

# Multimodal Interaction













**Human Computer Interaction**

Luigi De Russis, Fulvio Corno

Academic Year 2021/2022

# Designing for Diversity (recap)

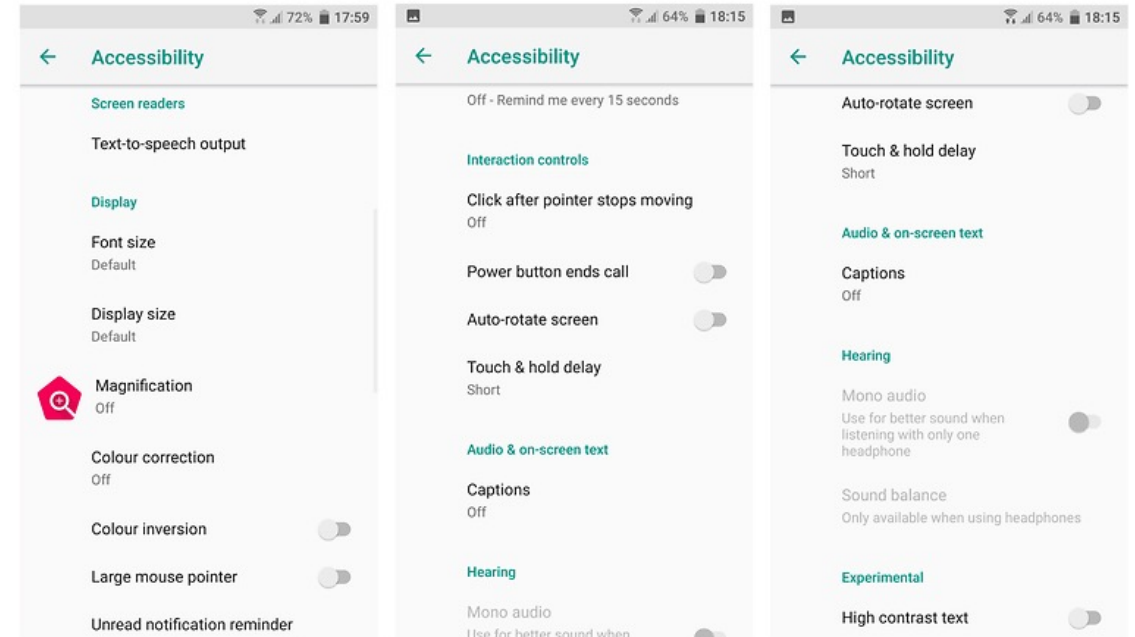
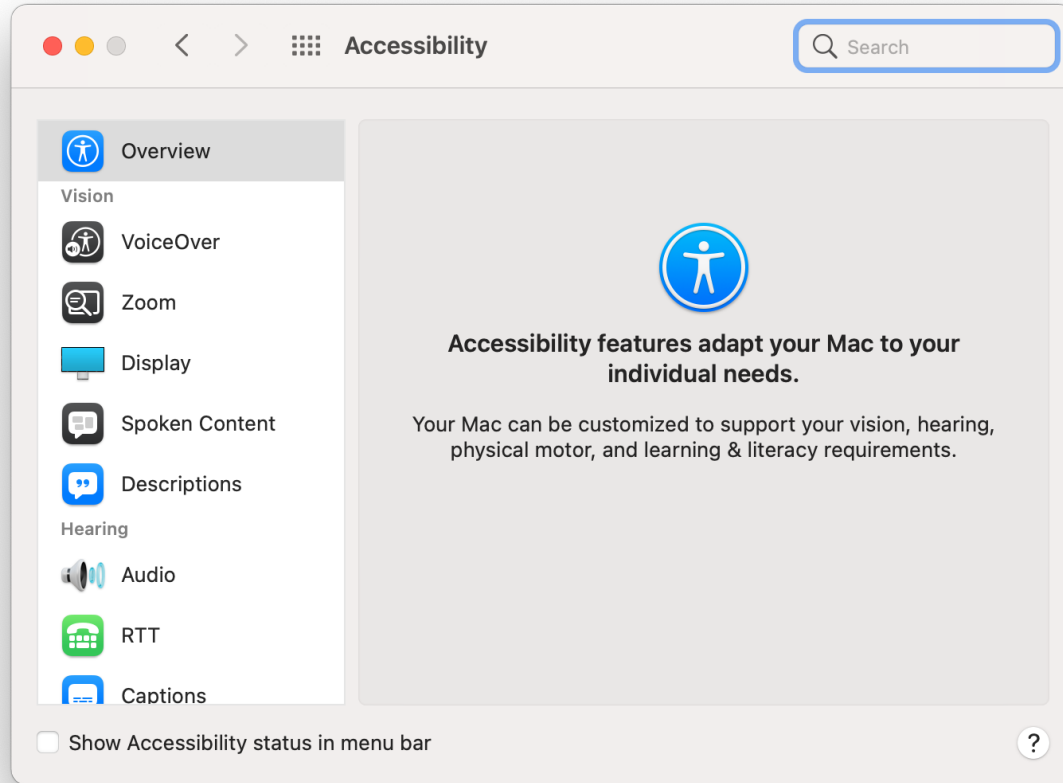
- The interactions we design with technology depend heavily on what we can understand/remember, see, hear, say, and touch
- Assuming all those senses and abilities are **fully** enabled **all** the time means ignoring several people
  - it also reflects how people really are, as "life happens"
- We want our designs to reflect that diversity

	Permanent	Temporary	Situational
Touch	 One arm	 Arm injury	 New parent
See	 Blind	 Cataract	 Distracted driver
Hear	 Deaf	 Ear infection	 Bartender
Speak	 Non-verbal	 Laryngitis	 Heavy accent

# Multiple Senses and Abilities

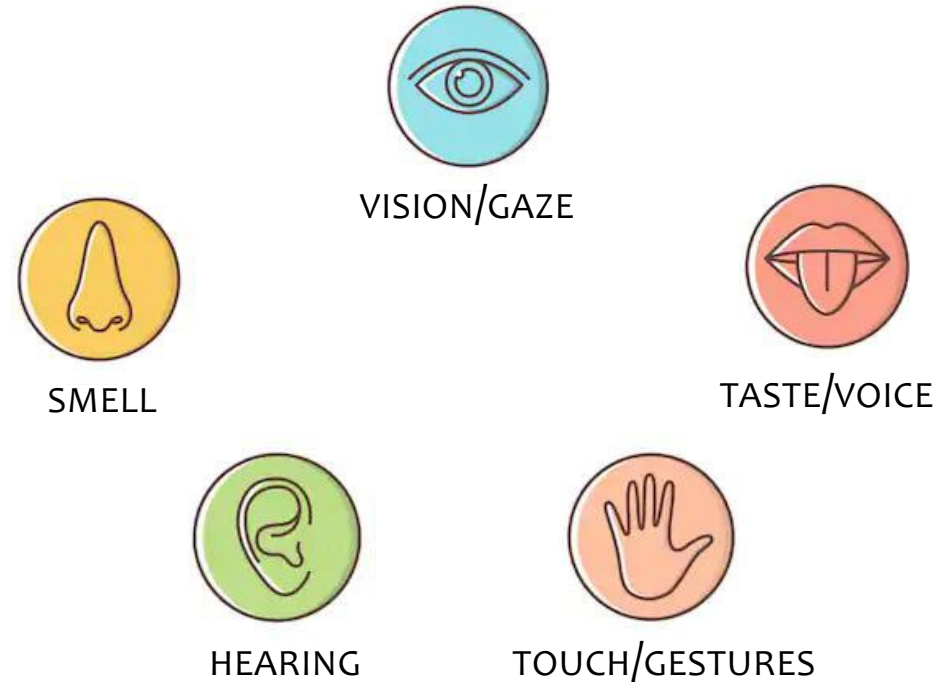
- Can we design an application or a system that leverages on multiple senses and abilities at the same time?
- Maybe providing different input/output mechanisms in different contexts and for different people?
- How?
  - redundancy
  - compatibility with assistive technologies
  - ...

# Example: Accessibility in OS



# Multimodal Interaction

**Definition:** To use more than one sensory channel or mode of interaction



Can we use all of these?

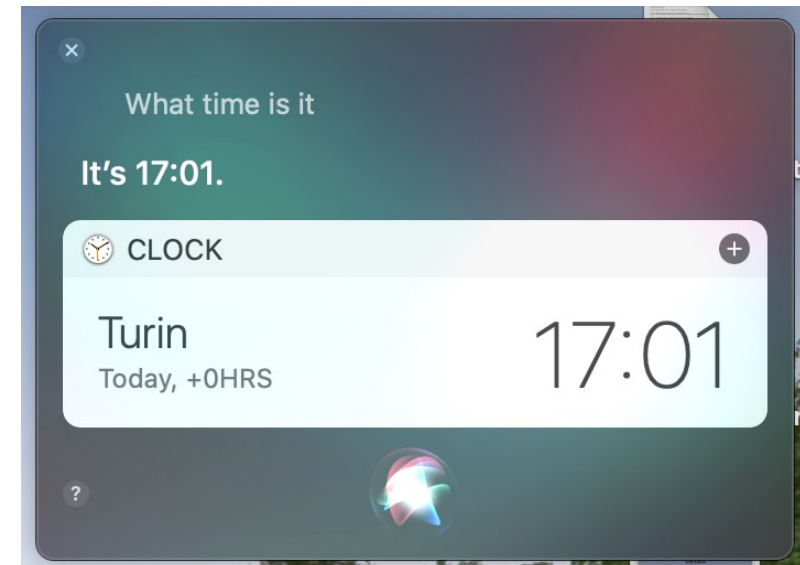
# Multimodal Interfaces Around Us

- Most interactive systems are predominantly **visual**
  - often WIMP based, they make use of simple sounds while adding more and more visual information to the screen
- As systems become more complex, the visual channel may be overloaded if too much information is presented at once
  - this may lead to frustration or errors in use
- Using multiple modes increases the *bandwidth* of the interaction
  - we should always remember that multi-modal interaction is not just about enhancing the richness of the interaction, but also about *redundancy*

# Multimodal Interfaces Around Us: Examples



Vision + Gesture + Hearing + Speech



Vision + Gesture + Hearing + Speech

# Multimodal Interfaces: Why (Not)? — LIVE

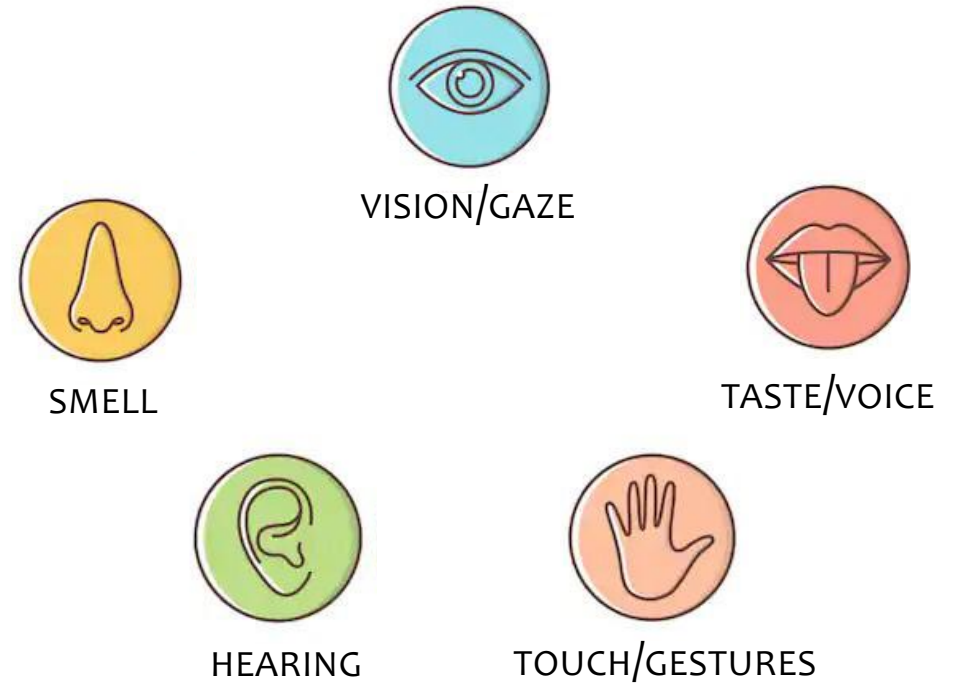
## Advantages

- Enabling more people to successfully use our application
  - improving the application for many
  - e.g., accessibility
- Making the system available in different situations and environments
- Redundancy and adaptation
- Emphasizing critical feedback and fixing in memory specific critical actions

## Disadvantages

- Too many senses (redundant) can bring to overload
  - e.g., with people with autism or in specific contexts
- Difficulty to understand from where a feedback/output comes from
  - where my attention is needed in multi-tasking app
- Since we are accustomed to it, the learning curve could be steep
- More effort is needed to design and develop a multimodal interface w.r.t. a GUI





# In depth...

Characteristics of the various senses/modes of interactions

# Vision



- **Human vision:** a highly complex activity, often the main source of information about the world
- The **eye:** a mechanism for receiving light and transforming it into electrical energy
  - light reflected from objects in the world and their image is focused upside down on the back of the eye
  - then, the receptors in the eye transform it into electrical signal which are passed to the brain
  - the brain detects, finally, pattern and movements
- *Disclaimer:* we are going to ignore "vision", since most of the course focused on it so far...

# Gaze



- Can we control a computer solely with the eyes?
  - Yes, with **eye-tracking** (or eye-gazed) interfaces
- A person's gaze can be used in a variety of ways to control user interfaces, typically through dedicated hardware components
  - alone or in combination with other input modalities (mouse, keyboard, ...)
- Main application areas:
  - to find more efficient and novel ways to facilitate the interaction for users with disabilities, who can use only their eyes for input
  - to use real-time eye tracking data to understand human behaviors, and to explore novel user interfaces

# Eye Trackers

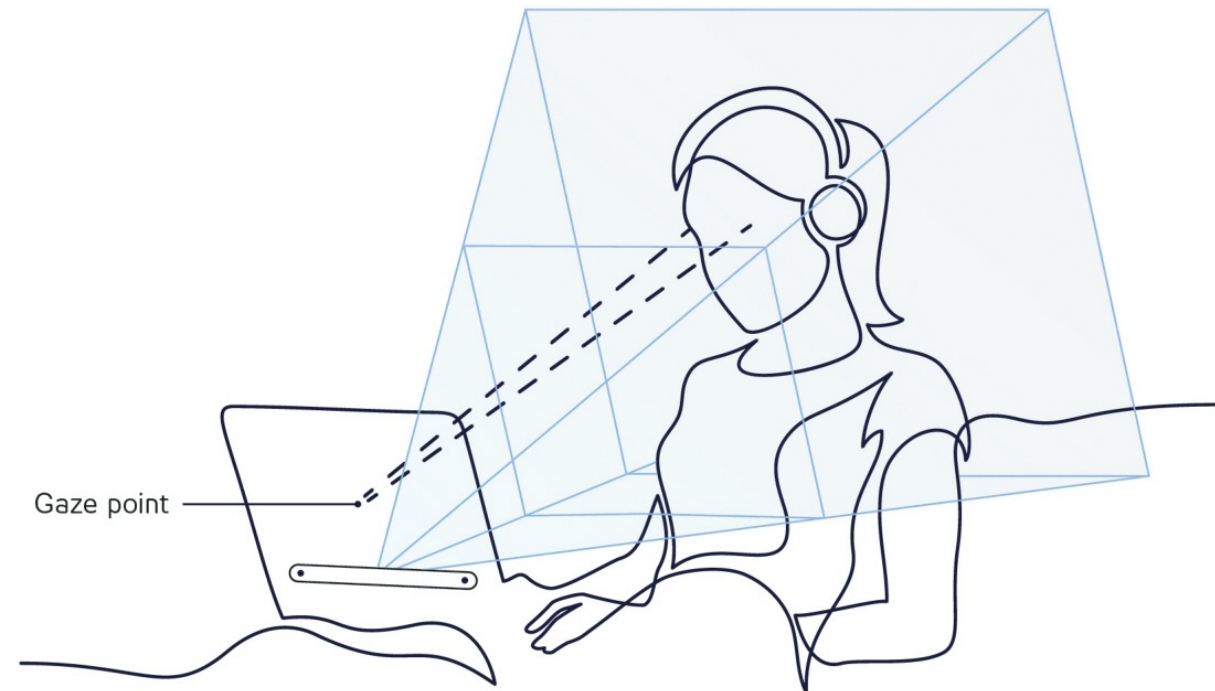


source: <https://www.tobii.com>

# Eye Tracker: How It Works



- 1** An eye tracker consists of cameras, projectors and algorithms.
- 2** The projectors create a pattern of near-infrared light on the eyes.
- 3** The cameras take high-resolution images of the user's eyes and the pattern.
- 4** Machine learning, image processing and mathematical algorithms are used to determine the eyes' position and gaze point.

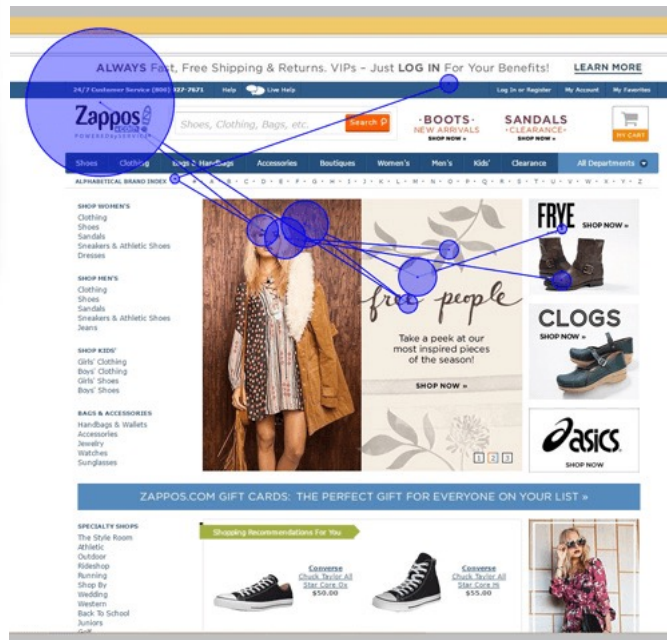


source: <https://www.tobii.com/group/about/this-is-eye-tracking/>

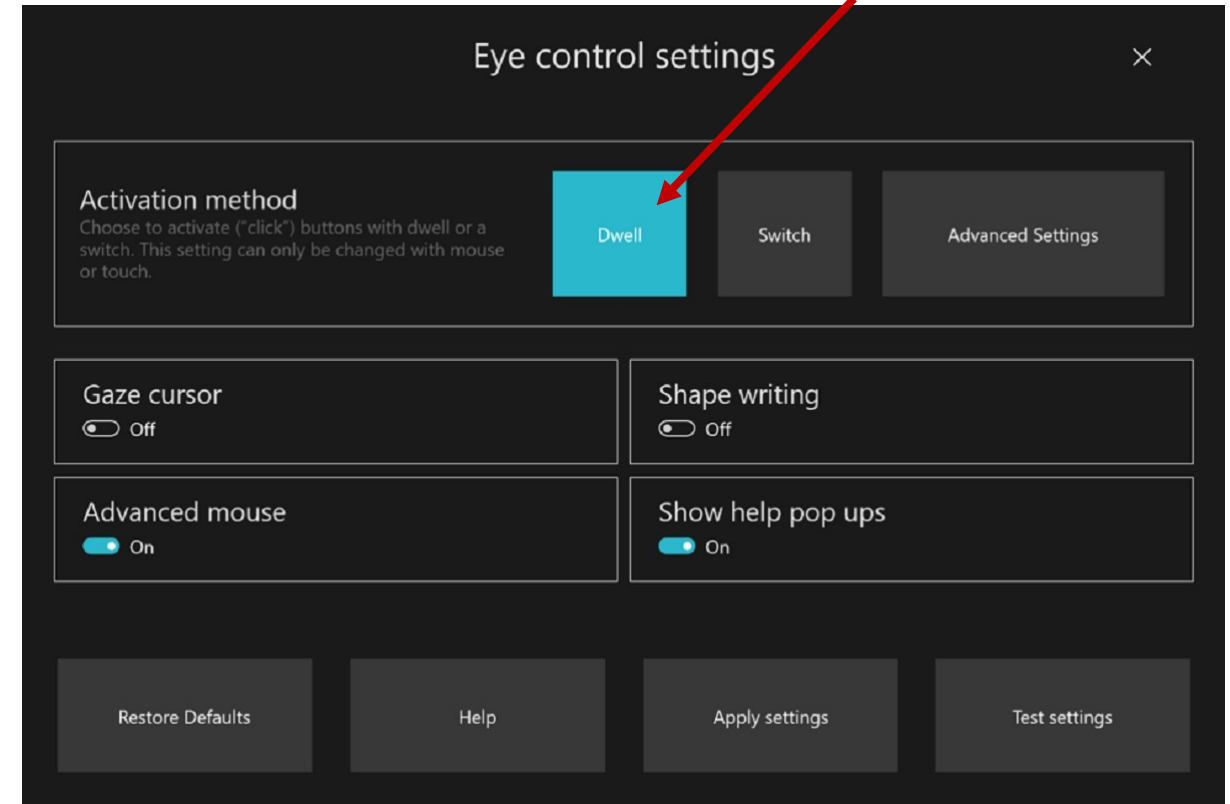
# Eye Tracker: Examples



Heatmap and scan path analysis

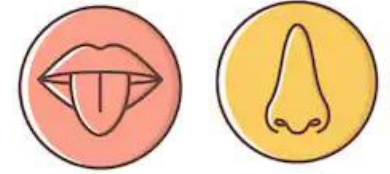


Beware the Midas' Touch!



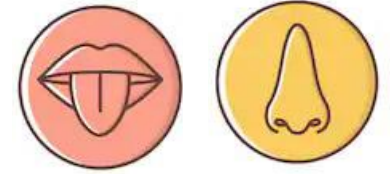
Windows 10 Eye Control Settings

# Smell and Taste



- **Traditionally**, they do not play a role in HCI
- However, humans have around 12 million **olfactory** receptor cells, able to detect around 10,000 different odors
- We are born with 10,000 taste buds on our tongue, the roof of the mouth, and in our throats. Each taste bud has about 10-50 cells are responsible for starting the action of taste and are replenished about every 7 to 10 days.
  - Bad news: we naturally start to lose these taste buds around 50 to 60 years of age
- Both smell and taste provides us with a very important **early-warning system** when it comes to objects or situations that may cause us harm
  - think about expired or... burning food

# Multi-Sensory Interaction (HCI Research)

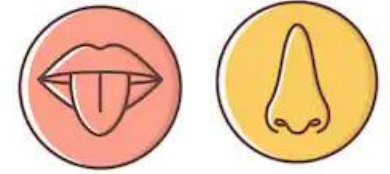


- In-car Olfactory Interaction
  - Photo: <https://multi-sensory.info>
- Delivering different scents to car drivers to indicate danger or points of interest

Dmitrijs Dmitrenko, Emanuela Maggioni, Chi Thanh Vi, and Marianna Obrist. 2017. What Did I Sniff?: Mapping Scents Onto Driving-Related Messages. In Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '17). ACM, New York, NY, USA, 154-163. DOI: <https://doi.org/10.1145/3122986.3122998>



# Multi-Sensory Interaction (HCI Research)



- TastyFloats
  - Photo: <https://multi-sensory.info>
- A novel system that uses acoustic levitation to deliver food morsels to the users' tongue

Chi Thanh Vi, Asier Marzo, Damien Ablart, Gianluca Memoli, Sriram Subramanian, Bruce Drinkwater, and Marianna Obrist. 2017. TastyFloats: A Contactless Food Delivery System. In Proceedings of the 2017 ACM International Conference on Interactive Surfaces and Spaces (ISS '17). ACM, New York, NY, USA, 161-170. DOI: <https://doi.org/10.1145/3132272.3134123>



# Touch and Gestures

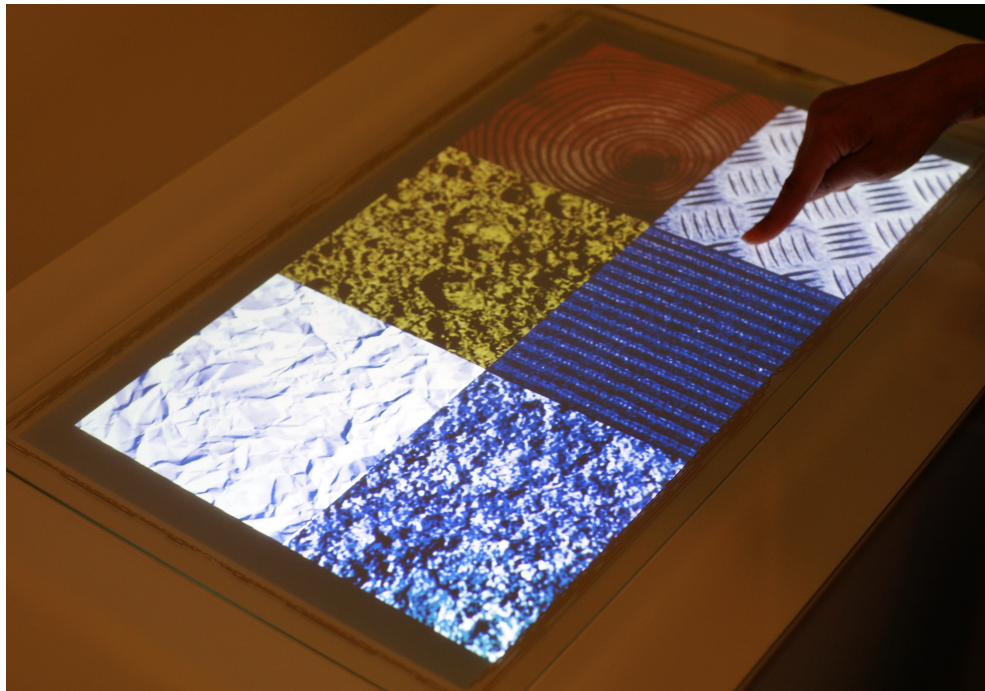
- Touch, or *haptic perception*, acts as means of feedback and it provides us with information about our environment:
  - information on shape, texture, resistance, temperature, comparative spatial factors
- Gestures (and hands/body movements, in general) are more and more used to control and provide inputs to computers
  - think about the "device with touch screen" in your pocket

# Haptic Interaction: Examples



- Braille Displays
  - They allow people who are blind to read, through haptics, the braille code present on screen
  - They exist in various formats: the most used are the 32 and 40 characters
  - All displays need a dedicated software, a screen reader

# Haptic Interaction: Examples



- TeslaTouch
  - Electro vibration for Touch Surfaces
  - A new technology for enhancing touch interfaces with tactile feedback
  - It can simulate a feeling of textures and materials by sliding fingers over their images (as in the picture)

Olivier Bau, Ivan Poupyrev, Ali Israr, and Chris Harrison. 2010. TeslaTouch: electrovibration for touch surfaces. In Proceedings of the 23rd annual ACM symposium on User interface software and technology (UIST '10). ACM, New York, NY, USA, 283-292. DOI: <https://doi.org/10.1145/1866029.1866074>

# Gesture: Examples (Camera-based)



BMW in-car gesture control

<https://www.youtube.com/watch?v=ttTBJkE-6fs>



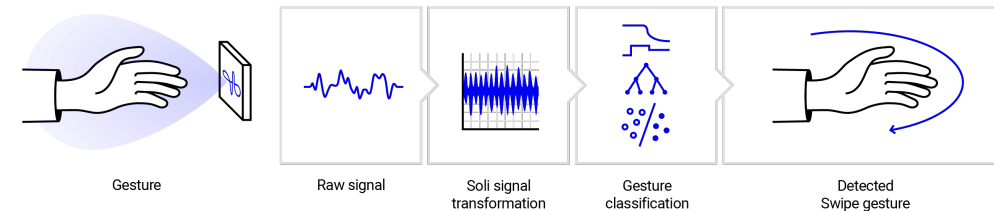
Leap Motion

<https://www.youtube.com/watch?v=rnlCGw-oR8g>

# Gesture: Examples (Radar-based)



- Motion Sense on the Pixel 4 phone
  - Project Soli by Google ATAP
    - <https://atap.google.com/soli>
  - Soli is a miniature radar that understands human motions at various scales (powered by ML)





# Hearing

- Sound can convey a remarkable amount of information about our environment (try to close your eyes and listen...)
- The **ear** can differentiate quite subtle sound changes and can recognize familiar sounds without concentrating attention on the sound source
- Rarely used in UI design, mostly confined to notifications, warnings, and typing sounds
  - multimedia is, obviously, an exception



# Non-Vocal Sounds

- Often used to provide transitory information, such as warning and events
- They must be "learned" but they are language-independent
  - boing, ting, squeak, ...
- They are useful! Experiments demonstrate:
  - auditory cues are adequate and help navigating either on a screen or in an immersive virtual environment
  - fewer typing mistakes with key clicks
  - video games are harder without sound





# Non-Vocal Sounds: Which Ones?

- Use **natural** sounds to represent different types of object or action
  - *Auditory Icons*, born in the early 1980's by Bill Gaver for Apple's Finder
  - basically, they are caricatures of naturally occurring sound
- Natural sounds have associated semantics which can be mapped onto similar meanings in the interaction
  - problem: not all things have associated meanings
- Additional information can also be presented:
  - muffled sounds, if object is obscured or action is in the background
  - use of stereo allows positional information to be added

# Auditory Icons: Examples



- Let's fill this together!



# Non-Vocal Sounds: Which Ones?

- Use synthesized, **structured** sounds to represent a specific event or signal information
  - *Earcons*, a pun on the more familiar term icon. Icon sounds like "eye-con" and is visual, so "earcon" was coined as the auditory equivalent in 1985
- They have no direct relationship to the event/information
  - problem: their meaning need to be learned
- Earcons are composed of "motives"
  - short, rhythmic sequences of pitches and variable intensity, quality, register, and dynamics
- They are used to add context, helping the user maintain awareness

# Earcons: Examples



- Let's fill this together!

# Voice and Speech



- Human voice is an efficient input modality: it allows people to give commands to a computer quickly, on their own terms
  - speech is language dependent, and it may be ambiguous
- Fully understanding natural language remains a dream (for now)
- Voice and speech interaction became mainstream, in recent years
  - thanks to Siri, Google Assistant, Alexa, ...
- Such applications simulate a natural language interaction at different extents
  - they require users to speak a restricted set of spoken commands that users have to learn and remember

# Voice-based Interaction



- From a computer perspective, voice-based interaction is mainly:
  - speech recognition (speech-to-text)
  - speech synthesis (text-to-speech)
- Applications may leverage one or both
  - in some cases, Natural Language Processing (or Understanding, NLU) is added
- Examples:
  - <https://dictation.io/>
  - <https://translate.google.com>

# Voice-based Interaction: Opportunities



- Spoken interaction is successful in some cases...
  - When users have physical impairments (also temporary)
  - When the speaker's hands are busy
  - When mobility is required
  - When the speaker's eyes are occupied
  - When harsh or cramped conditions preclude use of a keyboard
  - When application domain vocabulary and tasks is limited
  - When the user is unable to read or write (e.g., children)

# Voice-based Interaction: Obstacles



- ... and it encounters some issues, as well
  - Interference from noisy environments (and poor-quality microphones)
  - Commands need to be learned and remembered
  - Recognition may be challenged by strong accents or unusual vocabulary
  - Talking is not always acceptable (e.g., in shared office, during meetings)... also for privacy issues
  - Error correction can be time consuming
  - Increased cognitive load compared to typing or pointing
  - Some operations (e.g., math or programming) are difficult without extreme customization
  - Slow pace of speech output when compared to visual displays
  - Ephemeral nature of speech



# Designing Voice-based Interaction



1. Initiation
  - pressing a button, saying a "wake word", ...
2. Knowing what to say
  - learnability is one of the main issues of technologies that mimics natural language
3. Recognition errors (speech-to-text)
  - they will happen... e.g., dime/time
4. Correcting errors
5. Mapping to possible actions
  - mapping the recognized sentence/context to the "right" action is one of most difficult parts
6. Feedback and dialogs
  - to recover from errors, to be sure to start the "right" action, ...

# Screen Readers



- Software application for *handling* vocal synthesis and/or Braille Displays
- Used by people who are blind or with severe visual impairments
- With it, a person can navigate in the operating system
  - by passing from an icon to another, from a window to another
  - and receiving info on the context in which she is, through the Braille Display or a vocal synthesizer
- Built-in in mobile OS and in some desktop OS
  - VoiceOver (Mac and iOS)
  - TalkBack (Android)



# Screen Readers

- Other "famous" screen readers
  - NVDA, <https://www.nvaccess.org/download/>
  - JAWS <https://www.freedomscientific.com/products/software/jaws/>
- Demo Videos
  - Screen Reader Basics (by Google a11y), <https://www.youtube.com/watch?v=5R-6WvAihms>
  - Screen Reader Demo, <https://www.youtube.com/watch?v=dEbl5jvLKGQ>
- **Homework**
  - Try it on a website of your choice
  - A lot of websites are inaccessible to screen readers' users! 😞

# References

- Alan Dix, Janet Finlay, Gregory Abowd, Russell Beale: Human Computer Interaction, 3rd Edition
  - Chapter 10
- Ben Shneiderman, Catherine Plaisant, Maxine S. Cohen, Steven M. Jacobs, and Niklas Elmqvist, Designing the User Interface: Strategies for Effective Human-Computer Interaction
  - Chapter 9: Expressive Human and Command Languages



# License

- These slides are distributed under a Creative Commons license “**Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)**”
- **You are free to:**
  - **Share** — copy and redistribute the material in any medium or format
  - **Adapt** — remix, transform, and build upon the material
  - The licensor cannot revoke these freedoms as long as you follow the license terms.
- **Under the following terms:**
  - **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
  - **NonCommercial** — You may not use the material for [commercial purposes](#).
  - **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the [same license](#) as the original.
  - **No additional restrictions** — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.
- <https://creativecommons.org/licenses/by-nc-sa/4.0/>

