Theia: Assistive Environmental Communication via Audio Patterns for the Visually Impaired

Alexander Mehta, Ritik Jalisatgi

Introduction

Blind users have issues with precisely understanding their environment. Many current devices that assist the blind rely on verbal descriptions such as ***"orange on your left side"*** or ***"chair on your upper right side"***. The issue here is that verbal descriptions are only effective when the environment is simple. Verbal descriptions have trouble communicating precise position, shape, and size in a simple format. However, we created a system that solves this issue!

Theia uses **sound effects** and **verbal cues** to precisely communicate a blind user's environment. We use instrumental sounds in combination with verbal descriptions for better communication with the blind. Instrumental sounds such as piano notes or drum noises, are used to communicate precise **position and depth**, while verbal cues communicate **object type**.

A blind user can press a button and understand precise details about their environment in ~5 seconds through **Theia**.

Literature Review

Blind or visually impaired humans can possess traits of echolocation according to a (N=37) preliminary study by Thaler (2013) and is backed up by Thaler, Arnott, and Goodale (2011) who measured brain data to confirm that blind individuals that echolocate effectively often use the same places of their brain stimulated by visual activity.

Object Detection and Segmentation based assistance by Jiang, Lin, and Qu (2016) addresses this through a real time system to provide users with descriptions of objects that are played at where the objects are located. Jiang et al. (2016) connects to a large server for computation, and relies on a vertical component of headphones that doesn't exist without use of audio illusions (or uses height virtualization which is expensive and ineffective).

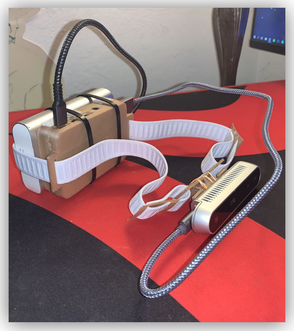
Text-based approaches have been introduced by Sarwar et al. (2022) by using OCR technology in order to describe signs to the blind. This technology can run on a Raspberry Pi, but fails to account for the overall scene, and non-text markers.

Engineering Goal

We set out to address previous design weaknesses by researchers. Specifically, a solution should be **low cost** (<$300) to ensure low barrier of entry, be **computationally efficient** to ensure consistent operation, address a **wide variety of objects**, have **precise spatial communication**, and have **quick communication**.

Physical Hardware Details

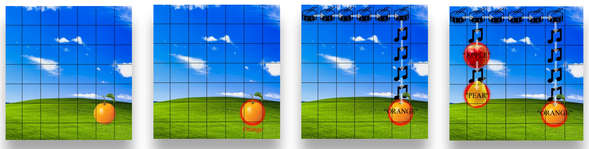
We use the **Intel Realsense D415** and a **Raspberry Pi 4** as a system under **$300** that has powerful enough to run our communication method. The Intel Realsense D415 gives us the depth data of a scene at a lower price point than LIDAR (**$100 vs $1,500**). All the hardware is connected to a power bank for *on-the-go* usage, and computations are ran on our Raspberry Pi 4.



Communication Method

Location Description through Instrumental Sounds

Method Visualized *(Sounds are played in the span of ~5 seconds)*



To communicate the user's environment, we used **piano sounds, verbal audio cues, and drum sounds**. We choose piano sounds and drum sounds because of to the user's familiarity with the sounds, which would allow them to understand the **pitches / frequencies** of the notes more easily.

Lower pitch sounds represent further points, and high pitch sounds represent close points. Along with this, they represent the vertical position of objects. Higher up objects have high pitch sounds and lower objects have low pitch sounds.

For communicating the horizontal positioning of objects, we would shift the audio between the user's left and right headphones. If the sound was *all the way on the right*, the object would be *all the way on the right.*

Finally, to communicate the type of object, we would just play a verbal audio for what the object was. If the object was an orange, we would play the sound of a man saying **"orange"**. *The verbal audios may be swapped for another language, or a sound effect pertaining to the object.*

Object Detection

We use the **YOLOv5** (You Only Look Once) model in order to detect objects. Specifically, we use a **downsized model** provided by Ultraytics with smaller model size and faster speed. The choice of YOLO over other models such as Swin Vision Transformer, R-CNN, and ConvNeXt comes down to **efficiency**. ConvNeXt models rely on transformers, which outperform YOLO at the cost of efficiency. A comparison table is below. Accuracy was tested on the COCO 2017 dataset.

Model	Accuracy	Parameters
YOLO Small	45.40%	7.2M
STT	48.50%	69M
ConvNeXt Tiny	49.60%	82M

The YOLO algorithm works through the following steps:

1. Initialize the image into a grid space.
2. For each cell grid:
   a. Detect probability of an object and the significance of the grid
3. Find intersection over union areas (IOUs) and remove the object from the grid in the less significant area, but expand the bounding boxes.

Using the cell method allows YOLO to compute all locations in one forward pass, while other methods such as R-CNN commonly detect objects one at a time. Splitting the image into a grid gets rid of the step of deciding where the algorithm should look for objects.

Human Testing Methods and Procedure

We use the following testing methodology:

1. Blindfold subject, and face them towards the table
2. Repeat the following process 3 times and record results
   a. Randomly select 1-3 objects
   b. Place items randomly on testing area
   c. Participants will guess location of object and object type (record responses and compare to actual position)
   d. Show them error in order to induce **pavlovian conditioning –** the concept that over time humans will learn patterns and associations of senses to results
   e. Repeat until have results for 1,2, and 3 objects.

Our procedure generalizes to situations such as locating a car in your environment, or understanding the location of people on a sidewalk.

Testing-trial images *(Participant consent given and form signed)*

Results

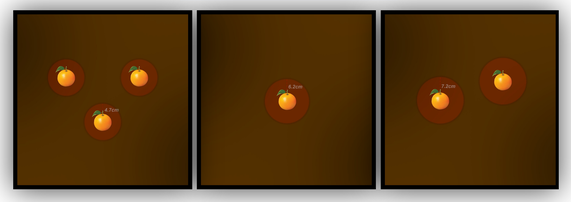
Results of the human experiment were recorded and statistical analysis was done to determine error. We found that in our testing of **9 participants** *(85 data points)*, Yolo correctly identified the object **94%** of the time. We normalized the distance by the size of the table (*61cm by 61cm*) to get a percent value which allows the data to be understood regardless of table size *(known as **NEA** or **Normalized Euclidean Accuracy**)*:

$$NEA = 1 - \frac{\text{error}}{\text{maximum error}}$$

Error Results

Object Count	Mean Error (cm)	σ (cm)	Mean NEA (%)
1	6.21	5.80	92.81%
2	7.20	5.90	91.67%
3	4.69	5.18	94.57%
Average	6.03	5.63	93.02%

Error visualized on a table:



Conclusion/Future Work

Results show that Theia helps visually impaired understand their environment accurately. In testing trials, Theia allowed users to understand objects near them with a **93.02% Mean NEA** and a **5.63cm SD** on a **61cm x 61cm table**.

Sound systems that are able to express sound with with more spatial precision would benefit Theia's communication method. Additionally, machine learning speedup methods such as **quantization** and **mixed precision** were not used (due to stability concerns), but could lead to speedups with ample testing. The machine learning method could also be improved by **utilizing the built in point-cloud data** from the real-sense sensors for better depth understanding. This was not done due to performance concerns, but as SoC's become more powerful and cheaper, it may be an option worth exploring.

References

Thaler, L., Arnott, S. R., & Goodale, M. A. (2011). Neural correlates of natural human echolocation in early and late blind echolocation experts. PLoS one, 6(5), e20162.

Massiceti D, Hicks SL, van Rheede JJ (2018) Stereosonic vision: Exploring visual-to-auditory sensory substitution mappings in an immersive virtual reality navigation paradigm. PLoS ONE 13(7): e0199389. https://doi.org/10.1371/journal.pone.0199389

Jiang, Rui & Lin, Qian & Qu, Shuhui. (2016). Let Blind People See: Real-Time Visual Recognition with Results Converted to 3D Audio.