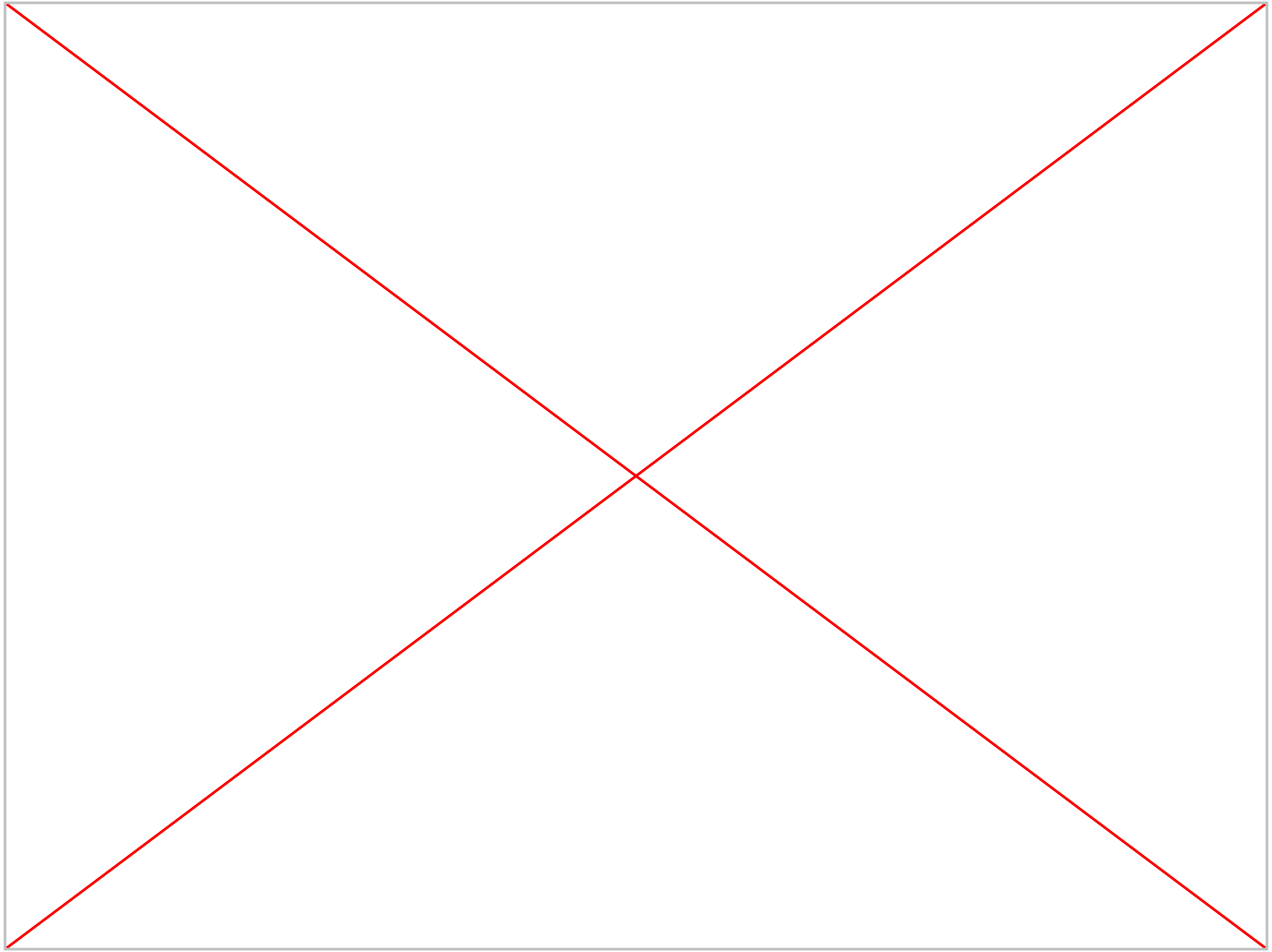
Visually Impaired Accessible Room

Alex Hafemeister, Arshia Dabiran, Dhruv Sirohi, Ethan Jiang



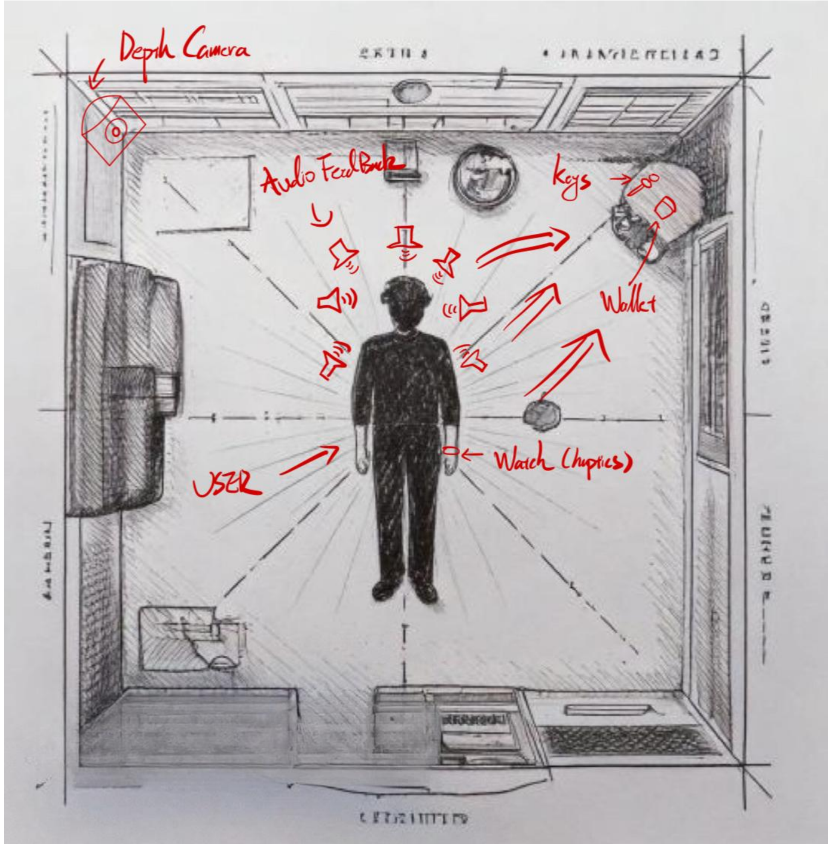
Motivation & Objectives

- Visually impaired people face highly uncertain environments when traveling.
- A smart room which can be installed quickly and easily in existing public establishments.
- Seamless guidance for the user in a new environment using haptic and audio feedback.

Relatively simple layout, but many moving parts.

Short term stay means user does not have time to acclimate.

System Overview



Technical Approach and Novelty

Current Methods and Limitations:

Head-Mounted Depth Cameras:

- Used for mapping and object detection
- Limitations: Uncomfortable for prolonged use, lacks intuitive feedback

Haptic Feedback in Object Retrieval:

- Guides precise object grasping
- Limitations: Minimal integration with real-time spatial navigation

Speech Recognition for Navigation:

- Enables hands-free interaction
- Limitations: Prone to errors in noisy environments

Moving Object Detection:

- Uses RGB-D data to track moving objects

Room Scanning and 3D Object Detection:

- Integrates RGB and depth data for precise 3D mapping
- Limitations: High computational requirements

Limitations of Current Practice:

- Single-modality feedback (visual, audio, or haptic alone)
- Head-mounted devices are often inconvenient
- Speech recognition affected by background noise

Our Approach and Novelty:

Stationary Depth Camera (RealSense L515):

- Mounted at an optimal height for stable, real-time room mapping

Multi-Modal Feedback Integration:

- Synchronizes haptic (Apple Watch) and auditory cues for intuitive guidance

Hands-Free Speech Commands:

- Uses speech to text for robust speech recognition and context-aware instructions and keyword detection

What's New:

Combines stationary depth sensing, multimodal feedback, and natural language processing for an intuitive experience that enhances user independence in object retrieval and navigation tasks.

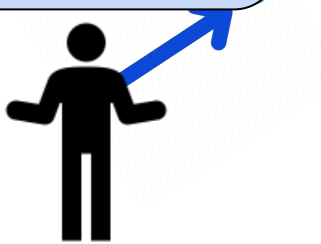
Overview

User request:
e.g. I need coffee,
find my laptop

User pose
detection, angles,
vectors
(L515 Depth
Camera)

Haptic proximity
guidance via
Apple Watch

Audio directional
guidance via
AirPods



Speech to text to capture
object of interest

Server uses camera data
to return vector to the app
(direction, hand-object
distance)

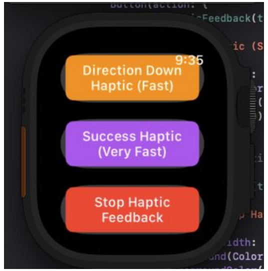
Midterm Status and Goals

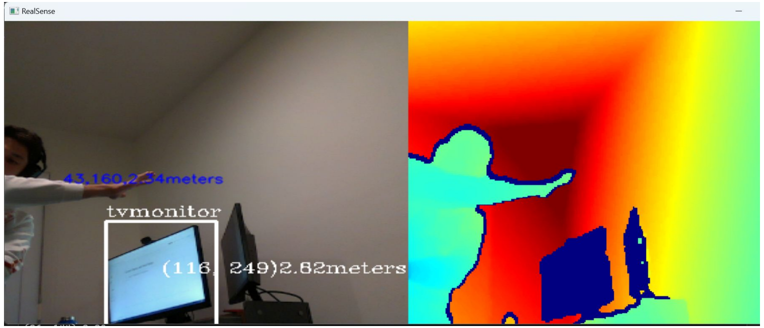
Midterm status

- Haptic feedback on solo watch app
- Depth map from camera
- Hand tracking and object distance detection
- Body pose landmarks and face mesh
- Face pose vector computation relative to camera

Next steps

- Speech to text recognition and keyword identification
- Directional audio feedback
- Directional determination and spatial angle computations
- Command conditions for directional audio and distance haptic
- Actuation command communication to watch + earphones
- UDP Server/Client Interface and payload structure
- Better object detection model
- Integrating all the different components together





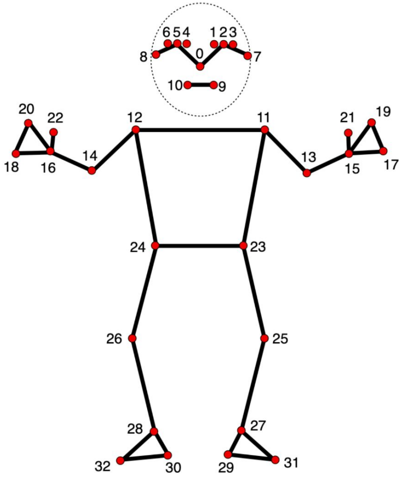
Pose and Head Angle

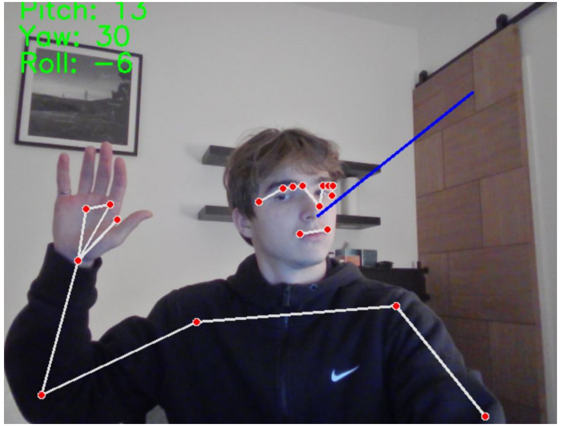
MediaPipe Pose Landmarker:

- Pre-trained ML kit, returns human pose landmarks
- Accurate landmark detection even back to camera
- Estimated, normalized z-axis value for each landmark

Forward angle computation:

- Pseudo-3D forward vector calculated from midpoint of eyes to nose
- Forward vector projected to x-z plane
- Angle of head relative to camera computed using arctan
- Direction head is facing relative to camera as angle in the range [0,360]

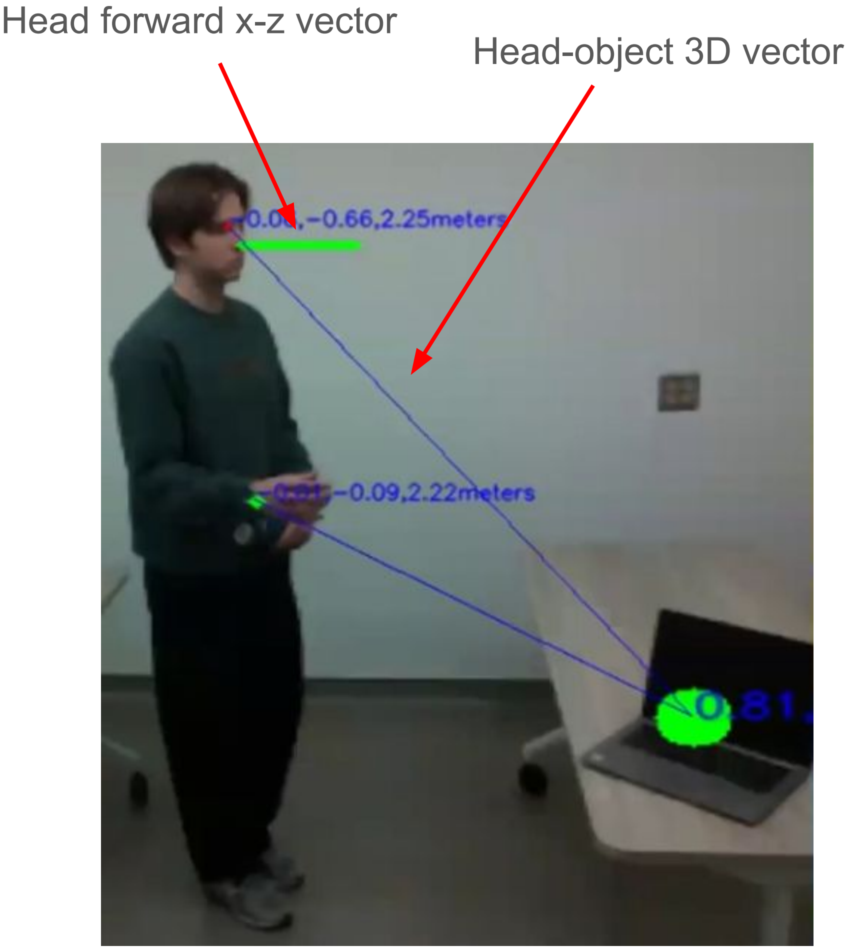




Head-Object Angle

Head-object angle computation:

- Head 3D position by deprojecting the midpoint of the eyes using depth and camera intrinsics
- Classified object 3D position by deprojecting in the same way
- 3D head-object vector obtained from subtracting head and object 3D points
- Head-object angle computed with arctan of x-z components of head-object vector
- [0,360] range like head forward angle for direct comparison



Directional Determination and Distance

Angle smoothing:

- Exponential moving average applied to head angle and head-object angle
- Low-pass filtering for less noisy directional determination

Directional determination:

- Direct comparison of the head angle and the head-object angle
- If absolute difference < 30, forward command
- If head angle < head-object angle - 30, right command
- If head angle > head-object angle + 30, left command

Wrist-object magnitude:

- Wrist pixel position extracted from Pose Landmarker
- Wrist position deprojected to 3D space using depth and camera intrinsics
- 3D wrist-object vector constructed, Euclidean distance of wrist-object vector is computed

Challenges:

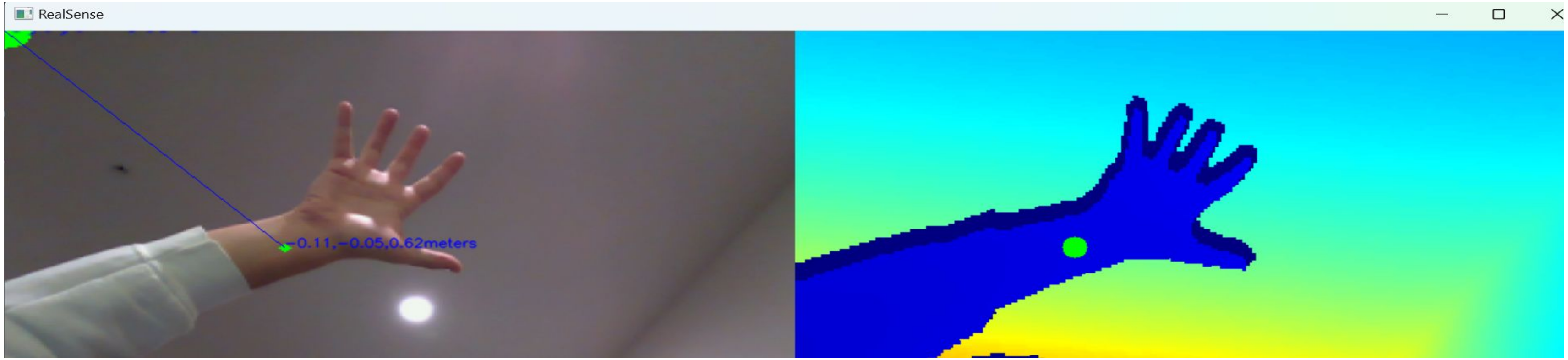
- Noisy depth map
- Non-optimized direction (left vs right) due to [0,360] angle mapping
- Inaccurate wrist 3D position when blocked by body

Laptop RealSense Camera Application

- Realsense Camera Color + Depth data collection
- UDP Server interfacing
  - Sending Direction, handDistance
  - Receiving object keywords
- MobileNetSSD/YOLOv7 Object Detection
  - Key Object detection
  - 2D to 3D Pixel to Point Deprojection
- Hand and Object Depth/3D deprojected point calculation
  - Using MediaPipe Pose/YOLOv7 landmarks

Laptop RealSense Camera Challenges

- Poor Object Detection Model
- 3D vector visualization
- Misaligned Frames(Depth and Color)
- Inaccurate Depth Measurements
- UDP Challenges on UCLA Network
- Data Quality Problems



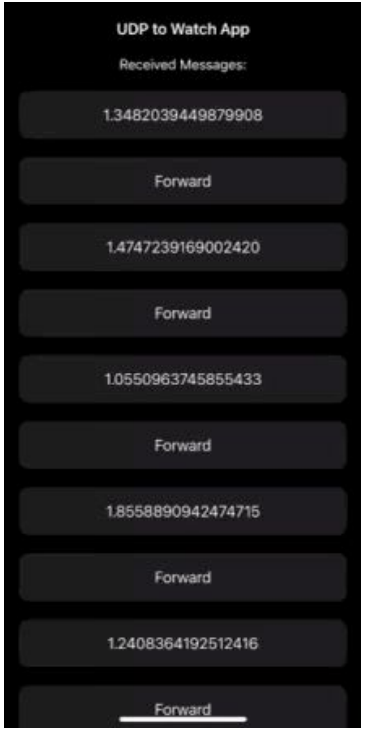
Speech-to-text & Audio Guidance

- Key Technical Details:
  - Transformer based Language Model for Speech Recognition
  - Generate data for model training (likely phrases with keywords)
  - Internet Connection required
  - Integration of Live Audio input and playback (AVFoundation AudioEngine)
  - On recognition of keyword, send request to laptop server that handles scene analysis
  - Wait for server to start broadcasting directional guidance, then play audio accordingly
  - 4 distinct audio files for instructions: left, right, forward, wait
  - Communication over Wi-Fi
- Challenges:
  - Spatial Audio Track generation
  - Maintaining audio coherence with rapidly changing directional guidance
- Future Goals:
  - Add on the fly spatial audio generation / modulation

Haptic Guidance

- Key Technical Details:
  - Powered by Taptic Engine
    - Based on an LRA design, which creates vibrations using a moving mass driven by an electromagnetic coil. (better haptic precision and energy efficiency)
  - Haptic Waveforms
    - Customizable patterns, called haptic waveforms, can emulate taps, pulses, or continuous vibrations.
- Method:
  - Haptic feedback only starts when wrist is within 1.5m of the object
  - Actuated using input from server that broadcasts UDP packets, each containing direction and a distance metric

Phone App Overview



Speech-to-text detects commands like "bottle" via Speech Recognition.

UDP communication ensures low-latency navigation data exchange.

Voting system processes directional data every 0.5 seconds.

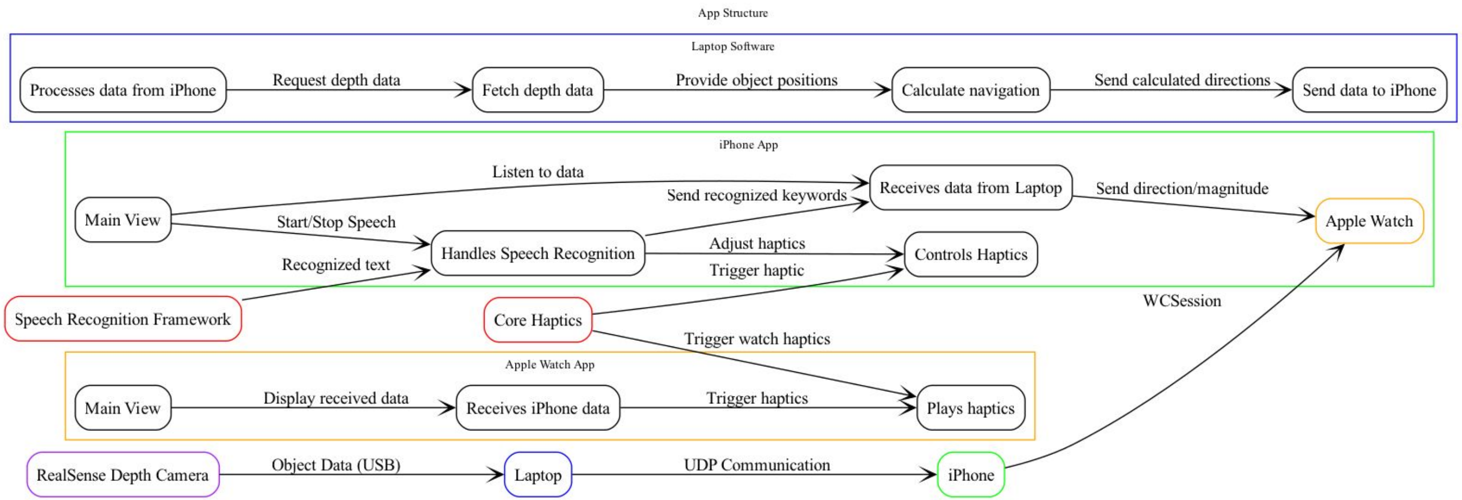
Magnitude filtering refines haptic feedback intensity.

Audio cues guide navigation, with headphone support for privacy.

Watch App Overview

- Vibrations strengthen as the hand nears the object.
- Fine-motor haptics complement phone's audio instructions.
- Built with WatchKit and WCSession for fast communication.
- Focused on precision guidance for blind users, no visuals needed.

Integration



- Combines UDP communication, WCSession, speech recognition, haptics, and depth camera data for navigation.
- Depth Camera sends 3D spatial data to the laptop for precise positioning and distance computation.
- Laptop processes Depth Camera data to calculate directions and sends commands to the phone via UDP.
- Phone relays commands to the watch through WCSession for haptic and audio feedback.
- Modular design ensures scalability, with each component (depth camera, laptop, phone, watch) integrated seamlessly.
- Speech-to-text on the phone enables real-time command adjustments and personalized navigation.

Results/Findings/Lessons Learned

- Generally accurate head/object/hand positioning + orientation computation
- Angle and vector computations give accurate directional commands to guide subject towards object
- Directional audio feedback is effective at guiding subject towards object (macro movements)
- Distance haptic feedback could be more nuanced to guide wrist to object (micro movements)
- Filtering and majority voting systems improve directional command reliability, reducing noise in real-time feedback.
- Data Metrics
  - Haptic Feedback Precision: Haptic feedback helped users locate objects within a 1-meter radius in 70% of trials, based on 20 tests measuring both time and accuracy.
  - UDP Latency: Recorded an average round-trip latency of 8.0 milliseconds, with values ranging between 6.8 ms and 9.2 ms across 20 trials.These measurements were obtained by logging timestamps at message transmission and reception on both devices
  - Distance Feedback Calibration: Tested vibration strength scaling by measuring user perception at different hand distances, validated across 20 trials to ensure intensity changes. (0.2m strong feedback, 1.0m medium feedback, 1.5m weak feedback)
  - Command to Response Delay(Auditory command -> iPhone -> Laptop -> Object position computation -> Iphone -> Watch) - 3.53 seconds

Conclusion/Future Direction

- Viable proof of concept prototype
  - Latency from real-time direction loop manageable
  - Precise and moderately accurate head/object/hand 3D positioning
  - Accurate orientation computation
  - Quick object detection and command interpretation
- Future Direction
  - More complex object classification (multiple objects, more object types)
  - Expand x-z direction computation to y-axis too
  - More complex commands (e.g. “take me to the concierge”, “tell me where the elevator is”)
  - Spatial audio feedback
  - Multiple cameras for better point cloud generation
  - Complex path planning

Implementation and Collaboration

Alex	Pose detection and spatial computations	Pose landmarker, head angle, head-object angle, direction command conditions, wrist-object distance
Arshia	Object 3D positioning + Server wrapper	Realsense Camera Data Collection, Implemented 2 object detection models, created laptop server UDP interface, generated object/hand/head 3D deprojected point
Dhruv	Speech Recognition	Implemented the speech recognition system, enabling accurate keyword detection and integration with the navigation pipeline
Yiteng	Watch & phone App	Developed the watch and phone apps, integrating UDP communication and audio/haptic feedback for navigation.