Evolving Real-Time Audio Description Solutions for Visually Impaired Users: From HMD to Mobile Platforms

Claudio Roberto M. Mauricio
UNIOESTE
Foz do Iguaçu, Brazil
claudio.mauricio@unioeste.br

Gustavo Camargo Domingues
UNIOESTE
Foz do Iguaçu, Brazil
cdomingues.gustavo@gmail.com

Vitor Luiz Binsfeld Vieira
UNIOESTE
Foz do Iguaçu, Brazil
vitorlbvieira@gmail.com

Ivã Padua
UNIOESTE
Cascavel, Brazil
ivan.padua@unioeste.br

Fabiana Frata Furlan Peres
UNIOESTE
Foz do Iguaçu, Brazil
fabiana.peres@unioeste.br

João Marcelo X. N. Teixeira
Universidade Federal de Pernambuco
Recife, Brazil
joao.teixe@ufpe.br

***Abstract*—Assistive technologies for visually impaired individuals are crucial for enhancing their independence and quality of life. This paper discusses the evolution of a real-time audio description solution for visually impaired users from using a Head-Mounted Display to a more mobile and affordable platform utilizing a Raspberry Pi Zero 2W with a camera and battery. This second solution offers the same level of performance as the first, with added advantages of lower cost and enhanced mobility. The new system has been evaluated by the same visually impaired user from before, who provided valuable feedback on the user experience.**

***Keywords*—Visually Impaired, Mobile Platform, Accessibility**

I. Introduction

Visual impairment affects a significant portion of the global population, leading to challenges in navigation, social interaction, and access to information. Traditional assistive tools such as canes and guide dogs, while helpful, often fall short in providing comprehensive support. Considering technological advances, various academic research ([1]–[4]) and new commercial solutions (OrCam[1] and Envision Glasses[2]) have emerged. Unfortunately, these commercial solutions are accessible to only a minority of the blind population due to their high cost.

A previous study developed by us introduced a solution using the Meta Quest 3 HMD to generate real-time audio descriptions, enhancing the spatial awareness of visually impaired individuals through AI-driven software.

The main objectives of this work are to develop a more affordable and portable audio description system using the Raspberry Pi Zero 2W [3], evaluate its effectiveness in providing real-time audio descriptions, and compare its performance and user experience to the initial HMD-based solution.

The structure of this paper is as follows: Section 2 reviews related works in the field of assistive technologies for the visually impaired, particularly focusing on audio description systems and their applications. Section 3 provides an overview of the technology, highlighting the technical challenges and the proposed solution using the Raspberry Pi Zero 2W. Section 4 describes the experimental setup, participant profile, selected venues, and different audio description modes tested. Section 5 presents user feedback and insights from the evaluation. Finally, Section 6 concludes the paper with a discussion on the findings, implications, and future directions for further enhancing the system.

II. Related Works

Companies such as Verbit [4] and Amberscript[5] provide professional audio description services. Despite the advertisement found on their web pages, their main product is focused on audio transcription and captioning.

Hardware-based solutions that use real time visual information exist, though. The Rabbit R1[6] ChatGPT-based personal

[1]https://www.orcam.com/
[2]https://www.letsenvision.com/glasses
[3]https://www.raspberrypi.com/products/raspberry-pi-zero-2-w/
[4]https://verbit.ai/
[5]https://www.amberscript.com
[6]https://www.rabbit.tech/rabbit-r1

assistant is one of them. One of its functionalities, while interacting to the user by voice, is to answer questions about the image being captured by its camera at the moment.

Ai Pin[7], a product from hu.ma.ne, claims to be the first multimodal device acting as a "second brain", being able to make calls, send messages, seek answers, capture moments, take notes, and so on. In practice, it is very similar to the Rabbit R1 platform, with the difference it is presented in a smaller form factor and has no screen, as illustrated in Figure 1 (it has a micro projector for displaying content instead).

Fig. 1. Ai Pin (top) vs Rabbit R1 (bottom).
Source: https://humane.com/ and https://www.rabbit.tech/.

OrCam Solutions, designed for visually impaired users, offer a range of features that significantly enhance daily life. These devices, such as the OrCam MyEye (Figure 2), can read text aloud from any surface, recognize faces, identify products, and even detect colors, all through a small, discreet camera that attaches to eyeglasses. The strong points of OrCam include its portability, ease of use, and the ability to operate offline, ensuring privacy and functionality even without an internet connection. Additionally, the intuitive interface and voice-activated commands make it accessible for users with varying levels of tech-savviness. However, the high cost of OrCam devices can be a barrier for many potential users, and the need for regular software updates might be challenging for those less comfortable with technology. While the device is highly effective in controlled environments, it may struggle with accuracy in poorly lit areas or with heavily stylized fonts. Despite these limitations, OrCam Solutions represent a significant advancement in assistive technology, empowering visually impaired individuals with greater independence and confidence in navigating their surroundings.

[7]https://humane.com/aipin

Envision Glasses, designed for visually impaired users, offer a suite of advanced features that substantially improve everyday experiences. These smart glasses can read text aloud from various surfaces, recognize faces, describe scenes, and identify objects, leveraging the power of artificial intelligence.

One of the major strong points of Envision Glasses is their integration with Google Glass, providing a lightweight and stylish design that users can wear comfortably throughout the day (Figure 2). The device supports multiple languages and offers a companion app for further customization and ease of use. Moreover, Envision Glasses provide real-time updates and cloud connectivity, ensuring users have access to the latest features and improvements. However, the reliance on internet connectivity for some functionalities might be a drawback in areas with poor or no internet access. Additionally, the high price point can be prohibitive for many potential users, and some may find the battery life limiting for extended use. Despite these challenges, Envision Glasses stand out as a powerful tool for enhancing independence and accessibility for visually impaired individuals, bringing cutting-edge technology into their daily lives.

Fig. 2. OrCam MyEye 3 Pro (left) vs Envision Glasses (right).
Source: https://www.orcam.com/en-us/orcam-myeye-3-pro and https://www.letsenvision.com/glasses/home.

Table I provides a comparison between the four aforementioned solutions regarding price, form factor, if they require an active internet connection to work, battery time and presence of a screen. It is important to note that Rabbit R1 and AI Pin have some sort of visualization solutions (screen and projector)

TABLE I
COMPARISON BETWEEN THE AVAILABLE SOLUTIONS.

	Rabbit R1	Ai Pin	OrCam MyEye 3 Pro	Envision Glasses
Price	∼$500	∼$150	∼$4,500	∼$3,500
Size	Small, handheld device	Small, wearable device	Small, attaches to glasses	Small, (Google Glass-based)
Has a screen	Yes (OLED display)	Yes (projection)	No	Yes (Head-up display)
Requires Internet	Yes	Yes	No	Yes
Features	Text reading, object identification, navigation assistance	Text reading, voice control, smart assistant	Text reading, face recognition, product identification, color detection	Text reading, face recognition, object identification, scene description, real-time updates
Battery Time	Up to 12 hours	Up to 8 hours	Up to 6 hours	Up to 8 hours

because they were not originally designed for visually impaired users. Envision Glasses, on the other side, has a screen because it was originally based on the Google Glasses platform.

When it comes to free software, Be My Eyes[8] is by far the closest solution to the proposed work. First launched in 2015 as a free app, it allows volunteers to assist blind and low-vision users. According to their website, Be My Eyes users can request assistance in over 180 languages making the app the biggest online community for blind and low-vision people as well as one of the largest micro-volunteering platforms in the world.

On September 25, 2023, OpenAI launched its GPT-4 model with vision capabilities, often referred to as GPT-4V. The updated model was capable of analyzing and interpreting image inputs in addition to text, marking a significant advancement in its multimodal capabilities. As of May 13, 2024, OpenAI launched its GPT-4o model[9], now capable of processing and generating outputs across multiple modalities - text, audio, and image - in real time. It integrates various input and output types under a unified model, providing rapid response times comparable to human reactions and improved performance on non-English languages, and that's how Be My Eyes comes back to the stage. Its new feature, Be My Eyes' Virtual Volunteer (currently under tests), will be able to describe scenes in real time for visually impaired users without the need of human volunteers at all, as shown in this online video[10].

In the video, it is possible to see the user holding his phone and asking questions to it related to scene understanding. Inspired by that possibility, we decided to perform an experiment to validate the possibility of using HMDs to capture first person views and use the newest ChatGPT models to describe them, in real time. This would augment the human senses using hearing as input to compensate the lack of vision. One of the main advantages of using an HMD to do so is the ability of not having to hold anything on hands and also capture an image closer to the user's real field of view. The technicalities of the proposed solution/experiments will be described in the next sections.

While this section provides a comprehensive overview of current academic and commercial advancements in real-time audio descriptions for visually impaired individuals, it is important to note that the field continues to evolve rapidly with ongoing research and development. Review papers, such as those referenced here, serve as valuable resources for identifying trends, gaps, and emerging technologies in the domain of assistive technologies for visual impairments [5]–[9]. As technology progresses and new methodologies are explored, further investigations and collaborations across interdisciplinary fields will be essential in pushing the boundaries of what is achievable in providing inclusive and effective solutions for the visually impaired community.

III. TECHNOLOGY OVERVIEW

The new system utilizes the Raspberry Pi Zero 2W, a compact and cost-effective computing platform, combined with a camera and a portable battery, as shown in Figure 3.

The Raspberry Pi Zero 2W setup includes a camera module for capturing visual data and a battery pack for portability. Instead of running the AI model locally, the captured image is sent to the cloud where AI processing is performed. This cloud-based approach allows for more powerful and sophisticated image processing and audio description generation. The audio output is then delivered back to the user through Bluetooth headphones, ensuring a seamless and unobtrusive user experience.

Fig. 3. Main components of the proposed solution.Source: Authors.

[8]https://www.bemyeyes.com/
[9]https://openai.com/index/hello-gpt-4o/
[10]https://www.youtube.com/watch?v=Zq710AKC1gg

The transition from the Meta Quest 3 HMD to the Raspberry Pi Zero 2W platform involved overcoming several technical challenges. These included ensuring sufficient processing power on the limited hardware, changing the original python code to access the native raspberry pi zero 2w camera, and integrating the components into a compact and wearable form factor.

The process of integrating audio description capabilities into the Raspberry Pi Zero 2W involves two main phases: initialization and looping. In the initialization phase, the system is powered up and the python code is executed automatically. In the loop phase, images are captured every second and the system decides if it must be processed according to user input.

Since both image processing and audio generation are performed in the cloud, there is a need for internet connection (through wifi). Also, the audio output is done using bluetooth headphones, so that these steps are done as soon as the raspberry wakes up.

We use Raspberry Pi Imager to generate a custom Raspberry Pi OS (64bit) already with information regarding wifi ssid and password and SSH connections enabled. When running the system for the first time, we connect to it through SSH and install the required Python packages (venv, Picamera2, Openai). Also, we assemble a startup script that runs a command line bluetooth command to automatically connect to the bluetooth headphone (bluetoothctl connect MAC_ADDRESS) and than call the python script of the application.

After the developed Python script is started[11], it executes the following steps in a loop:

- It captures an image from the native raspberry pi zero 2w camera using Picamera2 library. We use as trigger the amount of brightness of the image. Based on a luminance threshold, we decide if the next steps of the audio description processing should be executed, or if a new image must be captured. Whenever the user wants to take a photo to be processed, he must cover the camera with his hand. This makes the image being captured very dark, which indicates to the system that the next bright image should be sent for processing.
- The captured image is converted to Base64 encoding to be transmitted to the ChatGPT API. This is done using the cv2.imencode (from OpenCV) and base64.b64encode functions to get the image bytes and encode them to Base64, respectively. A prompt is then created providing both system and user roles, and a request is made to the API.
- After receiving the text response from the ChatGPT API, another request is made to convert the text into voice. For this, the text-to-speech model (tts-1) is selected along with one of the five available voices (in this case, the "alloy" voice), and the text is provided as input to the API.
- The final phase of the iteration is to play the received MP3 file to the user. Since the platform is already connected to the Bluetooth headphones, we simply start mpg123 music player passing as parameter the mp3 file name of the corresponding audio description downloaded. After playing the audio, the process starts again by capturing a new image.

IV. Experiments

The single participant in the experiment is a 37-year-old male who lost his vision at the age of eight due to glaucoma, retaining few visual memories such as primary colors and the faces of his parents. He has extensive experience with various assistive technologies for the visually impaired, having started using the DosVox[12] system long ago. His familiarity with these technologies has made him particularly insightful in evaluating and appreciating advancements in accessibility solutions, such as the real-time audio description system integrated with the Meta Quest 3 and ChatGPT-4o. After becoming acquainted with the research, he was invited to contribute as a co-author to this work. His feedback reflects a deep understanding of how technologies can enhance independence and quality of life for blind individuals.

A. First Solution Experiment

To evaluate the effectiveness of the audio description system integrated with Meta Quest 3, we conducted experiments using four different variations of audio description prompts. For all experiments, the system role was set as: "You are a person that provides professional audio description services. Help me describe the images I show you in Brazilian Portuguese". The experiments aimed to test different aspects of audio description, including general scene description, aiding locomotion, visual characteristics of people, and reading textual content.

The chosen venue for the experiments was a real-world setting — a university exhibition of student projects combined with a café/restaurant environment within the event. This setting was selected deliberately to replicate scenarios from the participant's daily life, offering an environment without controlled conditions where audio description could be effectively applied. The venue's dual nature allowed for testing in both crowded and relatively quieter areas, providing valuable insights into how the audio description system could assist in navigating varied social and practical situations. This choice aimed to ensure that the experimental outcomes were not only technically sound but

[11] https://github.com/Void-NullPointer/AudioDescription-for-blind-people

[12] https://intervox.nce.ufrj.br/dosvox/

also relevant and impactful in real-world settings commonly encountered by visually impaired individuals.

1) General Audio Description:

User Prompt: "Please describe this image."

In the General Audio Description mode, the system was tasked with describing the overall scene captured by the camera. This included identifying objects, people, and activities within the environment, providing visually impaired participants with a comprehensive understanding of their surroundings. The aim was to evaluate how well the system could convey contextual information that would be useful in everyday settings. Two samples of image inputs and the corresponding audio description results are listed in Figure 4.

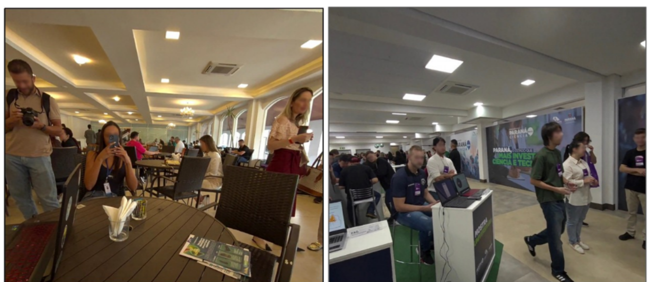

Fig. 4. General audio description samples. Source: Authors.

2) Locomotion Aid:

User Prompt: "Please describe this image focusing in aiding locomotion for a blind person. I want to know if the way ahead is free for walking. Describe the scene to help me to walk with more confidence, knowing if there is any obstacle in front of me. Be brief, I want a short sentence telling me only the necessary to step forward."

The Locomotion Aid mode focused on assisting participants in navigating through the environment. The system provided real-time audio cues to help avoid obstacles, identify pathways, and guide the participants safely from one location to another. This mode aimed to assess the system's capability to enhance the mobility and independence of visually impaired individuals. Two samples of image inputs and the corresponding audio description results are listed in Figure 5.

3) Visual Characteristics of People:

Fig. 5. Audio description samples for locomotion aid. Source: Authors.

User Prompt: "Please describe this image. Do not start the phrase with 'A imagem '. Describe the person in the center of the image. Focus on the facial features. Describe as gently as you can. I'm blind and want to know the person by her face."

This experiment aimed to provide descriptions focusing on the visual characteristics of individuals within the user's field of view. Two samples of image inputs and the corresponding audio description results are listed in Figure 6.

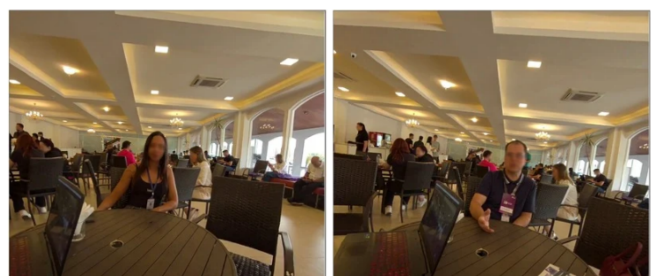

Fig. 6. Person description experiment samples. Source: Authors.

4) Reading Textual Content:

User Prompt: "Please read any textual content in front of you."

The goal of this prompt was to capture and relay any textual content present in the user's view, providing a valuable service for visually impaired users needing access to written information. Two samples of image inputs and the corresponding audio description results are listed in Figure 7.

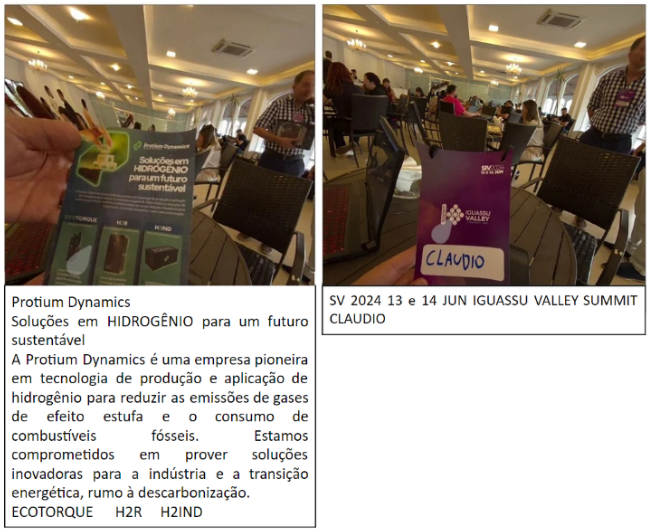

Fig. 7. Textual content-related experiment samples. Source: Authors.

B. Second Solution Experiment

To evaluate the effectiveness of the audio description system integrated with the Raspberry Pi Zero 2W with a camera, we conducted experiments using only two different modes, as previously mentioned: General Audio Description and Locomotion Aid. It is worth noting that the prompts and role were not changed in order to maintain similar and comparative results. The experiments aimed to test different aspects of audio description, specifically focusing on providing general scene descriptions and aiding in locomotion through varied environments, as proposed in the first experiment.

The chosen venue for the second experiment was a university building, encompassing a variety of settings such as rooms, laboratories, hallways, and other common areas, including outdoor sidewalks. This venue was selected to provide a realistic and diverse environment, mirroring the complexity of spaces that visually impaired individuals might encounter in their daily lives. This choice allowed us to test the system in different contexts within the same building, from quiet and controlled laboratory environments to more dynamic and unpredictable hallways and common areas. By conducting the experiments in such a setting, we aimed to gather insights into the system's performance in both structured and unstructured spaces, offering a comprehensive evaluation of its practical applicability.

1) General Audio Description:

As detailed in Section IV-A1, the objective was to obtain a general audio description of the scene being visualized. This prompt was designed to capture the overall context and key elements within the user's field of view. In the Figure 8, two different situations are presented: one shows the description of a busy corridor, and the other just the description of part of a room.

Fig. 8. General Description of an internal corridor (left) and of part of a room (right). Source: Authors.

2) Locomotion Aid:

As explained in Section IV-A2, this mode is specifically focused on providing descriptions that aid in locomotion, helping the user understand if the path ahead is clear and safe for walking. Figure 9 presents two results collected in different contexts, one in an outdoor environment and the other in an indoor environment.

V. User Feedback

In the feedback session, the visually impaired participant also shared his experiences and perspectives on using the real-time audio description solution, as he did with the first solution experiments. Figure 10 illustrates the user participating on both experiments (first solution (left) and second solution (right)).

The feedback from the visually impaired participant highlighted the enhanced mobility and convenience of the new solution. The user appreciated the lightweight and portable

Fig. 9. Locomotion aid in an outdoor setting (left) and indoor setting (right) using the Raspberry Zero 2 W solution. Source: Authors.

Fig. 10. Visually impaired person during experiments: using the HMD-based solution (left) and the Raspberry-based one (right). Source: Authors.

design, which allowed for greater freedom of movement compared to the HMD-based system. Additionally, the choice of using hand gestures to capture images and perform audio description was praised, emphasizing the importance of keeping the visually impaired person's hand free. The accuracy and detail of the audio descriptions were commended, affirming that the new setup maintained the high standards of the previous solution. Suggestions for further improvements included longer battery life, additional customization options for audio settings, and different interactions to switch between assistance and description modes

VI. Conclusion

This study demonstrates the successful evolution of our real-time audio description system into a more mobile and affordable solution using the Raspberry Pi Zero 2W. The new platform retains the effectiveness of the initial HMD-based approach while offering significant advantages in terms of cost and portability. Future work will focus on refining the hardware integration, extending battery life, and exploring additional applications to further enhance the accessibility and usability of the system for visually impaired individuals. We also intend to use a more dedicated hardware, developing a solution based on an ESP32 platform, to reduce even more the final commercial price to the general public. This will certainly bring new challenges due to the processing and memory restrictions of such platforms, but will allow the miniaturization of the solution as well, for less than half of the current raspberry pi zero based one.

References

[1] B. N. Mohapatra, R. Mohapatra, and P. Panda, "Path guidance system for blind people," *International journal of open information technologies*, vol. 7, no. 5, pp. 29–32, 2019.

[2] D. T. Yulianto, H. Maulana, and T. N. Nizar, "Smart stick as an assistive device for blind people using internet of things," in *2023 9th International Conference on Signal Processing and Intelligent Systems (ICSPIS)*. IEEE, 2023, pp. 1–7.

[3] A. Pardasani, P. N. Indi, S. Banerjee, A. Kamal, and V. Garg, "Smart assistive navigation devices for visually impaired people," in *2019 IEEE 4th international conference on computer and communication systems (ICCCS)*. IEEE, 2019, pp. 725–729.

[4] M. M. Soto-Cordova, F. Criollo-Sánchez, C. Mosquera-Sánchez, and A. Mujaico-Mariano, "Prototype of an audible tool for blind based on microcontroller," in *2020 IEEE Colombian Conference on Communications and Computing (COLCOM)*. IEEE, 2020, pp. 1–4.

[5] Y. Li, K. Kim, A. Erickson, N. Norouzi, J. Jules, G. Bruder, and G. F. Welch, "A scoping review of assistance and therapy with head-mounted displays for people who are visually impaired," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 15, no. 3, pp. 1–28, 2022.

[6] E. Gkini, I. Voyiatzis, and C. Sgouropoulou, "Head-mounted display systems as visual aids for the visually impaired: a survey," in *Proceedings of the 25th Pan-Hellenic Conference on Informatics*, 2021, pp. 323–327.

[7] H. M. Htike, T. H. Margrain, Y.-K. Lai, and P. Eslambolchilar, "Ability of head-mounted display technology to improve mobility in people with low vision: A systematic review," *Translational Vision Science & Technology*, vol. 9, no. 10, pp. 26–26, 2020.

[8] L. Hakobyan, J. Lumsden, D. O'Sullivan, and H. Bartlett, "Mobile assistive technologies for the visually impaired," *Survey of ophthalmology*, vol. 58, no. 6, pp. 513–528, 2013.

[9] P. Xu, G. A. Kennedy, F.-Y. Zhao, W.-J. Zhang, and R. Van Schyndel, "Wearable obstacle avoidance electronic travel aids for blind and visually impaired individuals: A systematic review," *IEEE Access*, vol. 11, pp. 66 587–66 613, 2023.