frontiers | Frontiers in Artificial Intelligence

TYPE Brief Research Report
PUBLISHED 25 October 2024
DOI 10.3389/frai.2024.1424371

OPEN ACCESS

EDITED BY
Samiya Khan,
University of Greenwich, United Kingdom

REVIEWED BY
Diego Vilela Monteiro,
ESIEA University, France
Luigi Di Biasi,
University of Salerno, Italy

*CORRESPONDENCE
Theodoros G. Soldatos
✉ theos.soldatos@gmail.com

RECEIVED 27 April 2024
ACCEPTED 04 October 2024
PUBLISHED 25 October 2024

CITATION
Ates Ö, Pandey G, Gousiopoulos A and Soldatos TG (2024) A brief reference to AI-driven audible reality (AuRa) in open world: potential, applications, and evaluation.
Front. Artif. Intell. 7:1424371.
doi: 10.3389/frai.2024.1424371

A brief reference to AI-driven audible reality (AuRa) in open world: potential, applications, and evaluation

Ömer Ates[1], Garima Pandey[1], Athanasios Gousiopoulos[2,3] and Theodoros G. Soldatos[1]*

[1]School of Information, Media, and Design, SRH Hochschule Heidelberg, SRH University of Applied Science, Heidelberg, Germany, [2]Department of Library, Archives and Information Systems, School of Social Sciences, International Hellenic University, Thessaloniki, Greece, [3]Department of Accounting and Information Systems, School of Economics and Business Administration, International Hellenic University, Thessaloniki, Greece

Recent developments on artificial intelligence (AI) and machine learning (ML) techniques are expected to have significant impact on public health in several ways. Indeed, modern AI/ML methods have been applied on multiple occasions on topics ranging from drug discovery and disease diagnostics to personalized medicine, medical imaging, and healthcare operations. While such developments may improve several quality-of-life aspects (such as access to health services and education), it is important considering that some individuals may face more challenges, particularly in extreme or emergency situations. In this work, we focus on utilizing AI/ML components to support scenarios when visual impairment or other limitations hinder the ability to interpret the world in this way. Specifically, we discuss the potential and the feasibility of automatically transferring key visual information into audio communication, in different languages and in real-time—a setting which we name '*au*dible *re*ality' (AuRa). We provide a short guide to practical options currently available for implementing similar solutions and summarize key aspects for evaluating their scope. Finally, we discuss diverse settings and functionalities that AuRA applications could have in terms of broader impact, from a social and public health context, and invite the community to further such digital solutions and perspectives soon.

KEYWORDS

digital health, public health, object recognition, text to speech, visual aid companion, vision impairment, real world decision support, biomedicine and healthcare informatics

1 Introduction

Recent artificial intelligence (AI) and machine learning (ML) developments are expected to significantly impact public health. Applications range from drug discovery and disease diagnostics to personalized medicine, healthcare operations, and evidence-based real-world (RW) analytics (e.g., Bohr and Memarzadeh, 2020; Vamathevan et al., 2019; Schneider et al., 2020; Lee and Yoon, 2021; Goecks et al., 2020; Elemento et al., 2021; Soldatos et al., 2022; Soldatos et al., 2019; Liu et al., 2023; Brock et al., 2022; Brock et al., 2023; Mullowney et al., 2023; Wang et al., 2024; Olawade et al., 2023). While AI/ML advancements may increase access to health services and improve the quality-of-life for many, challenges may persist for some, particularly in emergency situations (e.g., Kuriakose et al., 2023; Chenais et al., 2023; Grant et al., 2020; Ahmed et al., 2023). One such group that can benefit from modern AI solutions on computer vision and text-to-speech (TTS) technologies is visually impaired individuals (e.g., Kuriakose et al., 2023).

Computer vision is concerned with the development of algorithms and techniques that allow machines to analyze, process, and interpret digital images and videos. A small list of tasks pertaining image/video processing are listed in Table 1 (upper part). For example, in recent years object detection/recognition (ODR) is increasingly leveraged to assist in visually navigating environments (e.g., in autonomous vehicles; Tapu et al., 2017; Masmoudi et al., 2019; Malligere Shivanna and Guo, 2023). However, relying solely on visual input can be limiting for individuals with significant visual impairments, as it may not provide a comprehensive understanding of their surroundings. To address this challenge, researchers have explored the use of alternative modalities, such as of audio (e.g., Park and Fine, 2023; Maimon et al., 2023; Shvadron et al., 2023; Gamage et al., 2023; Neugebauer et al., 2020), which can enhance ODR and improve the experience and independence of visually impaired persons.

TTS technology is one such modality that can provide auditory support to users of ODR (e.g., Hao et al., 2024; Orynbay et al., 2024; Hemavathy et al., 2023), or (Pooja et al., 2024). Typically, TTS is used in daily digital communication to convert written text into spoken words, making it easier and faster to consume information (by simultaneously hearing the words). This technology has many practical applications (see Table 1, lower part), including helping people who are unable to read (or have difficulty in reading) to access written content.

Fortunately, recent advancements in modern deep learning (DL) techniques have improved our ability to perform these tasks. Recent developments in ODR technology use improved DL models such as

TABLE 1 Common image/video processing tasks and popular of text-to-speech (TTS) applications.

Common image/video processing tasks	
Task	**Description of goal**
Object detection (or object recognition)	To determine what objects are present. Given an input image, class labels and probabilities of the objects likely contained in that image are extracted (see, Figure 1). Boxes determining the position of these objects in the image may also be extracted.
Image classification	To assign a label or class to an entire image. Given an input image, a prediction about which class the image belongs to is returned.
Segmentation	To assign labels to the pixels of an image. Specifically, semantic segmentation is the process of classifying every pixel in an image into a specific class or category, instance segmentation is the process of detecting individual objects in an image (while also distinguishing between objects of the same class), whereas panoptic segmentation is used to assign a unique label to every pixel in an image (with pixels that do not belong to any object instance being assigned to 'background').
Text information	To detect regions in an image which contain text. The text mentioned in these regions may be processed further to extract the mentioned information (e.g., speed limit from a sign board).
Action recognition	To recognize which (human) actions or activities are performed, in which sequence of frames, in which time interval, and in which location in the scene the acting person(s) is/are. Spatio-temporal action detection is focused on determining the regions and times of an action class, whereas skeleton-based action detection is concerned with capturing (human) actions as represented by motions of skeleton.
Object tracking	To trace the location of objects in images or video frames. Single object tracking is concerned with a single object throughout a video sequence, while multiple object tracking involves tracking multiple objects simultaneously in the same video sequence. In single object tracking, the goal is to maintain the identity of the object being tracked over time, whereas in multiple object tracking, the goal is to simultaneously track and maintain the identities of multiple objects in the same scene. In a sense the goal is to trace the instances of object(s) appearing across different frames, even when undergoing changes in appearance, orientation, and movement.
Popular applications of TTS technology	
Application	Description of scope
Voice assistants	Software to execute tasks via voice command. Some of the most popular virtual voice assistants include Apple's Siri (Siri, 2024), Amazon's Alexa (Amazon Alexa Voice AI | Alexa Developer Official Site, 2024), Microsoft's Cortana (Cortana, 2024), and Google's Assistant (Google Assistant, 2024). Smart TTS technology ensures smooth and efficient communication between users and application, providing a more natural and human-like interaction between them and the used devices.
Customer service automation	Interactive voice response systems may combine pre-recorded messages or TTS technology to engage with customers, allowing them to provide and access information without employing a human agent.
Education and learning	TTS technology can help education and learning in various ways. It can make reading and understanding text easier for individuals who struggle with traditional reading methods, such as those with dyslexia or visual impairments. TTS can also help learners with limited reading proficiency to access digital content, including textbooks and online resources, as well as improve their pronunciation and fluency in a foreign language. In addition, TTS can provide a multi-sensory learning experience, which can be helpful for learners with different learning styles.
Audio books	Converting written text (such as books, news, magazine articles or websites) into spoken words can be particularly helpful for those with visual impairments or reading difficulties.
Navigation tools	Provision of instructions and directions in voice to users (e.g., to a driver or to a pedestrian) or reading out the names of streets and landmarks makes it easier to navigate without having to look at a map or screen. Additionally, TTS can be used to alert the user of upcoming turns, changes in direction, and other important information related to navigation.
Travel and tourism	TTS technology can help travelers to cope with real-time information (e.g., travel announcements in airports and train stations) by providing audio narratives translated in the language of their choice.

R-CNN, SSD, and YOLO that are more accurate and faster (see Supplementary Table 1). These models are trained on large datasets of annotated images, such as the COCO (Common Objects in Context; COCO, 2023) and the ImageNet (ImageNet, 2023) collections and are capable of detecting objects with high accuracy in real-time (RT). Characteristically, the COCO dataset contains photos of almost hundred object types, whereas the full ImageNet contains 20 K+ categories (see COCO, 2023; ImageNet, 2023; Lin et al., 2015; Russakovsky et al., 2015; Deng et al., 2009). Moreover, several ODR and TTS algorithms are available today as libraries of commonly used programming languages (Supplementary Table 2 lists some such popular ODR and TTS options; in Python; Python.org, 2023). Several of those libraries offer a set of pre-trained models for ODR making it easier nowadays for developers to implement own algorithms and custom applications [e.g., TensorFlow (2023) or OpenCV (2024)]. Similarly, TTS libraries make it easier to generate speech from text data by offering a variety of features, including the ability to customize voice parameters, adjust speaking rate, and control pitch and volume, as well as conversion in multiple languages [e.g., like pyttsx3 (Bhat, 2021); gTTS (Durette, 2022), and espeak (Asrp, 2024)], making them versatile tools for developers who need to create speech-enabled applications.

Main ways to combine ODR and TTS into integrated speech synthesis systems of spoken descriptions include the:

- Stepwise approach: first using ODR models that output bounding boxes, and then using TTS modalities to convert object labels into speech, or
- Descriptive approach: using ODR models that output detailed information about objects (such as size, shape, or color) and then use TTS systems to generate more detailed spoken descriptions, or
- Hybrid approach: creating single, end-to-end models that are specifically trained to directly output spoken descriptions of detected objects, eliminating the need for combining separate components; this approach can built on the 'multi-modal' capabilities that more recent AI increasingly enables—allowing to input one modality (e.g., video or text) and output another (e.g., image or audio).

In comparison, the first approach is more straightforward and easier to implement, but the spoken descriptions may be limited to object labels only. The second approach generates more detailed descriptions but may require more complex ODR models. The third approach has the potential to be more efficient and accurate, but it requires more complex training, which could be a limitation for some programmers developing real-world applications and may not be as interpretable as the other two approaches.

Considering these advancements, we reflect on how effective could a generic solution be today, that is able to transfer key visual information into audio communication. Importantly, such a general solution should be able to apply also in RW and for any language.

To describe this setting, we decided to use the broad phrase '*au*dible *rea*lity' (or AuRa) to denote a variety of options related to using sound perception as a means of experiencing or understanding reality. While this term is not a widely recognized (or commonly used) term in mainstream language or technology, we want to distinguish it from related topics, such as auditory analysis, virtual acoustics, binaural audio, sonification, and so on. Like virtual reality (VR) and augmented reality (AR) create a spatial sense of presence in a (digital) world, AuRa encompasses the use of sound to represent and interact with the physical world. However, in contrast to VR, AR or other mixed reality technologies (e.g., Real and Araujo, 2021), AuRa does not intend to be immersive or to represent digital environments. Moreover, we are interested in (AuRa) solutions that are portable and/or wearable (e.g., Kuriakose et al., 2023; Liu et al., 2018; Wang et al., 2021; Real and Araujo, 2019), without requiring multiple devices, interfaces or advanced neuroscience components (see Wang et al., 2021; Schinazi et al., 2016). Nonetheless, the AuRA solution we envision should be straightforward and able to be used together with other independent wearables referring to further sensory options (e.g., Kilian et al., 2022; Zhu et al., 2023).

During this work we also prototyped a proof-of-concept (PoC) and searched for key characteristics to evaluate AuRa performance in RW (see Figure 1). Our PoC was aimed to be deployable also on a smartphone with camera and be able to support users from diverse backgrounds in RT (including both visually impaired individuals and not). Based on this experience, we discuss relevant perspectives (potential, limitations, and challenges) and search for options available today. Importantly, we present a simple way to characterize similar solutions in a self-assessment reflection summary (see Supplementary Table 3). We anticipate that our work will add to the efforts of the community toward the development of more effective and accessible aids, particularly for individuals with visual difficulties.

2 Materials and methods

To build our PoC, we combined modern DL components in an integrated solution using ODR and TTS modules in a single pipeline (see Supplementary Data sections 2.1, 2.2, 2.3). Figures 1A,B provide an overview of the whole process.

3 Results

We wanted to implement functionalities that can be important in numerous occasions. In specific, we utilized modern DL techniques toward a PoC that could:

a) capture main objects present in live video stream frames,
b) announce them in voice, by
c) using the user's language of choice.

While this modular approach has been examined previously (e.g., in specific or in limited settings like in Kuriakose et al., 2023; Tapu et al., 2017; Guravaiah et al., 2023; Makhmudov et al., 2020; Alahmadi et al., 2023; Chen, 2022), or (Vijetha and Geetha, 2024), there are not many tools available today that combine these tasks together toward a complete solution that is suitable for the broad audience, for any language. There are several reasons for this dearth, including perhaps commercial prospects and restrictions, amount of effort required, access to resources, as well as maturity of underlying DL models and the rapid changes in the AI landscape.

We expect that as more developments take place, a multitude of programmable options will be examined more consistently spanning both application and user design options (e.g., language selection that the text will be translated into, rate at which the video is sampled or how frequently should the speaker summarize the view, techniques or rules for determining the spoken description, how many and which objects should be prioritized as important, text transformations, output features, or end of process), as well as technical capabilities (e.g., visual capture ability, number of objects and categories that can be detected, view resolution, refresh speed, distance, disk size, memory, etc.). We find that systematic examining of inherently underlying limitations and tradeoffs is important to determine an 'optimal, default' setting, especially when it comes to application in non-controlled conditions.

The main difficulty lies in the fact that features that may severely impact functionality and user experience include both quantitative and qualitative aspects that are not easy to measure. Some such examples include input capture (e.g., camera focus or distance of objects may sometimes limit usefulness of live, open world application), accuracy (e.g., errors in ODR, not natural-sounding or easy-to-understand TTS output), performance (e.g., coordination of object capture, detection, and vocal description speed can be a bottleneck leading to out-of-sync visual and audio), possible underlying model bias (e.g., training datasets may not be diverse enough to account for the desired RW scenarios, models may be biased toward recognizing certain types of objects only or toward providing only few speech patterns), hardware (e.g., RT requirement may limit availability and use in certain devices or environments), but also context and understandability aspects (e.g., not all objects may be detected in complex scenes or in scenes with occlusions, TTS models may not always be able to generate speech for a given context, and so on).

Fortunately, modern DL developments provide constant improving to each of those individual aspects and several commercial AI efforts exist, dedicated to relevant tasks (e.g., for video transcription or TTS generation, Video Transcription, 2024; VEED.IO, 2024; LOVO AI, 2024; AI Voice Generator, 2024a; AI Voice Generator, 2024b; Murf AI, 2024; Synthesys.io, 2024; ElevenLabs, 2024; Home, 2024; AI Studios, 2024; Fliki, 2024; Resemble AI, 2024; Descript, 2024; Maestra, 2024; Sonix, 2024; Media.io., 2024; Dubbing Tool, 2024, and so on). However, to put such capabilities in perspective together, in the context of an AuRA scenario is not straightforward. For this reason, we compiled a set of such characteristics that can be considered in combination when it comes to reflecting on the *AuRa*-bility of a tool under development (see Supplementary Table 3). Figure 1C provides a high-level summary of some of those aspects, either in taxonomy or circular format. Importantly, we anticipate that our summary will facilitate a more straightforward comparison and on par evaluation of AuRA tools. Moreover, it offers flexibility in its usage, allowing for the highlighting of different aspects each time. For instance, one might choose to summarize selected measures using a matrix or tabular form (with text- or color-coded cell descriptions), whereas for others a radar or spider representation may be more appropriate (see Figure 1C).

We find that to properly assess the use of current and future (e.g., multimodal) AI/ML based models in open world requires the development of appropriate benchmarks [(e.g., Liu et al., 2018)], probably tailored to specific use-scenarios. Another central source in assessing RW performance of any one integrated AuRa implementation

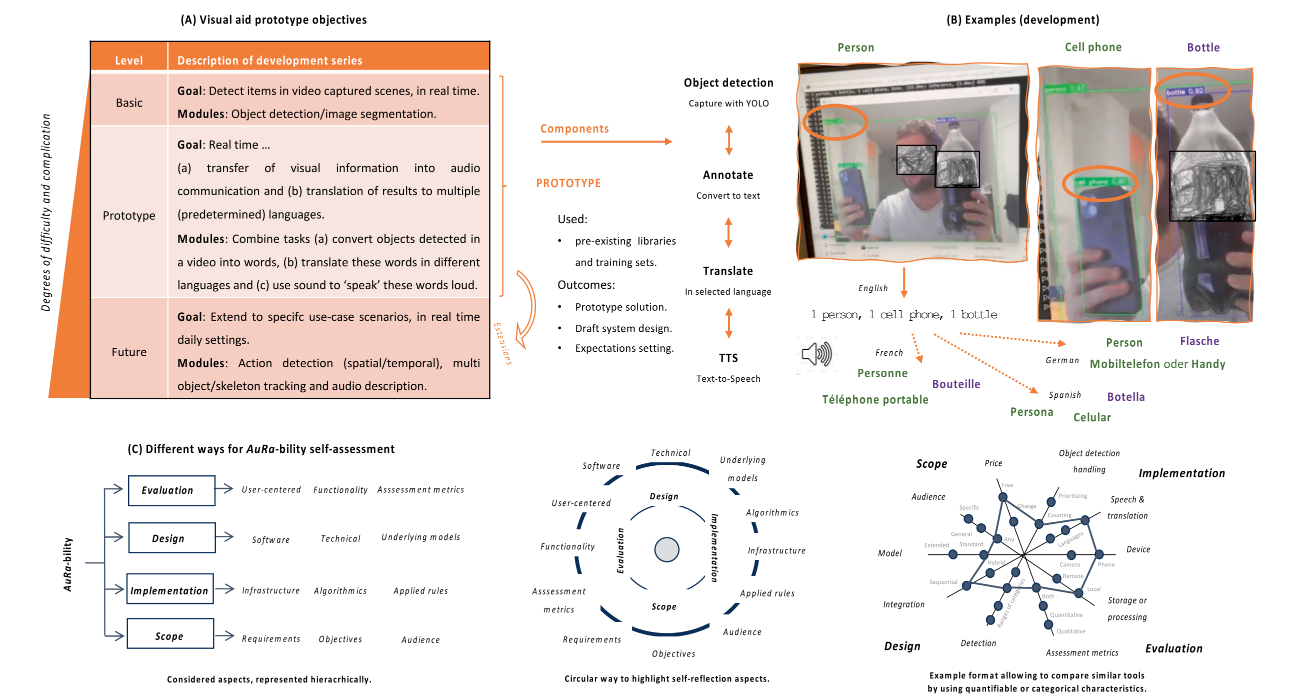

FIGURE 1
Summary of our approach and results. To build our visual aid system, we focused on being able to annotate image (video frames) in real-time (see, upper left part). **(A)** To do this we use object detection modules that can identify and locate objects within an image (or video), and output bounding boxes around those objects along with a text label that specifies the class of each object. We then automatically extract the names of these categories and pass them though a text-to-speech (TTS) module. Eventually, the generated summary can be 'spoken' in voice, in different languages (see, upper middle part). **(B)** Examples of tests during development (see, upper right part; three detected items translated in voice in selected languages)—developer's face and bottle's brand blurred. **(C)** Similar implemented technologies can be tested by users in real world settings and their *AuRa*-bility can be summarized in different self-assessment formats (see, lower part).

can be direct user evaluation, helping gather essential qualitative feedback. For example, a relevant questionnaire could help determine which customizable parameters are better according to users' preferences, whether users need training, what functionalities meet best the needs and of each target group, what is the scope of expectations (e.g., regarding input, language, understandability, or speed options) and, importantly, allow to vote what objects and actions/activities are important to be captured (see Supplementary Data section 2.4). We find that such surveys are necessary to determine the development of these systems further, providing timely feedback to improve and extend in terms of RW application.

4 Discussion

Inspired by the capabilities enabled by modern AI developments, we sought to explore the potential of transferring key visual information into audio communication toward the development of an audible visual aid. We approached this question from a 'higher-level' perspective aiming for fast, portable solutions that can apply in RW, in different languages and in RT. Without neglecting the extensive work done in the broader field, we avoid engaging in extensive historical review or performing in-depth comparison and analysis of previous approaches and current AI developments. Instead, we shortly review available options today and search for pragmatic approaches to combine recent ML/DL models with the objective of building 'light' standalone companion-applications that can support situations when visual limitations may hinder the ability to interpret the world.

One key setting where this would be important is for individuals with visual impairments who often face challenges in navigating their environment, in identifying objects, and in interacting with their surroundings. ODR technology has demonstrated potential in tackling some of these difficulties by automatically recognizing and localizing objects within digital images and videos (see Supplementary Table 1). However, interacting with such information alone may be limiting for some, as it may not be easily comprehensible. Speech-based interfaces, on the other hand, provide an additional format that may allow communicating information about objects in the environment (see Table 1), which might not be completely discernible through visual cues alone (e.g., Kuriakose et al., 2023). Combining, thus, these technologies can help improve quality of life both in terms of independence and safety.

Ultimately, automated content capture from live image/video streaming and converting this into spoken descriptions can have numerous important uses. Few such implications include, among others, the improved ability to:

- Navigate unfamiliar indoor/outdoor environments, by detecting obstacles and providing audio feedback on surroundings (e.g., by notifying the user of objects in near proximity in RT, while they are walking).
- Monitor for safety issues, by detecting and alerting the user about potential hazards (such as approaching vehicles, pedestrians, or other objects tracked) the system can help prevent avoidable accidents.
- Enhance independent life capabilities, by interacting with the surroundings more confidently and by performing everyday tasks (such as household activities) more effectively.

Overall, such systems are applicable in several business, industrial and other settings addressing important problems (see also Table 1; Supplementary Table 4), such as:

- Autonomous vehicles (e.g., automated detection of obstacles or of traffic conditions on the road to alert passengers or other vehicles in vicinity).
- Law, security, and surveillance (e.g., detection of people or objects in a specific area to provide alerts or instructions to security personnel).
- Smart homes (e.g., recognizing specific objects or people to provide personalized message greetings, alerts, or instructions to residents).
- Art, entertainment, and education (e.g., RT transcription and translation of visual content to make it more accessible and easier to understand for a wider range of people).

However, in this work we are interested in the support of individuals in cases where visual input may not be otherwise available. Such cases do not only include visual impairment, but also potential emergencies and scenarios for those who prefer speech to text-based interfaces (e.g., due to literacy or language limitations). Currently, there exist several options available for the general public that support similar cases. Supplementary Table 4 lists some such platforms for a variety of tasks (from nutrition to plants): characteristically, Google Lens (2024) has today 10B+ downloads, whereas Narrator's Voice (2024), Voice Aloud Reader (TTS) (2024) or T2S (2024) have 10 M+ downloads each. Typically, these systems tackle the two tasks (ODR and TTS) mostly separately. More importantly, we find that many of these apply in rather comfortable (controlled) situations, or on restricted settings, and are triggered mainly on demand. Even though this applies also for mobile apps that support visually impaired individuals, we notice that the latter are in comparison reasonably more adjusted for dynamic use in RW, providing RT feedback in non-specific environments (see Table 2).

All apps examined in Table 2 can recognize objects, people, and texts. Some offer more specific options such as describing color and lighting conditions, identifying currency, or locating objects (e.g., TapTapSee, 2023; Lookout, 2024), or (Seeing AI, 2024). However, we find that RW functionality may in some cases be semi-dependent in that user interaction is required to activate (or to determine respective) tasks after the initial launch of the application. This may be cumbersome in some situations, hindering full autonomy and may require a controlled setting or assistance from (or cooperation with) another person for optimal performance. Moreover, we find that more advanced features (like exploring surroundings, describing scenes or emotions, recognizing specific persons, understanding handwriting, determining approximate distance, or using audio AR to navigate in the world) are often in 'experimental' or 'beta' mode, requiring more improvement and research (e.g., Liu et al., 2018; TapTapSee, 2023; Lookout, 2024; Seeing AI, 2024).

Despite these limitations, it is expected that the constant technological progress observed today (e.g., updating pre-trained AI models, with more data or new architectures) will help make novel features available soon. Some forward-looking options may include also the repurposing of the multimodal capabilities of modern, integrated large language model (LLM) approaches [e.g., such as OpenAI's GPT-4 (GPT-4, 2024), Anthropic's Claude (Claude, 2024), or Google's DeepMind Gemini (Google DeepMind, 2024) models among others]. While direct human feedback remains the best in aiding the visually impaired, virtual AI enabled modules already complement such services that connect people needing sighted support with volunteers (e.g., Be my eyes, 2024) offers a virtual assistant integrating automated image-to-text technology and OpenAI GPT-4 features (OpenAI's GPT-4, 2024). Regarding design, we find that most Table 2 apps face similar challenges, irrespective of their AI modalities—examples include:

- World setting and input quality (e.g., each smartphone app is only able to recognize objects that are in focus and within the camera's scope; lighting conditions are also important for the quality of the identification).
- Visual capture (e.g., use of the phone's camera to take a picture or a video; some require the activity to take place upon demand, whereas others require slow speed while moving to allow for RT assessment, like Lookout, 2024); video stream might be limited in size—e.g., TapTapSee allows for videos that are up to 10 sec long to be captured each time (TapTapSee, 2023).
- Functionality grouping (e.g., whether the task is concerned with texts, objects, people, or other specific goal).
- Activation method (e.g., depending on the app, the task at hand—whether video capture or image analysis—could be triggered in different ways, like by tapping or via voice command; e.g., TapTapSee, 2023; Sullivan Plus, 2024).
- Cloud based services (i.e., some require online access to work; e.g., TapTapSee, 2023; Seeing AI, 2024).
- Multi-lingual support [i.e., number of languages that can be used may vary depending on task or on phone's OS—e.g., Seeing AI supports 20 languages (Seeing AI, 2024)], whereas Envision AI can read up to 60 languages (Envision, 2024); other apps use the language setting of phone's OS, like Lookout (2024).

Finally, even when users change language settings, this may not apply equally well for all languages or to all tasks [e.g., Lookout (2024) has a separate functionality for text reading and for food labels]. As result, some features may not be available in all languages or may perform better in some than in others. Nevertheless, multilanguage support is constantly evolving and we expect that this gap will close over time. One characteristic example of relevant rapid developments in recent years is the growth of LLMs, including their ability to translate between languages—some notable LLM models include Google's Bert

TABLE 2 Popular mobile apps with many downloads that combine object detection and TTS in real-time to support visually impaired[(+)].

Name	Downloads[D]	URL	Citation
TapTapSee[B]	500 K+	taptapseeapp.com	TapTapSee (2023)
Lookout[A]	500 K+	goo.gle/lookout	Lookout (2024)
Envision AI[B]	100 K+	letsenvision.com	Envision (2024)
Sullivan+[B]	100 K+	mysullivan.org	Sullivan Plus (2024)
Seeing AI[B]	50 K+	seeingai.com	Seeing AI (2024)

[(+)]More similar apps exist to support visually impaired with less than 100 K downloads such as Supersense (Supersense, 2024) or MyEyes (MyEyes, 2024), whereas some efforts may no longer be available for download and use, or may not be updated regularly [e.g., Vision (Vision, 2024) or Aipoly (V7 Aipoly, 2024)]; also, available options may be different depending on geographical regions. K, denotes thousands; TTS, short for text-to-speech; [D]information available from Google Play on April, 2024; [A]available only in Android; [B]available both for Android and iOS.

(see BERT, 2024; Devlin et al., 2019; Open Sourcing BERT, 2024; Google-research/bert, 2024), T5 (see T5, 2024; Raffel, 2023; Python. Google Research, 2024; Google-research/t5x, 2024), LaMDA (see Google, 2024; Thoppilan, 2022), PaLM (Google AI, 2024) and the more recent Bard (now Gemini, 2024), OpenAI's GPT series (v4, GPT-4, 2024), Anthropic's Claude (Claude, 2024), Microsoft's Copilot (Microsoft Copilot, 2024), Meta's LLaMa family (see Meta Llama, 2024; Meta-llama/llama3, 2024) and Mistral's AI models (Mistral AI, 2024). Uniting efforts away from a suite of AI language translation models toward a single speech model supporting multiple languages is endorsed by several larger corporations and open source contributions, like Meta (e.g., Meta AI, 2024a; Pratap, 2023) and the No Language Left Behind (NLLB) initiative (see Meta AI, 2024b; NLLB Team, 2022; NLLB, 2024; NLLB-MOE, 2024).

Ultimately, RT video object descriptions should provide valuable information in a variety of settings, not only to support individual's decision making in daily life but also to improve safety, quality of care, efficiency, and social inclusion (see Table 3). The development of such applications can benefit not only visually impaired persons but also other individuals who may have difficulty interpreting visual information, such as individuals with cognitive disabilities, with language or social barriers, or with limited literacy skills. Table 3 mentions few such cases when accessible AuRa experience can be important, with potential applications spanning various fields, including healthcare, education, and entertainment.

While the simplest AuRa form can thus be potentially of interest to a wide range of circumstances, in any setting in which a user's main input sensor is hearing (from remote device controlling to the safety monitoring of children or pets), it can be easily combined or enhanced with additional or with more advanced technologies (e.g., VR/AR extensions may also apply).

Nevertheless, despite significant progress, there are still challenges to address. One challenge is the need for robust ODR that can accurately perform in a wide range of environments and lighting conditions. In that aspect, we believe that modern generative AI techniques have the potential to be effective in addressing cases of fuzzy image capture in open world (e.g., by enhancing resolution or by generating simulated frames in incomplete video). Another challenge is the need for natural and expressive TTS systems that can convey object information in a clear and understandable manner. Beyond the stepwise approach and basic architecture of our PoC, there are today several opportunities to benefit from current multimodal breakthroughs [such as GPT-4, 2024, Claude, 2024, or Gemini (Google DeepMind, 2024)] that can handle both text and vision inputs, or the other way round (e.g., Imagen, 2024, Parti, 2024, DALL·E3, 2024, Sora, 2024), Make-A-Video (Meta AI, 2024c), Gen-2 (Runway, 2024), or Lumiere (2024). However, this progress must also be considered with caution, especially when it comes to specific tasks [e.g., see intricacies in LLM performance regarding scientific context, like Galactica Demo (2024) and the biochemical domain (Zhang, 2024)] or modalities, and it is unclear how much of AuRability such united AI efforts could address today already. We find that more user-centered design and evaluation studies are needed to ensure that the needs and preferences of users (whether visually impaired or not) are adequately met. Direct comparisons of such tools are also not straightforward and may have to consider several dimensions (e.g., Kuriakose et al., 2023; see also Figure 1C; Supplementary Table 3; Supplementary Section 2.4).

With our prototype experience discussion, we do not attempt to make a new proposal, but rather to reflect on the applicability of this given architecture and stepwise approach. We want to draw the

TABLE 3 Potential AuRA applications to personal and community well-being.

Area/Topic	Description (examples, scenarios)[H]
Social equality and integration	There are several groups of people who would benefit from real-time audio description, promoting participation and interaction. Key audience is people with visual disabilities who require more accessible and understandable content. Examples: • Social media: generated high-quality audio transcripts can assist in creators and content becoming more accessible to a wider audience. • Politics and businesses: marketing, advertising, and e-commerce content can become more accessible and appealing to a wider audience. • Family and friends: enabling of team-based activities with independent participation by overcoming sight limitations.
Arts, culture, entertainment	There are several groups of people who would benefit from real-time audio description, enhancing both experience and understanding. Examples: • Description of landscapes, statues, paintings, architectures. • Audible captions for live events, such as public speeches, presentations, and other visual broadcasts; applies to theater, museum, cinema, video gaming as well as outdoor activities.
Education and research	Consider researchers and educators who want to create searchable content for easier reference and analysis, or learners interested in understanding content for educational purposes. Examples: • Audible captions for live events, such as public speeches, presentations, and other visual broadcasts, like in classroom or scientific conferences. • Children learning the names of objects (in own or other language). • Opportunities for multilingual learning play and simulation-based training.
Health and safety	Remote counseling in isolated places or in limited visibility settings (e.g., due to distance, geographical or extreme weather conditions). Examples: • Clinical operation: healthcare professionals to identify and track medical equipment and supplies in remote site in real-time. • Telesurgery: audio confirmation of tools available on-site during operation progress. • Child safety: description of child's environment (identification of harmful objects) to avoid accidents, improving their independence, confidence, and mobility.
Emergency response	Remote support (e.g., via drones) to help efficient and effective search and rescue operations (e.g., fires, floods, earthquakes), ultimately saving lives and reducing the risk to rescue workers. Examples: • Provision of valuable information, improving response times, safety, and chances of survival. • Quick identification and localization of critical equipment, supplies, and victims or individuals who need assistance, in real-time, and in low-light or obscured environments. • Provision verbal instructions to individuals who may be disoriented or unable to communicate effectively (e.g., help a person in distress understand the situation by providing information while waiting for rescue). • Audio description of drone imagery transmitted to rescuers regarding the source or identification of hazardous materials in a critical area—gathering of information and assessment of the situation while in another area from safe distance, reducing exposure to danger.

[H]Highlighted scenarios are intended to refer to circumstances when audio (listening) is more admissible than visual (or other sensory) mediums. Examples of such circumstances may include multitasking, lack of personnel, or remote monitoring (e.g., consider a first responder with broken microphone transmitting live video to another field coordinator whose display-monitor has been broken).

attention toward the creation of better performing solutions and to establish an easy-to-use, adaptable, and transferable setting for extended use by the broader community. We believe that personalization and adaptations tailored to the specific circumstances of potential users will be warranted in future. For example, many of the freely available software (open) libraries and models today (e.g., see Supplementary Table 2), are pre-trained with 'relatively limited' image datasets—e.g., the 'COCO' dataset (COCO, 2023) is composed of only less than hundred different object classes (cars, persons, sport balls, bicycles, dogs, cats, horses, etc.)—that may be too generic for specific tasks. On the other hand, determining extended image datasets (e.g., ImageNet, 2023) should consider also the level of detail or abstractness that respective category labels will be described with. For example, using specific only dictionaries or ontologies, might interfere with effective multilingual support since some terminology or words may not be obvious to (unambiguously) translate in any language. Dictionary independent translation and TTS models—from any language to any language [e.g., NLLB (Meta AI, 2024b)]—are therefore important to be considered, especially when it comes to capturing titles of actions or of activities. From this perspective, action recognition poses challenges not only regarding technological implications (e.g., input, DL architecture and datasets), but also regarding the types (and number of predefined) action classes that should be (adequately) recognizable. Automated AI enabled 'audio description' projects require also further attention and standardization, especially when it comes to low vision users, to content that is not expressed via sound (e.g., a dialog) and to the diversity of possible context settings [e.g., Kuriakose et al., 2023; Brack, 2024; AD Lab Pro, 2024; Web Accessibility Initiative (WAI), 2024; Wang et al., 2021; Van Daele et al., 2024; Jain, 2023; Ning, 2024]. On some occasions, haptic, physical, or hardware support may contribute to improved detection and description performance (e.g., via marked labels or fixed QR codes placed at determined locations helping the AuRa system). Direct community support may also largely help extend AuRabilty scope (e.g., by engaging in image labeling activities, by prioritizing actions or objects deemed important to be handled first in different scenarios, and so on).

In future, we expect implementations with optimized sub-components and extended functionalities that revolve around (a) the decision-making support (e.g., for avoiding obstacles) and (b) the description of activities captured in longer, continuous 'single-take' video stream frames. These may also come as sophisticated versions of 'hybrid' approaches (e.g., integrated multi-modal systems)—indeed, present-day state-of-the-art provides a lot of opportunity to timely explore the extent that such capabilities can catalyze AuRa applications, contributing to several improvements that range from more efficient management of lengthy content (e.g., handling longer durations) to more advanced question answering and complex understanding skills (e.g., suitable for expert or domain-specific application). Ultimately, an advanced live AuRa description system will be an extended solution that can generate (sub-) title like audio text descriptions (or question responses) to provide an augmented experience.

To set priorities of future developments more appropriately, we invite the community and the public to engage in organized feedback projects (e.g., via questionnaires) providing regularly structured information guiding the specific goals and user requirements (e.g., about input parameters, languages, speed, understandability, scope, functionalities, types of objects, device components, etc.) that newer, targeted solutions should address. We also expect that such efforts can be more efficiently supported by the active involvement of the community in the preparation of datasets (e.g., training examples) from the collective contribution of crowd collaboration projects. Several platforms exist today that can enable exchange of image labeling information and annotation initiatives (e.g., Make Sense, 2024; V7, 2024; Dutta and Zisserman, 2019; CVAT, 2024; Open Source Data Labeling, 2024; Dataloop, 2024; Data Labeling, 2024; SuperAnnotate, 2024; Encord, 2024; LabelMe, 2024; Roboflow, 2024).

Altogether, we find that a straightforward architecture comprised of four main steps (i.e., video capture, object identification, description in text, translation in different languages) is a generic approach capable today already to help with the goal of building a working AuRa framework (see Figures 1A,B). With our work, emphasize, in addition, the importance of soliciting feedback directly from potential target groups to better guide tailored preferences and to inform future developments (see also Supplementary Table 3; Supplementary Section 2.4). To our opinion, the field can be dealt with more systematically, particularly given the technological capabilities demonstrated recently. In strong support of this direction is also the example of Project Astra (2024), a very freshly released initiative by Google's DeepMind toward a universal digital AI assistant. We believe that the community, even when with limited resources, should not miss the chance to more actively aid, together with larger organizations, in the structured evaluation and benchmarking of such advanced AI agents that are capable of more complex reasoning and multimodal interactivity. For these reasons, we anticipate that our discussion will be seminal, influencing some of the coming efforts of the community toward the development of more effective and accessible digital solutions for visually impaired persons, but also inspiring tools for important applications in medical, health or other emergency settings (see also Table 3). Importantly, our discussion underscores the role and impact of such digital interventions in protecting and improving broader public health and policies in terms of both personal and community well-being.

5 Conclusion

Combining object detection and speech conversion technologies has the potential to significantly enhance accessibility of information for visually impaired individuals. Beyond integrating separate distinct modules, we envisage more dynamic, open world applications, performing in RT, for any place and language. Many of today's portable mobile solutions are potentially able to contribute into breaking both visual impairment inequalities and restrictive language barriers. While there are still challenges to be addressed, the progress made in this area has been significant, and there is a strong foundation laid for continued development and optimization of these technologies. The community should also take advantage of the opportunity to explore the possibilities enabled by repurposing modern AI advancements (multimodal capabilities, improved interactivity, and more complicated reasoning) to tackle everyday situations. In addition, we expect that in near future more studies will take place to examine underlying trade-offs and that coming tools will enable functionalities that are tailored to more specific scenarios and audiences (e.g., imagine an AuRa agent answering to a visually impaired person reliably whether a street is safe to cross at a certain moment). We, therefore, invite the community to gather this information in an organized manner and to create appropriate performance benchmarks, which can inform decisions regarding model selection and system optimization strategies. We anticipate that broader, collaboratively sourced feedback can serve as an effective guide to the characteristics of future data focus and training efforts. Finally, we aspire that our comments and discussion will help raise more awareness on the challenges of visual impairment as well as will be influential to multiple such efforts.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

ÖA: Software, Resources, Writing – original draft. GP: Project administration, Resources, Writing – original draft. AG: Resources, Writing – review & editing. TS: Conceptualization, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2024.1424371/full#supplementary-material

References

AD Lab Pro. (2024). Audio description: a laboratory for the development of a new professional profile. Available at: https://www.adlabpro.eu/ (Accessed on June 02, 2024).

Ahmed, M. I., Spooner, B., Isherwood, J., Lane, M., Orrock, E., and Dennison, A. (2023). A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus* 15:e46454. doi: 10.7759/cureus.46454

AI Studios. (2024). "AI Avatar Video Generator." Available at: https://www.deepbrain.io/aistudios (Accessed on June 01, 2024)

AI Voice Generator. (2024a). Text to speech, #1 best AI voice. Available at: https://speechify.com/ (Accessed on April 08, 2024)

AI Voice Generator. (2024b). Realistic text to speech and AI voiceover. Available at: https://play.ht (Accessed on June 01, 2024)

Alahmadi, T. J., Rahman, A. U., Alkahtani, H. K., and Kholidy, H. (2023). Enhancing object detection for VIPs using YOLOv4_Resnet101 and text-to-speech conversion model. *Multimodal Technol. Interact.* 7:77. doi: 10.3390/mti7080077

Amazon Alexa Voice AI | Alexa Developer Official Site. (2024). Amazon Alexa. Available at: https://developer.amazon.com/en-US/alexa.html (Accessed on June 04, 2024)

Asrp, P. (2024). Python-espeak: Python C extension for the eSpeak speech synthesizer. Available at: https://github.com/asrp/python-espeak (Accessed on June 01, 2024)

Be my eyes. (2024). See the world together. Available at: https://www.bemyeyes.com/ (Accessed on April 08, 2024)

BERT. (2024). Pre-training of deep bidirectional transformers for language understanding. Available at: https://research.google/pubs/bert-pre-training-of-deep-bidirectional-transformers-for-language-understanding/ (Accessed on June 02, 2024)

Bhat, N. M. (2021). pyttsx3: Text to speech (TTS) library for Python 2 and 3. Works without internet connection or delay. Supports multiple TTS engines, including Sapi5, nsss, and espeak. Python. MacOS:: MacOS X, Microsoft:: Windows, POSIX. Available at: https://github.com/nateshmbhat/pyttsx3 (Accessed on June 01, 2024)

Bohr, A., and Memarzadeh, K. (2020). The rise of artificial intelligence in healthcare applications. *Artificial Intell. Healthcare* 2020, 25–60. doi: 10.1016/B978-0-12-818438-7.00002-2

Brack, F. (2024). "The audio description project," The Audio Description Project. Available at: https://adp.acb.org/index.html (Accessed on June 02, 2024).

Brock, S., Jackson, D. B., Soldatos, T. G., Hornischer, K., Schäfer, A., Diella, F., et al. (2023). Whole patient knowledge modeling of COVID-19 symptomatology reveals common molecular mechanisms. *Front. Mol. Med* 2. doi: 10.3389/fmmed.2022.1035290

Brock, S., Soldatos, T. G., Jackson, D. B., Diella, F., Hornischer, K., Schäfer, A., et al. (2022). The COVID-19 explorer—an integrated, whole patient knowledge model of COVID-19 disease. *Front. Mol. Med* 2. doi: 10.3389/fmmed.2022.1035215

Chen, J. (2022). "A Real-time 3D object detection, recognition and presentation system on a Mobile device for assistive navigation". Available at: https://academicworks.cuny.edu/cc_etds_theses/1069 (Accessed on June 04, 2024).

Chenais, G., Lagarde, E., and Gil-Jardiné, C. (2023). Artificial intelligence in emergency medicine: viewpoint of current applications and foreseeable opportunities and challenges. *J. Med. Internet Res.* 25:e40031. doi: 10.2196/40031

Claude. (2024). *Meet Claude*. Available at: https://www.anthropic.com/claude (Accessed on June 02, 2024)

COCO. (2023). Common objects in context. Available at: https://cocodataset.org/#home (Accessed on April 27, 2023)

Cortana. (2024). Your personal productivity assistant. Available at: https://www.microsoft.com/en-us/cortana (Accessed June 04, 2024).

CVAT. *Open Data Annotation Platform*. (2024). Available at: https://www.cvat.ai/ (Accessed on April 12, 2024).

DALL·E3. (2024). *DALL·E 3*. Available at: https://openai.com/index/dall-e-3/ (Accessed on June 02, 2024).

Data Labeling. (2024). Hive AI. Available at: https://thehive.ai/data-labeling (Accessed on April 12, 2024).

Dataloop. (2024). "Dataloop | let the builders build". Available at: https://dataloop.ai/ (Accessed on April 12, 2024).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. pp. 248–255.

Descript. (2024). All-in-one video and podcast editing, easy as a doc. Available at: https://www.descript.com/ (Accessed on June 01, 2024)

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. *arXiv* 1810:04805. doi: 10.48550/arXiv.1810.04805

Dubbing Tool. (2024). "Leading AI video localization and dubbing tool." Available at: https://www.rask.ai/ (Accessed on June 01, 2024)

Durette, P. N. (2022). gTTS: gTTS (Google text-to-speech), a Python library and CLI tool to interface with Google translate text-to-speech API: Python. Available at: https://about.readthedocs.com

Dutta, A., and Zisserman, A. (2019). "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM international conference on multimedia, Nice France: ACM*, pp. 2276–2279.

Elemento, O., Leslie, C., Lundin, J., and Tourassi, G. (2021). Artificial intelligence in cancer research, diagnosis and therapy. *Nat. Rev. Cancer* 21, 747–752. doi: 10.1038/s41568-021-00399-1

ElevenLabs. (2024). "Text to speech and AI voice generator". Available at: https://elevenlabs.io (Accessed on June 01, 2024)

Encord. (2024). "The complete data development platform for AI". Available at: https://encord.com/ (Accessed on April 12, 2024).

Envision. (2024). Perceive possibility. Available at: https://www.letsenvision.com/ (Accessed on April 08, 2024)

Fliki. (2024). "Fliki: AI video generator - turn ideas into videos," Available at: https://fliki.ai (Accessed on June 01, 2024)

Galactica Demo. (2024). Available at: https://galactica.org/?via=topaitools (Accessed on June 02, 2024).

Gamage, B., Do, T.-T., Price, N. S. C., Lowery, A., and Marriott, K. (2023). "What do blind and low-vision people really want from assistive smart devices? Comparison of the literature with a focus study," in *Proceedings of the 25th international ACM SIGACCESS conference on computers and accessibility*, in ASSETS '23. New York, NY, USA: Association for Computing Machinery, pp. 1–21.

Gemini. (2024). "Gemini - chat to supercharge your ideas." Available at: https://gemini.google.com (Accessed on June 02, 2024).

Goecks, J., Jalili, V., Heiser, L. M., and Gray, J. W. (2020). How machine learning will transform biomedicine. *Cell* 181, 92–101. doi: 10.1016/j.cell.2020.03.022

Google. (2024). "LaMDA: our breakthrough conversation technology". Available at: https://blog.google/technology/ai/lamda/ (Accessed on June 02, 2024).

Google AI. (2024). "Google AI PaLM 2". Available at: https://ai.google/discover/palm2/ (Accessed on June 02, 2024).

Google Assistant. (2024). Your own personal Google Assistant. Available at: https://assistant.google.com/ (Accessed on June 04, 2024).

Google DeepMind. (2024). "Gemini". Available at: https://deepmind.google/technologies/gemini/ (Accessed on June 2, 2024)

Google Lens. (2024). Search what you see. Available at: https://lens.google/ (Accessed on April 01, 2024)

Google-research/bert. (2024). Python. Google Research. Available at: https://github.com/google-research/bert (Accessed on June 02, 2024).

Google-research/t5x. (2024). Python. Google research. Available at: https://github.com/google-research/t5x (Accessed on June 02, 2024).

GPT-4. (2024). *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses*. Available at: https://openai.com/index/gpt-4/ (Accessed on June 2, 2024)

Grant, K., McParland, A., Mehta, S., and Ackery, A. D. (2020). Artificial intelligence in emergency medicine: surmountable barriers with revolutionary potential. *Ann. Emerg. Med.* 75, 721–726. doi: 10.1016/j.annemergmed.2019.12.024

Guravaiah, K., Bhavadeesh, Y. S., Shwejan, P., Vardhan, A. H., and Lavanya, S. (2023). Third eye: object recognition and speech generation for visually impaired. *Procedia Comput. Sci.* 218, 1144–1155. doi: 10.1016/j.procs.2023.01.093

Hao, Y., Yang, F., Huang, H., Yuan, S., Rangan, S., Rizzo, J. R., et al. (2024). A multi-modal foundation model to assist people with blindness and low vision in environmental interaction. *J. Imag.* 10:103. doi: 10.3390/jimaging10050103

Hemavathy, J., Shree, A. Sabarika, Priyanka, S., and Subhashree, K., (2023). "AI based voice assisted object recognition for visually impaired society," in *2023 International conference on data science, agents and artificial intelligence (ICDSAAI)*, pp. 1–7.

OpenCV. (2024). Home. Available at: https://opencv.org/ (Accessed on June 01, 2024).

Home. (2024). WellSaid Labs. Available at: https://wellsaidlabs.com/ (Accessed on June 01, 2024)

Imagen. (2024). Text-to-image diffusion models. Available at: https://imagen.research.google/ (Accessed on June 02, 2024).

ImageNet. (2023). Available at: https://www.image-net.org/index.php (Accessed on April 27, 2023)

Jain, G. (2023). "Front row: automatically generating immersive audio representations of tennis broadcasts for blind viewers," in *Proceedings of the 36th annual ACM symposium on user Interface software and technology, in UIST '23. New York, NY, USA: Association for Computing Machinery*, pp. 1–17.

Kilian, J., Neugebauer, A., Scherffig, L., and Wahl, S. (2022). The unfolding space glove: A wearable Spatio-visual to haptic sensory substitution device for blind people. *Sensors (Basel)* 22:1859. doi: 10.3390/s22051859

Kuriakose, B., Shrestha, R., and Sandnes, F. E. (2023). DeepNAVI: A deep learning based smartphone navigation assistant for people with visual impairments. *Expert Syst. Appl.* 212:118720. doi: 10.1016/j.eswa.2022.118720

LabelMe. (2024). The open annotation tool. Available at: http://labelme.csail.mit.edu/guidelines.html (Accessed on April 12, 2024).

Lee, D., and Yoon, S. N. (2021). Application of artificial intelligence-based Technologies in the Healthcare Industry: opportunities and challenges. *Int. J. Environ. Res. Public Health* 18:E271. doi: 10.3390/ijerph18010271

Lin, T.-Y., Maire, M., Belongie, S., and Hays, J. (2015). Microsoft COCO: common objects in Context. *arXiv* 1405:0312. doi: 10.48550/arXiv.1405.0312

Liu, Y., Stiles, N. R., and Meister, M. (2018). Augmented reality powers a cognitive assistant for the blind. *eLife* 7:e37841. doi: 10.7554/eLife.37841

Liu, M., Wu, J., Wang, N., Zhang, X., Bai, Y., Guo, J., et al. (2023). The value of artificial intelligence in the diagnosis of lung cancer: A systematic review and meta-analysis. *PLoS One* 18:e0273445. doi: 10.1371/journal.pone.0273445

Lookout. (2024). Assisted vision - apps on Google play. Available at: https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en_US (Accessed on April 8, 2024)

LOVO AI. (2024). AI voice generator: realistic text to Speech and Voice Cloning. Available at: https://lovo.ai/ (Accessed on June 01, 2024)

Lumiere. (2024). Google research. Available at: https://lumiere-video.github.io/ (Accessed on June 02, 2024).

Maestra. (2024). "Maestra AI - transcription, subtitling and voiceover." Available at: https://maestra.ai (Accessed on June 01, 2024)

Maimon, A., Wald, I. Y., Ben Oz, M., Codron, S., Netzer, O., Heimler, B., et al. (2023). The topo-speech sensory substitution system as a method of conveying spatial information to the blind and vision impaired. *Front. Hum. Neurosci.* 16. doi: 10.3389/fnhum.2022.1058093

Make Sense. (2024). Available at: https://www.makesense.ai/ (Accessed on April 11, 2024).

Makhmudov, F., Mukhiddinov, M., Abdusalomov, A., Avazov, K., Khamdamov, U., and Cho, Y. I. (2020). Improvement of the end-to-end scene text recognition method for 'text-to-speech' conversion. *Int. J. Wavelets Multiresolution Inf. Process.* 18:2050052. doi: 10.1142/S0219691320500526

Malligere Shivanna, V., and Guo, J.-I. (2023). Object detection, recognition, and tracking algorithms for ADASs—A study on recent trends. *Sensors (Basel)* 24:249. doi: 10.3390/s24010249

Masmoudi, M., Ghazzai, H., Frikha, M., and Massoud, Y. (2019). "Object detection learning techniques for autonomous vehicle applications," in *2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pp. 1–5.

Media.io. (2024). "Media.io - Online Video, Audio, Image AI Tools," Available at: https://www.media.io/ (Accessed on June 01, 2024)

Meta AI. (2024a). "Introducing speech-to-text, text-to-speech, and more for 1,100+ languages," Available at: https://ai.meta.com/blog/multilingual-model-speech-recognition/ (Accessed on June 02, 2024).

Meta AI. (2024b). Research topic - no language left behind. Available at: https://ai.meta.com/research/no-language-left-behind/ (Accessed on June 02, 2024).

Meta AI. (2024c). Make-A-video. Available at: https://makeavideo.studio/ (Accessed on June 02, 2024).

Meta Llama. (2024). Meta Llama. Available at: https://llama.meta.com/ (Accessed on June 02, 2024).

Meta-llama/llama3. (2024). Python. Meta Llama. Available at: https://github.com/meta-llama/llama3 (Accessed on June 02, 2024).

Microsoft Copilot. (2024). "Microsoft copilot: your everyday AI companion." Available at: https://ceto.westus2.binguxlivesite.net/ (Accessed on June 02, 2024).

Mistral AI. (2024). "Mistral AI technology." Available at: https://mistral.ai/technology/ (Accessed on June 02, 2024).

Mullowney, M. W., Duncan, K. R., Elsayed, S. S., Garg, N., van der Hooft, J. J. J., Martin, N. I., et al. (2023). Artificial intelligence for natural product drug discovery. *Nat. Rev. Drug Discov.* 22, 895–916. doi: 10.1038/s41573-023-00774-7

Murf AI. (2024). AI voice generator: versatile text to speech software. Available at: https://murf.ai (Accessed on June 01, 2024)

MyEyes. (2024). AI assistant for blind and low-vision people. Available at: https://myeyes.app/ (Accessed on April 10, 2024)

Narrator's Voice. (2024). TTS - apps on Google play. Available at: https://play.google.com/store/apps/details?id=br.com.escolhatecnologia.vozdonarrador&hl=en (Accessed on April 8, 2024)

Neugebauer, A., Rifai, K., Getzlaff, M., and Wahl, S. (2020). Navigation aid for blind persons by visual-to-auditory sensory substitution: A pilot study. *PLoS One* 15:e0237344. doi: 10.1371/journal.pone.0237344

Ning, Z. (2024). "SPICA: interactive video content exploration through augmented audio descriptions for blind or low-vision viewers," in *Proceedings of the CHI conference on human factors in computing systems, in CHI '24. New York, NY, USA: Association for Computing Machinery*, pp. 1–18.

NLLB. (2024). Available at: https://huggingface.co/docs/transformers/en/model_doc/nllb (Accessed on June 02, 2024).

NLLB Team (2022). No language left behind: scaling human-centered machine translation. *arXiv* 2207:04672. doi: 10.48550/arXiv.2207.04672

NLLB-MOE. (2024). Available at: https://huggingface.co/docs/transformers/en/model_doc/nllb-moe (Accessed on June 02, 2024).

Olawade, D. B., Wada, O. J., David-Olawade, A. C., Kunonga, E., Abaire, O., and Ling, J. (2023). Using artificial intelligence to improve public health: a narrative review. *Front. Public Health* 11:1196397. doi: 10.3389/fpubh.2023.1196397

Open Source Data Labeling. (2024). Label Studio. Available at: https://labelstud.io/ (Accessed on April 12, 2024).

Open Sourcing BERT. (2024). State-of-the-art pre-training for natural language Processin. Available at: http://research.google/blog/open-sourcing-bert-state-of-the-art-pre-training-for-natural-language-processing/ (Accessed on June 02, 2024).

OpenAI's GPT-4. (2024). "Introducing be my AI (formerly virtual volunteer) for people who are blind or have low vision, powered." Available at: https://www.bemyeyes.com/blog/introducing-be-my-eyes-virtual-volunteer (Accessed on April 08, 2024)

Orynbay, L., Razakhova, B., Peer, P., Meden, B., and Emeršič, Ž. (2024). Recent advances in synthesis and interaction of speech, text, and vision. *Electronics* 13:1726. doi: 10.3390/electronics13091726

Park, W. J., and Fine, I. (2023). The perception of auditory motion in sighted and early blind individuals. *Proc. Natl. Acad. Sci.* 120:e2310156120. doi: 10.1073/pnas.2310156120

Parti. (2024). Pathways autoregressive text-to-image model. Available at: https://sites.research.google/parti/ (Accessed on June 02, 2024).

Pooja, S., Urs, A. S., Raj, V. B., Madhu, B. R., and Kumar, V. (2024). "Cognitive object detection: A deep learning approach with auditory feedback," in *2024 IEEE International Conference for Women in Innovation, Technology and Entrepreneurship (ICWITE)*, pp. 162–167.

Pratap, V. (2023). Scaling speech technology to 1,000+ languages. *arXiv* 2305:13516. doi: 10.48550/arXiv.2305.13516

Project Astra. (2024). Google DeepMind. Available at: https://deepmind.google/technologies/gemini/project-astra/ (Accessed on June 03, 2024).

Python. Google Research. (2024). Google-research/text-to-text-transfer-transformer. Available at: https://github.com/google-research/text-to-text-transfer-transformer (Accessed on June 02, 2024).

Python.org. (2023). "Welcome to Python.org." Available at: https://www.python.org/ (Accessed on April 27, 2023)

Raffel, C. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv* 1910:10683. doi: 10.48550/arXiv.1910.10683

Real, S., and Araujo, A. (2019). Navigation Systems for the Blind and Visually Impaired: past work, challenges, and open problems. *Sensors (Basel)* 19:3404. doi: 10.3390/s19153404

Real, S., and Araujo, A. (2021). VES: A mixed-reality development platform of navigation Systems for Blind and Visually Impaired. *Sensors (Basel)* 21:6275. doi: 10.3390/s21186275

Resemble AI. (2024). "AI voice generator with text to speech converter." Available at: https://www.resemble.ai/text-to-speech-converter/ (Accessed on June 01, 2024)

Roboflow. (2024). Give your software the power to see objects in images and video. Available at: https://roboflow.com/ (Accessed on April 11, 2024).

Runway. (2024). "Gen-2 by runway," Available at: https://research.runwayml.com/gen2 (Accessed on June 02, 2024).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* 115, 211–252. doi: 10.1007/s11263-015-0816-y

Schinazi, V. R., Thrash, T., and Chebat, D.-R. (2016). Spatial navigation by congenitally blind individuals. *Wiley Interdiscip. Rev. Cogn. Sci.* 7, 37–58. doi: 10.1002/wcs.1375

Schneider, P., Walters, W. P., Plowright, A. T., Sieroka, N., Listgarten, J., Goodnow, R. A.Jr., et al. (2020). Rethinking drug design in the artificial intelligence era. *Nat. Rev. Drug Discov.* 19, 353–364. doi: 10.1038/s41573-019-0050-3

Seeing AI. (2024). Talking camera for the blind. Available at: https://www.seeingai.com/ (Accessed on April 8, 2024)

Shvadron, S., Snir, A., Maimon, A., Yizhar, O., Harel, S., Poradosu, K., et al. (2023). Shape detection beyond the visual field using a visual-to-auditory sensory augmentation device. *Front. Hum. Neurosci.* 17:1058617. doi: 10.3389/fnhum.2023.1058617

Siri. (2024). Apple. Available at: https://www.apple.com/siri/ (Accessed on June 04, 2024)

Soldatos, T. G., Kaduthanam, S., and Jackson, D. B. (2019). Precision oncology-the quest for evidence. *J Pers Med* 9:E43. doi: 10.3390/jpm9030043

Soldatos, T. G., Kim, S., Schmidt, S., Lesko, L. J., and Jackson, D. B. (2022). Advancing drug safety science by integrating molecular knowledge with post-marketing adverse event reports. *CPT Pharmacometrics Syst. Pharmacol.* 11, 540–555. doi: 10.1002/psp4.12765

Sonix. (2024). "Automatically convert audio and video to text: fast, accurate, and affordable". Available at: https://sonix.ai/ (Accessed on June 01, 2024)

Sora. (2024). Creating video from text. Available at: https://openai.com/index/sora/ (Accessed on June 02, 2024).

Sullivan Plus. (2024). *Discover the world with Sullivan Plus. Let it become your eyes to seeing the world.* Available at: https://mysullivan.org/ (Accessed on April 8, 2024)

SuperAnnotate. Empowering enterprises with custom LLM/GenAI/CV models. (2024). Available at: https://www.superannotate.com/ (Accessed on April 12, 2024).

Supersense. (2024). AI for blind / scan text, money and objects. Available at: https://www.supersense.app/ (Accessed on April 10, 2024)

Synthesys.io. (2024). "Unlock generative AI content at scale." Available at: https://synthesys.io/ (Accessed on June 01, 2024)

T2S. (2024). T2S. Available at: https://app-t2s.web.app/ (Accessed on April 8, 2024)

T5. (2024). "Exploring transfer learning with T5: the text-to-text transfer transformer." Available at: http://research.google/blog/exploring-transfer-learning-with-t5-the-text-to-text-transfer-transformer/ (Accessed on June 02, 2024).

TapTapSee. Blind and visually impaired assistive technology - powered by CloudSight. Ai image recognition API. (2023). Available at: https://taptapseeapp.com/ (Accessed on May 2, 2023)

Tapu, R., Mocanu, B., and Zaharia, T. (2017). DEEP-SEE: joint object detection, tracking and recognition with application to visually impaired navigational assistance. *Sensors (basel)* 17:2473. doi: 10.3390/s17112473

TensorFlow. (2023). TensorFlow. Available at: https://www.tensorflow.org/ (Accessed on April 27, 2023)

Thoppilan, R. (2022). LaMDA: language models for dialog applications. *arXiv* 2201:08239. doi: 10.48550/arXiv.2201.08239

V7. (2024). The AI data engine for Computer Vision and Generative AI. Available at: https://www.v7labs.com/ (Accessed on April 11, 2024).

V7 Aipoly. *Vision AI for the Blind and Visually Impaired.* (2024). Available at: https://www.aipoly.com/ (Accessed on April 10, 2024)

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* 18, 463–477. doi: 10.1038/s41573-019-0024-5

Van Daele, T., Iyer, A., Zhang, Y., Derry, J. C., Huh, M., and Pavel, A. (2024). "Making short-form videos accessible with hierarchical video summaries," in *Proceedings of the CHI conference on human factors in computing systems, in CHI '24. New York, NY, USA: Association for Computing Machinery*, pp. 1–17.

VEED.IO. (2024). AI video editor - fast, online, free," VEED.IO. Available at: https://www.veed.io (Accessed on June 01, 2024)

Video Transcription. (2024). "Video to text automation." Available at: https://letterdrop.com/video-to-text (Accessed on June 01, 2024)

Vijetha, U., and Geetha, V. (2024). Obs-tackle: an obstacle detection system to assist navigation of visually impaired using smartphones. *Mach. Vis. Appl.* 35:20. doi: 10.1007/s00138-023-01499-8

Vision. (2024). For blind people - apps on Google play. Available at: https://play.google.com/store/apps/details?id=com.talovstudio.vision&hl=en (Accessed on April 10, 2024)

Voice Aloud Reader (TTS). (2024). *Apps on Google play.* Available at: https://play.google.com/store/apps/details?id=com.hyperionics.avar&hl=en (Accessed on April 8, 2024)

Wang, Z., Li, H., Chen, J., Chai, X., and Zhai, Z. (2021). "A wearable vision-to-audio sensory substitution system based on deep learning for the visually impaired," in *2021 International Conference on Digital Society and Intelligent Systems (DSInS)*, pp. 283–286.

Wang, Y., Liang, W., Huang, H., Zhang, Y., Li, D., and Yu, L.-F. (2021). "Toward automatic audio description generation for accessible videos," in *Proceedings of the 2021 CHI conference on human factors in computing systems, in CHI '21. New York, NY, USA: Association for Computing Machinery*, pp. 1–12.

Wang, J., Xue, L., Jiang, J., Liu, F., Wu, P., Lu, J., et al. (2024). Diagnostic performance of artificial intelligence-assisted PET imaging for Parkinson's disease: a systematic review and meta-analysis. *npj Digit. Med* 7, 1–11. doi: 10.1038/s41746-024-01012-z

Web Accessibility Initiative (WAI). (2024). *Description of visual information*. Available at: https://www.w3.org/WAI/media/av/description/ (Accessed on June 02, 2024).

Zhang, Q. (2024). Scientific large language models: A survey on Biological and Chemical Domains. *arXiv* 2401:14656. doi: 10.48550/arXiv.2401.14656

Zhu, H. Y., Hossain, S. N., Jin, C., Singh, A. K., Nguyen, M. T. D., Deverell, L., et al. (2023). An investigation into the effectiveness of using acoustic touch to assist people who are blind. *PLoS One* 18:e0290431. doi: 10.1371/journal.pone.0290431