performance and limitations of the ARTASA system's text recognition capabilities at varying distances. On average, the latency from image capture to speech output is approximately 2.5 seconds. The system is capable of reading fonts as small as 10 points under favorable lighting conditions, maintaining an accuracy range of 85–95% for high-contrast text. The implemented OCR confidence filter effectively eliminates low-quality reads, enhancing the reliability of spoken output. However, system performance diminishes significantly at distances beyond four feet. Additionally, motion blur and suboptimal lighting conditions adversely affect OCR confidence, underscoring the importance of stable positioning and adequate illumination during use.

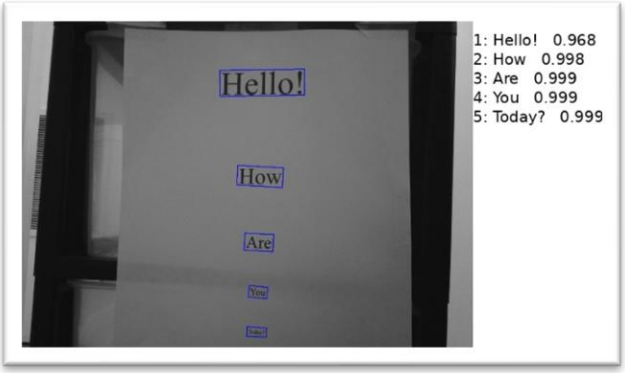


Figure 2: Image of OCR output

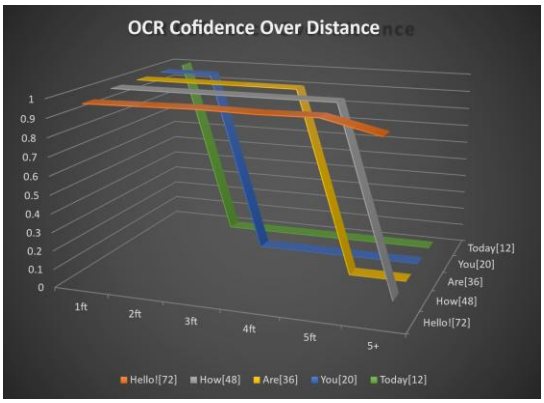


Figure 3: Graph of OCR Confidence

5. Conclusion and Future Work

ARTASA successfully demonstrates a functional, self-contained assistive reading system for visually impaired users. Its offline operation, simple interface, and modular architecture make it a compelling tool for real-world use. Future enhancements could include:

- Adding feedback to indicate OCR success/failure
- Incorporating translation support
- Deploying a higher-resolution camera module
- Building a custom ergonomic enclosure with tactile feedback

---

References

[1] PaddleOCR, *PaddleOCR: Optical Character Recognition with PaddlePaddle*. [Online]. Available: https://github.com/PaddlePaddle/PaddleOCR

[2] Piper TTS, *Piper: A fast, local neural text-to-speech system*. [Online]. Available: https://github.com/rhasspy/piper

[3] Espressif Systems, *ESP32-S3-DevKitC Hardware Reference*. [Online]. Available: https://docs.espressif.com/projects/esp-idf/en/latest/esp32/api-reference/index.html

[4] Arducam, *Arducam Mega 5MP OV5642 SPI Camera Module for Any Microcontroller*. [Online]. Available: https://docs.arducam.com/Arduino-SPI-camera/MEGA-SPI/MEGA-SPI-Camera/