Human-Computer Interaction

2024/2025

Lecture Class 14

Voice and Gesture Interaction

universidade de aveiro
deti departamento de electrónica, telecomunicações e informática

Speech is our most natural form of communication

Voice Input

Has advantages when the user:

- Has physical deficiency
- Must move around
- Has hands/eyes busy
- Is in a low visibility or cluttered environment

Voice Input

Has inherent disadvantages:

- Voice is transient
- May disturb other people
- May result in lack of privacy
- May be slower and more tiresome (overloading STM)

Command-based VUI

Users must learn specific phrases/keywords to trigger actions

- Predictable and reliable
- Require memorizing syntax
- Often follows: wakeup word + command + parameter
- Low flexibility, but good recognition

Okey Google, set timer for 10 minutes

Natural Language VUI

Users can express the same intent in multiple ways using everyday speech

- More intuitive and conversational
- Handles variations, synonyms, …
- Requires understanding what is being said
- More prone to errors

Can you remind me in 10 minutes?

Make it 15, instead.

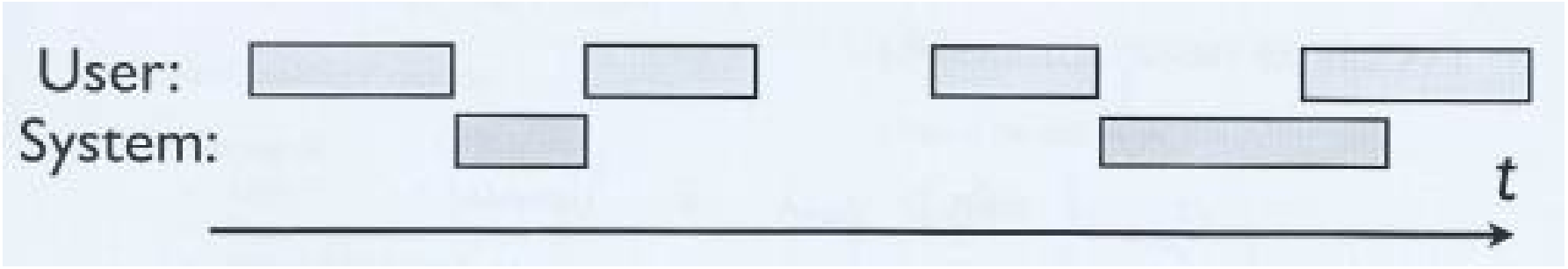
Human Conversation

C_1:	. . . I need to travel in May.
A_1:	And, what day in May did you want to travel?
C_2:	OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_2:	And you're flying into what city?
C_3:	Seattle.
A_3:	And what time would you like to leave Pittsburgh?
C_4:	Uh hmm I don't think there's many options for non-stop.
A_4:	Right. There's three non-stops today.
C_5:	What are they?
A_5:	The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
C_6:	OK I'll take the 5ish flight on the night before on the 11th.
A_6:	On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_7:	OK.

Human Conversation

Turn-taking

- Dialog is characterized by turn-taking
- Resource allocation problem (only one channel)



Dialogue Acts

	Utterance	Dialogue act
U:	Hi, I am looking for somewhere to eat.	`hello(task = find,type=restaurant)`
S:	You are looking for a restaurant. What type of food do you like?	`confreq(type = restaurant, food)`
U:	I'd like an Italian somewhere near the museum.	`inform(food = Italian, near=museum)`
S:	Roma is a nice Italian restaurant near the museum.	`inform(name = "Roma", type = restaurant, food = Italian, near = museum)`
U:	Is it reasonably priced?	`confirm(pricerange = moderate)`
S:	Yes, Roma is in the moderate price range.	`affirm(name = "Roma", pricerange = moderate)`
U:	What is the phone number?	`request(phone)`
S:	The number of Roma is 385456.	`inform(name = "Roma", phone = "385456")`
U:	Ok, thank you goodbye.	bye()

Grounding

Participants in conversation or any joint activity need to establish common ground.

Grounding: acknowledging that the hearer has understood

Grounding

- A: And you said returning on May 15th?
  C: Uh, yeah, at the end of the day.
  A: OK

- C: OK I'll take the 5ish flight on the night before on the 11th.
  A: On the 11th? OK.

- C: ...I need to travel in May.
  A: And, what day in May did you want to travel?

Confirmations
Explicit

User: I'd like to fly from Denver Colorado to New York City on September 21st in the morning on United Airlines

System: Let's see then. I have you going from Denver Colorado to New York on September 21st. Is that correct?

User: Yes

Confirmations

Implicit

U: *I'd like to travel to Berlin*

S: When do you want to travel to Berlin?

U: *Hi I'd like to fly to Seattle Tuesday morning*

S: Traveling to Seattle on Tuesday, August eleventh in the morning. Your name?

Implicit vs Explicit

- Complementary strengths

Explicit

- Easier for users to correct system's mistakes (Can just say "no")
- But is cumbersome and long

Implicit

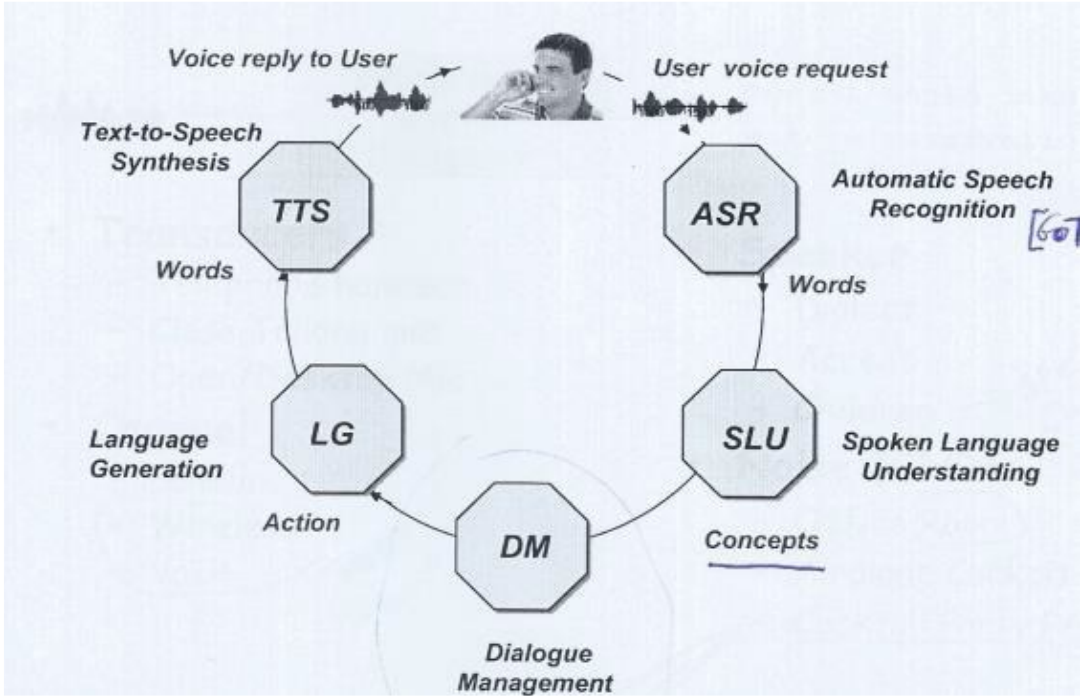
- Much more natural, quicker, simpler (if system guesses right).

Spoken Dialog

Between Humans and Machines

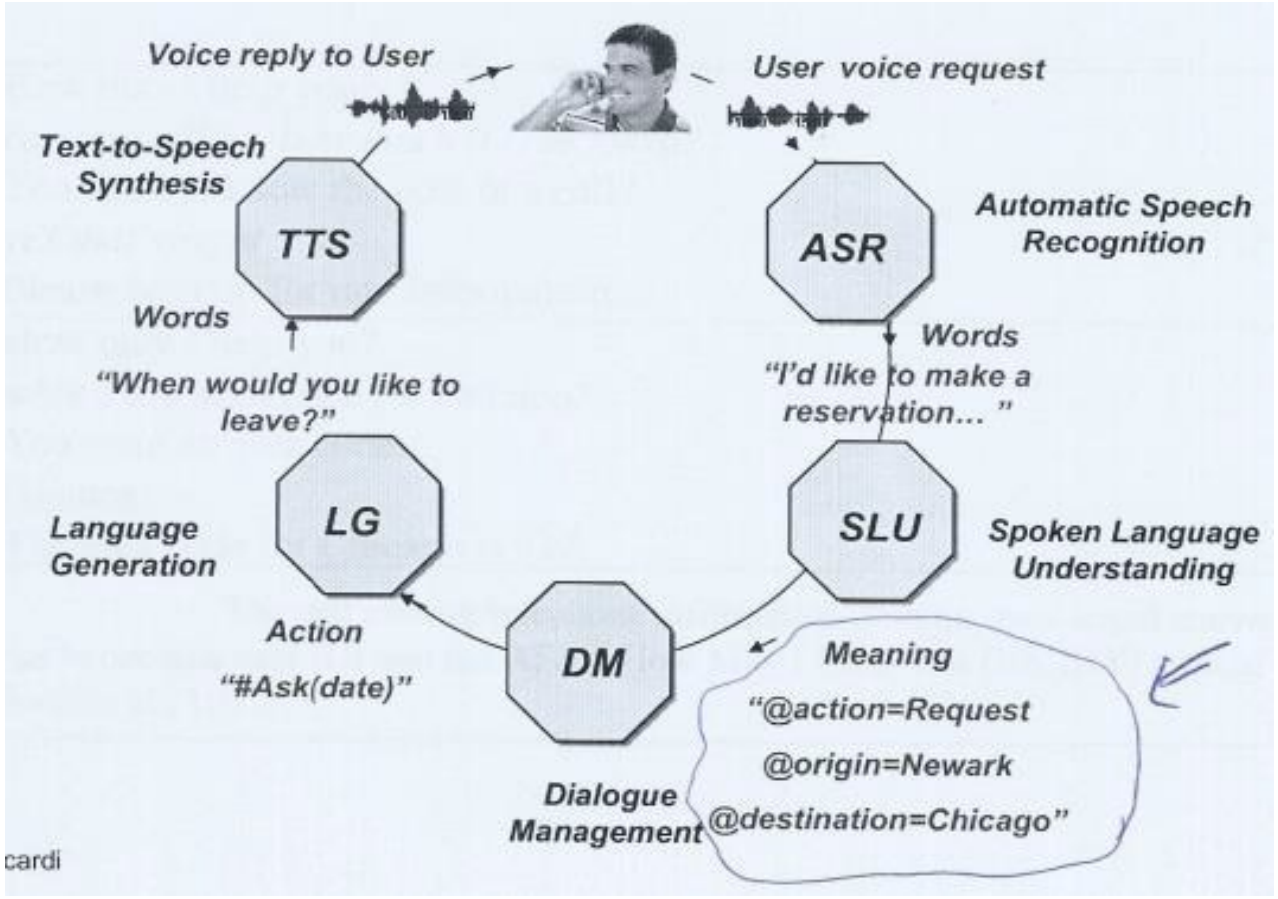
Challenges of Spoken Dialog

- detect that someone is speaking …
- … recognize the words spoken …
- … understand the meaning of words …
- … Interpret meaning in dialog context …
- … decide what the system should say …
- … formulate what is to be said …
- … articulate (or synthesize) the speech



Example

- **User**: yes I would like to make a reservation on the afternoon flight from New York to Chicago…
- **ASR**: I'd like to make a reservation on the afternoon fly from Newark to Chicago
- **SLU**:
  - Action Determination (e.g. @action=Request)
  - Named Entities (e.g. "@city=Newark")
  - Semantic Model associated with world model (e.g. database)
    - @departure_city="New Work"
    - @arrival_city="Chicago"



Dialogue Systems
(conversational agents, dialog agents, chatbots)

- Systems designed to interact with humans via conversations (either in text or speech)
- Personal Assistants on phones or other devices
  - SIRI, Alexa, Cortana, Google Assistant
- Playing music, setting timers and clocks
- Chatting for fun
- Booking travel reservations
- Clinical uses for mental health

Initiative

who has control of conversation

In normal human-human dialog, initiative shifts back and forth between participants

Systems that control conversation are known as system initiative or single initiative

System Initiative

- SYSTEM: Where do you want to go?
- USER: Paris
- S: From where do you want to go?
- U: London
- S: At what time do you want to leave?
- U: Three o’clock
- S: Which date do you want to leave?
- U: On the first of august

https://botnation.ai/site/wp-content/uploads/2022/01/chatbot-1-445x265.png

System Initiative

Pros and Cons

Advantages:

- Simple to build
- User always knows what they can say next
- System always knows what user can say next
  - Known words: Better performance from ASR
  - Known topic: Better performance from NLU
- Ok for very simple tasks
  - (entering a credit card, or login name and password)

Disadvantage:

- Too limited

User Initiative

User directs the system

Generally, user asks a single question, system answers

System can't ask questions back, engage in clarification dialog, confirmation dialog

Used for simple database queries

- User asks a question, system gives an answer

Example:

- Web search is user initiative dialog

Mixed Initiative

- Initiative can shift between system and user
- Simplest kind of mixed initiative:
  - use structure of a frame to guide dialog
- Goal is fill in the slots by asking the questions

Slot	Question
ORIGIN	What city are you leaving from?
DEST	Where are you going?
DEPT DATE	What day would you like to leave?
DEPT TIME	What time would you like to leave?
AIRLINE	What is your preferred airline?

Based on what the user provides, the system takes the initiative to fill in missing slots

Design and Development of Spoken Dialog Systems

https://media.istockphoto.com/id/1226991733/photo/successful-motivational-speaker-talking-about-happiness-self-success-empowerment-efficiency.jpg?s=612x612&w=0&k=20&c=TLkDQ_C0xFeQCJQM7UrVAEeKDY0HPEkaeQ-wOfh0Fbk=

Setting User Expectations

You need to be careful to keep the user in the “ right spot”

Don’t ask a question if you won’t be able to understand the answer

Example: After user finishes writing an email…

- SYSTEM: Do you want to send it or change it?
- USER: yes.
- SYSTEM: Errr what? Errr (this system will self destruct in 5 seconds)…

Method: Sample Conversations

How do we start development?

One of the best (and cheapest!) ways:

- Pick five of the most common use cases for your Assistant …
- Then write some “blue sky” dialogs for each case
- In addition, write a few for when things go wrong

Method: Sample Conversations

Try reading the conversations with a colleague impersonating an assistant

Simulate the system using a Wizard of Oz approach

Iteratively test the design

Confirmations

Making sure that users feel understood is an important part of any good VUI design: Explicit? Implicit? Hybrid ?

Onboarding

Introduce User to the capabilities of the conversational assistant – very important!! Helps moderate expectations

For example:

- “Hi, my name is Billy and I am here to assist you with your accessible touristic routes. I can help you find an hotel, such as “I want an hotel in Aveiro” or answer questions about accessibility as in “Is Hotel Meliá, in Aveiro, accessible using wheelchairs?”

Conversational Markers

VIRTUAL ASSISTANT

How many hours of sleep did you get last night?

USER

About seven.

VIRTUAL ASSISTANT

How many servings of fruits and vegetables did you eat yesterday?

USER

Maybe four.

VIRTUAL ASSISTANT

Did you take your medication last night?

USER

Yes.

VIRTUAL ASSISTANT

Goodbye.

VIRTUAL ASSISTANT

I'll be asking you a few questions about your health. First, how many hour of sleep did you get last night?

USER

About seven.

VIRTUAL ASSISTANT.

Good job. And how many servings of fruits and vegetables did you eat yesterday?

USER

Maybe four.

VIRTUAL ASSISTANT

Got it. Last question—were you able to take your medication last night?

USER

Yes.

VIRTUAL ASSISTANT

All right. That's it for now. I'll talk to you again tomorrow. Goodbye.

What are the differences?

Conversational Markers

Conversational markers are an important way to let the user know where they're at in the conversation, and that they are understood.

Include:

- Timelines ("First," "Halfway there," and "Finally")
- Acknowledgments ("Thanks," "Got it," "Alright," and "Sorry about that.")
- Positive feedback ("Good job," and "Nice to hear that")

Error Handling

Provide sensible strategies to tackle errors and things your system does not know how to answer.

“Sorry, I did not understand your last sentence, can you say it differently?”

“My apologies, but I do not have information about raising engineers in captivity.”

Error Handling

And you can take the chance to continue onboarding:

“At the moment I can only answer solicitations such as “I need an accessible museum in Aveiro”.

And don’t blame the user:

“Learn how to speak, dummy!”

vs

“I apologize, but I am still learning and was not able to understand your intentions. Could you say it differently, please?”

Voice Output

Advantages of using voice output:

- physical difficulties
- to be able to move around
- hands and eyes busy
- Adverse conditions: low visibility, high Gs

Disadvantages:

- tiresome and uncomfortable for long periods
- transient (taxes STM)
- May have privacy issues
- May disturb other people


And, often, speech communication is not alone

Gestures

Gestures are a form of
nonverbal communication

As speech, they allow
interacting at a distance

What is a Gesture

A gesture is any physical movement that a digital system can sense and respond to without the aid of a traditional pointing device such as a mouse or stylus.

- Examples: A wave, a head nod, a touch, a toe tap, and even a raised eyebrow can be a gesture.

Gestures include movement of the hands, face, or other parts of the body.

Types of Gestures

Touch gestures
Tap, swipe, pinch, rotate (direct manipulation)

In-air gestures
Hand/arm movements detected by cameras or sensors

Full-body gestures
Physical movements tracked by systems like Kinect

Micro-gestures
Subtle finger/hand movements for wearables

Public Installations

Hygiene concerns

Less mechanical parts for people to break...

Sophistication effect

Can potentially accommodate more users

Medical Settings

Infection control

- Surfaces are not touched and do not need to be disinfected

Maintaining sterile environment

- A surgeon can access information during a procedure without having to go through sterilization, again

Smart Homes

Interacting with multiple devices, quietly

Gaming

https://www.researchgate.net/publication/307894274_Diverse_Techniques_for_Geometric_Features_Extraction_and_Identification_of_Gesture_Recognition/figures?lo=1

eXtended Reality

Natural embodiment

Intuitive object manipulation

Controller-free interaction

Enhanced presence

Automotive

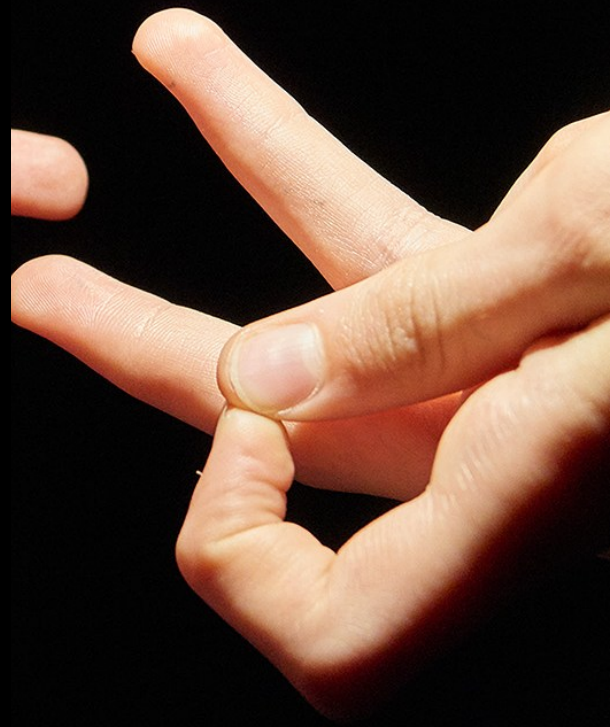
Control systems without taking eyes off the road

Reduce cognitive load.. If well designed

Customizable among drivers

BMW iDrive

Gesture Interaction Design Principles



https://freight.cargo.site/t/original/i/b5f3b33b224be42cf822798999cd5c89835c3d10801af6e2bf24eea9cb6b0105/gestured_opening_hands1_SW_REPLACE.jpg

Should this be a gestural Interface?

Yes!

When:

More natural interaction

- Humans are used to interacting directly w/ objects

Less cumbersome or visible hardware

- e.g., no need for keyboards, mice; small or no display

Flexibility

- e.g., gestures in a smartwatch to send messages

Fun

- Gesture interaction can be more engaging (e.g., Wiimote)

Should this be a gestural Interface?

No!

When:

Heavy data input
- e.g., keyboards are faster than gestures

Reliance on the visual
- Many gestural interfaces rely on visual feedback

Reliance on the physical
- Doing gestures can be more tiring and physically demanding

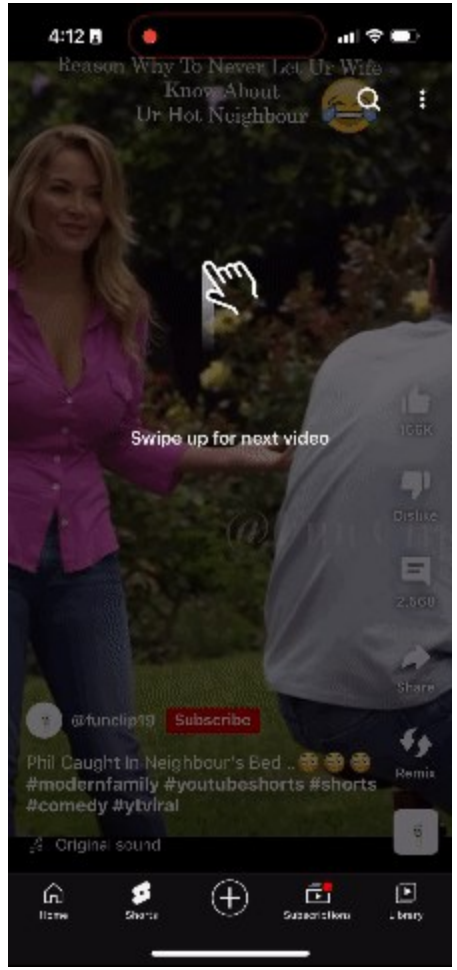
Inappropriate for context
- Doing gestures in public can be felt as embarassing

Discoverability

Visual affordances suggesting the possibility of gestures

Advanced gestures progressively introduced

On-boarding



Side note:

Findability vs Discoverability
https://www.nngroup.com/videos/findability-vs-discoverability/

Feedback

Visual confirmation of gesture recognition

Adequate response time between gestures and system response

Multimodal feedback

- Use multiple feedback channels including sound and haptics (when available)

Consistency

Similar gestures produce similar outcomes

Are there any platform conventions?

Consistent response times and thresholds for recognition

- e.g., a certain extent of a swipe movement should always produce the same outcome and take the same time to provide feedback and trigger the corresponding action

Memorability

Leverage physical metaphors connected to actions

Limit number of unique gestures

Spatial and directional consistency can help

- e.g., down always decreases

Cultural Differences

Avoid gestures with negative connotations for the target culture(s)

- E.g., stop/hi-five gesture is offensive in Greece

Cultural differences in spatial navigation

- E.g., swipe left to move to the next page; but what about in Arabic books?

Test gestures across cultures

- Even simple gestures (e.g., showing 3 with your hand) can have differences

Gesture for 3 across cultures

Social Acceptability

The use of gestures in public spaces needs to be carefully analyzed

Learning Curve

Working memory limitations

- Gestures need to be remembered

Mental model formation of spatial-temporal mappings

Onboarding strategies

- Progressive disclosure of gesture complexity
- Upfront training vs just-in-time
- …

Transfer effects from previous experiences

Error tolerance

Performance of gesture recognition depends on several factors

- Type of sensor, environmental conditions, static or dynamic gestures, continuous or segmented recognition, computational demand

Transferring existing gesture recognition methods across domains and user groups can result in poor performance

Consider how errors affect user experience

- Provide feedback about system confidence and recover strategies
- Doing gestures demands physical effort. Too much failures will lead to disengagement and frustration

Personalization

Users have different motor control and physical capabilities

- Not anyone can dance to a rhythm…

Handedness / dominant-side preferences

- Not only about the hands… feet too

Scaling gestures to different reach capabilities

- Also reducing motion extent for fatigue-prone users

Gestural Interface Design Stages

https://www.quantamagazine.org/wp-content/uploads/2019/03/Gesture-2880x1620_Lede-Poster.jpg

Understanding the Problem

Perform user research:

- How users interact with similar systems
- Physical capabilities and limitations of the target users
- Cultural differences and aspects limiting acceptability
- Environmental constraints (light, space, public place…)
- Explore scenarios and task analysis to understand the tasks and their context

Gesture Elicitation

What is the best match between a gesture set and system functions?

Place users in a relevant context and ask how they would gesture to do _____?

https://miro.medium.com/v2/resize:fit:1200/1*xRxPJ8Ft4Vg_lUumtkVofg.gif

Prototyping and Testing

- Sometimes, it can be useful to support gesture elicitation and testing through low fidelity prototypes
- Wizard of Oz approaches are common and useful
- Implement low cost solutions (e.g., using a webcam)
- Pay attention to fatigue…

Gorilla Arm Effect

Fatigue, discomfort and muscle strain experienced when interacting with mid-air gesture interfaces (or vertical touchscreens) for extended periods

If asked to suggest gestures, people tend to favour gestures where the forearm is closer to the body

For instance, for a mid-air swipe gesture, how much do you stretch your arm forward?

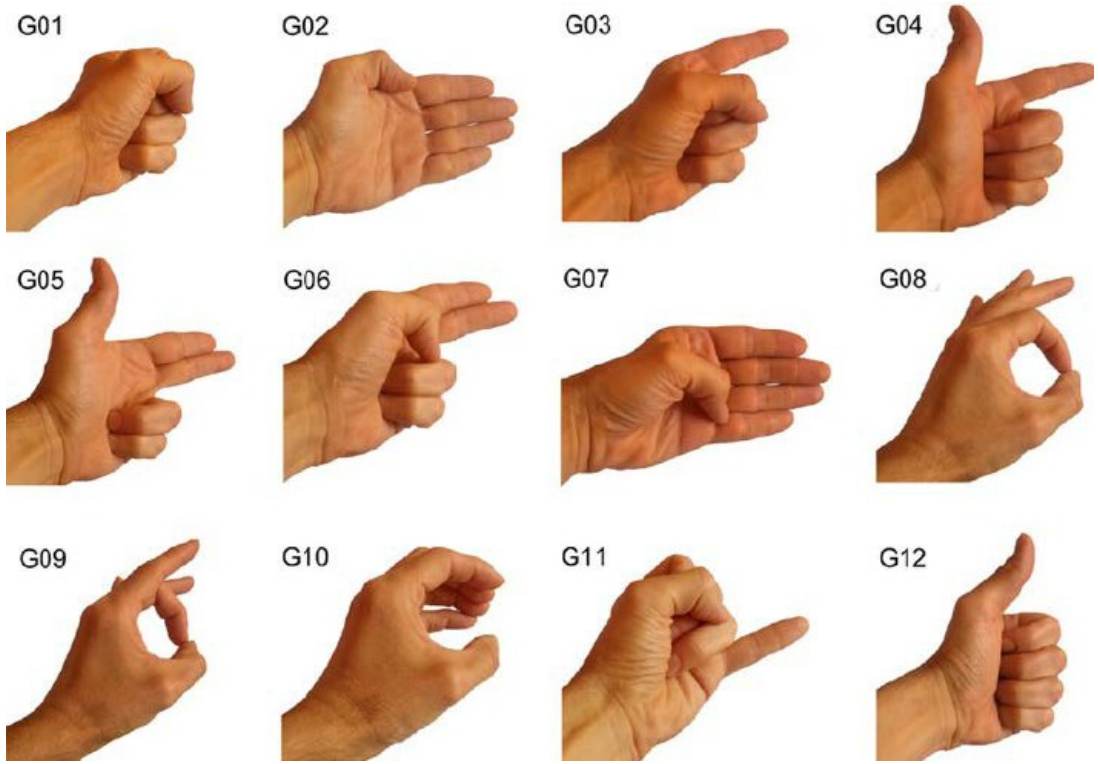
Gestures suggested by users to interact with the home in study @ IEETA

Documenting

Build a gesture dictionary for the design and development team and to ship with the product

Create gesture representations to explain them to users, during user research

Support user onboarding



Documenting

What should be documented?

How do users direct communication to a system?

How does the system establish that it is ready for input?

What can be done with the system?

How does the system respond?

How does the system avoid or recover from errors or misunderstandings?

Documenting

A first level of gesture documentation conveying information on their use context can be provided by methods you already know:

Scenarios

Task Analysis

However, these descriptions lack a visual component that is key for us to understand the nature of the gesture

Documenting

Storyboarding

Tell a “story” using still images of notable elements of user interaction in context

[Saffer, 2008]

Documenting

Swim Lanes

Complements storyboards with what is happening at different levels of the system

User approaches the kiosk		User touches anywhere on kiosk screen	User taps any location on the map	User moves away from the kiosk
Kiosk is in idle mode	Idle screen changes	Idle screen changes to main screen	Map zooms in	Kiosk returns to idle mode
Proximity sensor triggers at 3 meters	Display active idle screen	Display main map with location data	Display detail map with location data	Display idle screen if proximity sensor does not detect presence
Advertising space available		Advertising space available	Advertising space available	

[Saffer, 2008]

Top lane
The storyboard itself, including narrative

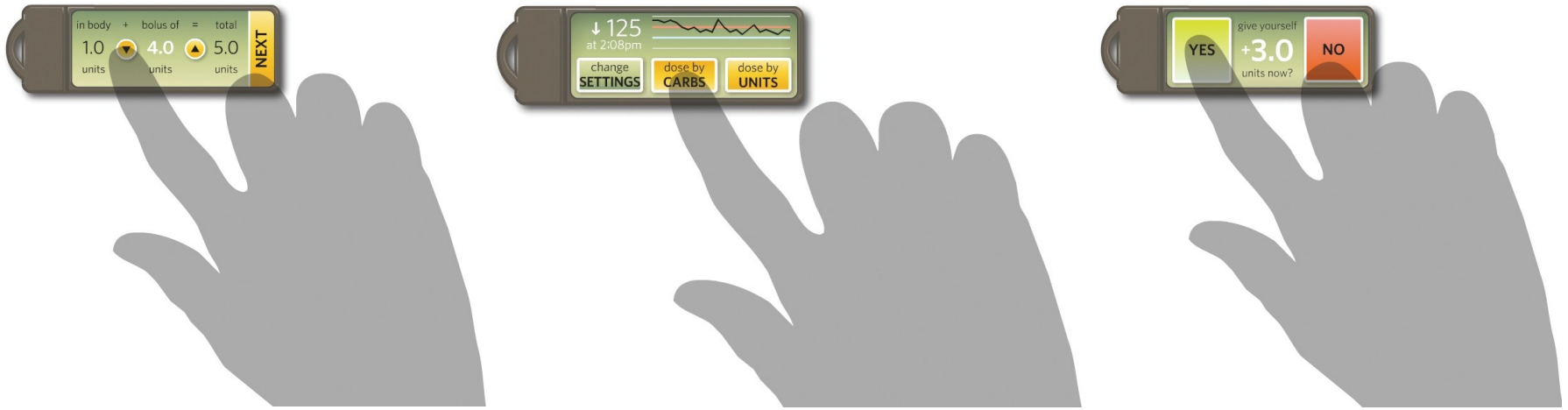
Second lane
On-screen changes

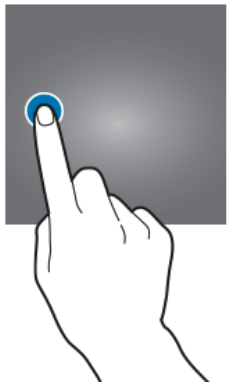
Third lane
System flow

Bottom lane
Business processes

Animations

Sometimes, gestures are better depicted through animations





Gesture Interaction Supporting Technologies

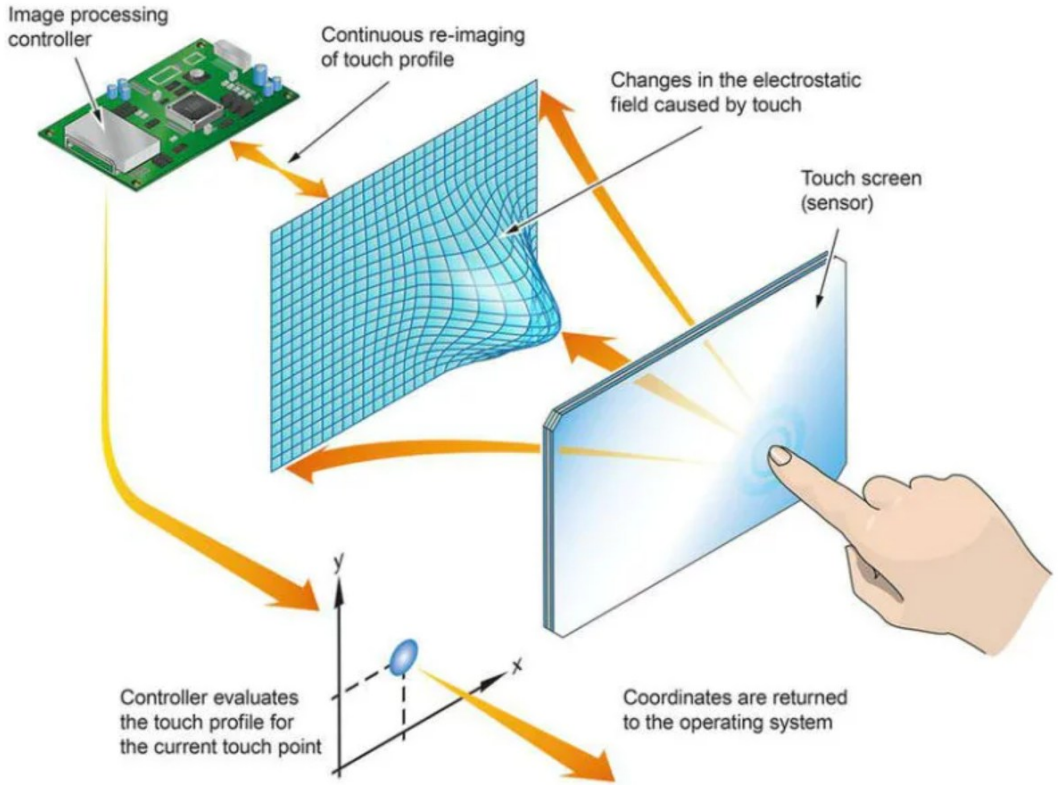
Capacitive/Resistive/IR

Capacitive: human body created distortion in electrostatic field

Resistive: Physical pressures causes layers to touch

Infrared grid interrupted by touch

Integration with haptic feedback



Computer Vision

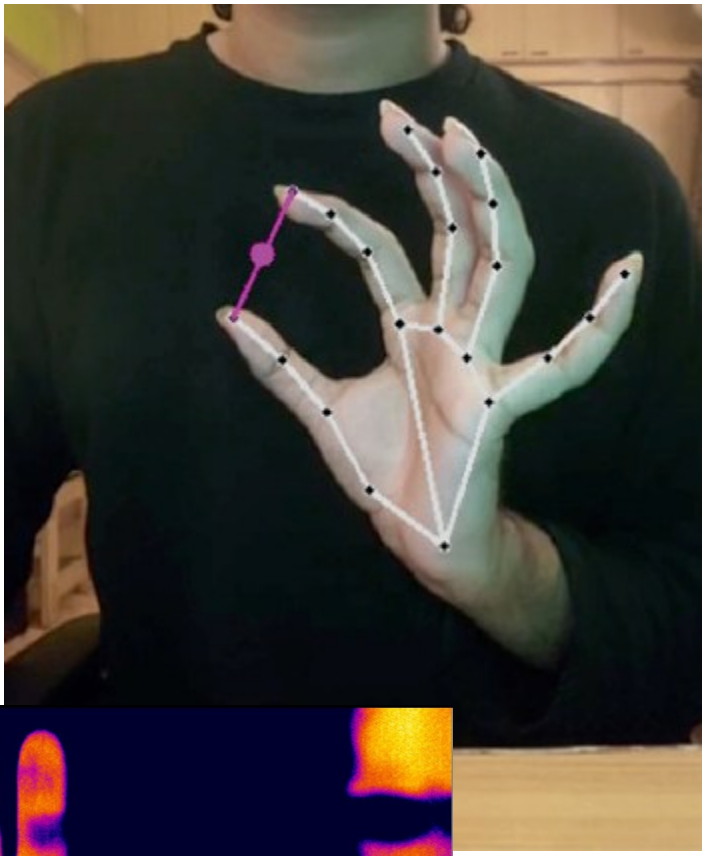
Camera-based tracking approaches

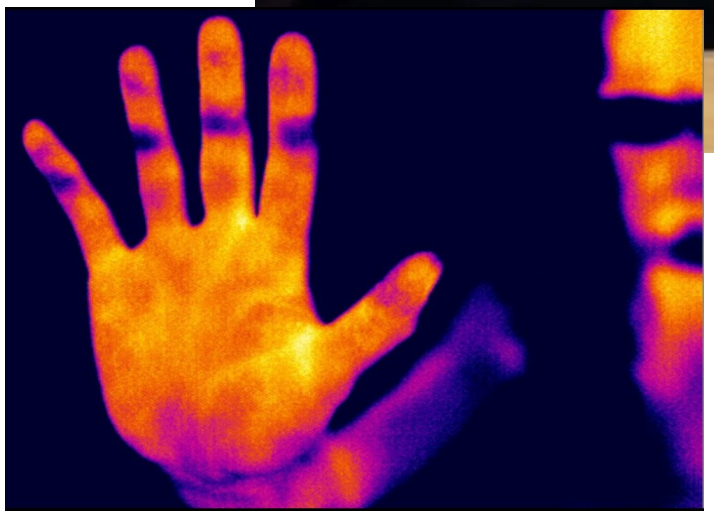
Options include thermal and RGB cameras

- mind environmental conditions and privacy

Entails hand/body detection, feature extraction (contours, fingers, …)

- Tracking across video frames
- Handling occlusion





Depth

- Provides 3D data to recognize gestures
- Different approaches: structured light, time-of-flight (Kinect), stereo depth (binocular disparity)
- More resilient to low ambient light
- Sometimes fused with RGB image

Greylevel value depicts distance to depth camera.
Darker is closer.

https://ieeexplore.ieee.org/document/9000925

Accelerometer/Gyroscope

Accelerometers, gyroscopes, and magnetometers

Wearables

- e.g., watch, finger or arm bands

Handheld controllers

Embedded in everyday objects

Muscular activity

Muscle motion generates electrical signal

Implies wearable (e.g., electrodes or band)

Can also detect subtle gestures (without visible movement, just muscle contraction)

https://www.embs.org/tnsre/articles/deep-learning-for-electromyographic-hand-gesture-signal-classification-using-transfer-learning/

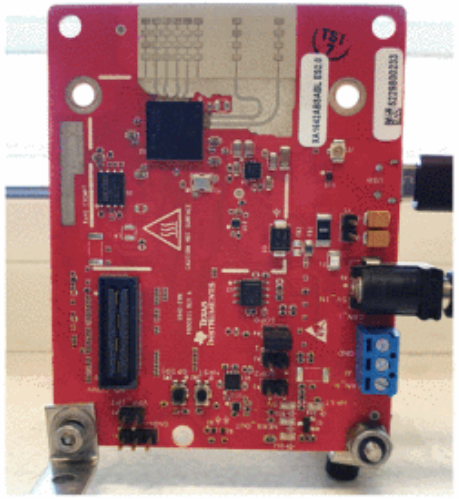
Radar

Leverages frequency-modulated waves to detect the position and velocity of targets

Works regardless of lighting conditions and through non-metallic objects

Interesting for privacy-sensitive spaces (e.g., bathroom)

More complex to work with than cameras



Research on gesture recognition using radar @ IEETA

Trends in Gestural Interfaces

Gestures in Context

- Activity recognition to predict relevant gestures
- Social context detection for appropriate gesture sets
- Proximity-based interaction zones
- Adaptation based on user behavior patterns

Ubiquitous Sensing

- Gesture recognition hardware and features as part of the environments (objects, rooms, fabric…)
- Thermal/radar sensing for privacy and through wall interaction
- Natural user movements as implicit interaction

Standardization

- Common gesture vocabularies across devices
  - “Turning on” gesture being the same for the TV, radio, lamp
- Reduce barriers between diverse hardware
  - Gestural interface design is still a lot about the hardware being used
- User testing protocols and performance metrics
  - There is a need for a common ground for assessing and comparing gestural interfaces

Integration with other Modalities

There is a strong synergy between gestures and speech

Multimodal interactive systems implement these features through what is called modality fusion*

Gestures can play different roles: they can be complementary (saying “put that there” while pointing to the object and location) or redundant (e.g., saying “I agree” while doing a thumbs up)

* Multimodal Interaction is an elective course available at DETI, dealing with this and other aspects of multimodal systems

Bolt’s seminal work: Put That There (1980)

https://www.youtube.com/watch?v=CbIn8p4_4CQ&ab_channel=MITMediaLab

https://www.youtube.com/watch?v=RyBEUyEtxQo&ab_channel=StefanMarti

https://www.media.mit.edu/publications/put-that-there-voice-and-gesture-at-the-graphics-interface/

Bibliography

- Dan Jurafsky, James H. Martin, ‘Speech and Language Processing’, 3rd ed, 2023 (chapter 15)
- Cathy Pearl, ‘Designing Voise User Interfaces’, O’Reilly Media, Inc., 2016 (chapter 2)
- Dan Saffer, Designing Gestural Interfaces, O’Reilly, 2008 https://learning.oreilly.com/library/view/designing-gestural-interfaces/9780596156756/
- Google Material Design: Gestures https://m3.material.io/foundations/interaction/gestures
- Cambridge Handbook of Gesture Studies, Cambridge University Press, 2024