You talkin’ to me? - An attention-aware embodied agent

Rahul R. Divekar\textsuperscript{1,2}, Jeffrey O. Kephart\textsuperscript{2}, Lisha Chen\textsuperscript{1}, Xiangyang Mou\textsuperscript{1}, Hui Su\textsuperscript{1,2}

\textsuperscript{1}Rensselaer Polytechnic Institute (RPI), Troy, NY
\textsuperscript{2}IBM, Yorktown Heights, NY
Conversational AI - The Spectrum

Interface: Text Only
Embodiment: Phones, Tablet, Laptop

Interface: Voice, Buttons
Embodiment: Special Hardware

Interface: Voice, Touch
Embodiment: Phones, Cars, etc.

Voice based

* Images from Messenger, Slack, Amazon, Google Home, Tesla, Android
Voice Conversational AI - Types of Conversations

- **One-shot commands**
  - Human: “Alexa, turn on the lights”

- **Multi-round dialogue**
  - Human: “Hey Google, remind me to buy groceries”
  - AI: “When would you like to be reminded?”
  - Human: “Tonight”
  - AI: “Sure”
Voice Conversational AI - You talkin’ to me?

- Wake up word - “Alexa, Hey Google, etc.”
- Extended attention span
Conversational AI - The Spectrum

**Interface:** Text Only
**Embodiment:** Cellphone, Tablet, Laptop

**Interface:** Voice, Buttons
**Embodiment:** Special Hardware

**Interface:** Voice, Touch
**Embodiment:** Phones, Cars, etc.

* Images from Messenger, Slack, Amazon, Google Home, Tesla, Android
Conversational AI - The Spectrum

**Interface:** Text Only
**Embodiment:** Cellphone, Tablet, Laptop

**Interface:** Voice, Buttons
**Embodiment:** Special Hardware

**Interface:** Voice, Touch
**Embodiment:** Phones, Cars, etc.

* Images from Messenger, Slack, Amazon, Google Home, Tesla, Android
Conversational AI - The Spectrum

Interface: Text Only
Embodiment: Cellphone, Tablet, Laptop

Interface: Voice, Buttons
Embodiment: Special Hardware

Interface: Voice, Touch
Embodiment: Phones, Cars, etc.

Interface: Text, voice, body
Embodiment: On screen in sensor-equipped environments

Current Focus

* Images from Messenger, Slack, Amazon, Google Home, Tesla, Android
Smart Environments - Typical Features

- **Interface**: Voice + Body + Text
- Multi-modal dialogue
- Multi-round dialogue
- Multi-person dialogue
Smart Environments - Typical Dialogue

Participants: H1, H2 and AI

H1 -> AI: Celia, show me a plot of exoplanets

H2 -> H1: I’d like to see planets bigger than earth

H2 -> AI: Celia, can you plot planets bigger than earth?

....
Smart Environments - Typical Dialogue

**Participants:** H1, H2 and AI

H1 -> AI: *Celia*, show me a plot of exoplanets

H2 -> H1: I’d like to see planets bigger than earth

H2 -> AI: *Celia*, can you plot planets bigger than earth?....

H2 -> AI: *Celia*, change the x-axis to log scale

H2 -> AI: *Celia*, tell me more about star x

H2 -> AI: *Celia*, show me planets closest to earth

H2 -> AI: *Celia*, plot the age against mass
Voice Conversational AI - You talkin’ to me?

- Wake-up word - “Alexa, Hey Google, Celia etc.”
- Extended attention span
Voice Conversational AI - You talkin’ to me?

- Wake-up word - “Alexa, Hey Google, Celia etc.” ------- Unnatural, cumbersome
- Extended attention span  Side conversation misinterpreted as command
Voice Conversational AI - You talkin’ to me?

- Wake up word - “Alexa, Hey Google, Celia etc.” Unnatural, cumbersome
- Extended attention span Side conversation misinterpreted as command

“Sorry, I do not know how to do that”
Question - How do you know someone is talking to you?
Solution - Find out where the user is looking

- Analogous to making eye contact in human-human conversations
- Use head orientation as a rough estimation
- If, user looks at AI agent
- Then, treat overlapped utterance as command (thresholded)
User Study Setup
Findings #1

- Technology shortcomings
  - Head orientation tracking failure
  - Lag in interpreting change of head orientation
  - Transcription inaccuracies

- User behaviour
  - Head orientation shifts when
    - Trying to recollect
    - Reading from a page
    - Looking for help from partner
    - Looking at other elements on screen (e.g. plot)
    - Stroking chin while thinking (self-occlusion from camera)
  - Example: Show me planets with radius greater than <change in headpose to human participant> umm... <chin stroking> Earth
Findings #2

Good news: Users are willing to change their behavior to help the system infer attention.

E.g. Change in articulation, wait to clear backlog of ASR, etc

Catch: Inform the user about the system’s interpretation of attention (analogous to eye-contact reciprocation)
Watson Avatar (no attention)  

Paying attention (Verbal or Head-orientation)  

Face Detection Error  

Face Detected! Look this way to get attention  

Transcript: I want the results of the meeting  

No Attention  

Transcript: Celia show me a plot of exoplanets  

Attention and Intent  

Transcript: Celia this is not a command  

Attention and No Intent
Transcript: Celia show me a plot of exoplanets.
Outcomes - A test bed for attention awareness

- Multi round interaction
- Multi people interaction
- Multi modal interaction
Further Direction

- Formal usability studies
- Explore other behavioral cues
- Sophisticated models to fuse multi-modal data in real time
- Data collection of interactions
- Using machine learning to make attention decisions
Acknowledgements

Rensselaer Polytechnic Institute (RPI)
Cognitive Immersive Systems Lab (CISL)
IBM
AI Horizons Network (AIHN)
Questions?