

# Conversational Agents

Human-AI Interaction

Luigi De Russis, Alberto Monge Roffarello

# Background: Voice and Speech

# 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 Conversational Interactions



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, ...

# Conversational Agents

... and their User Interfaces



# Voice User Interfaces

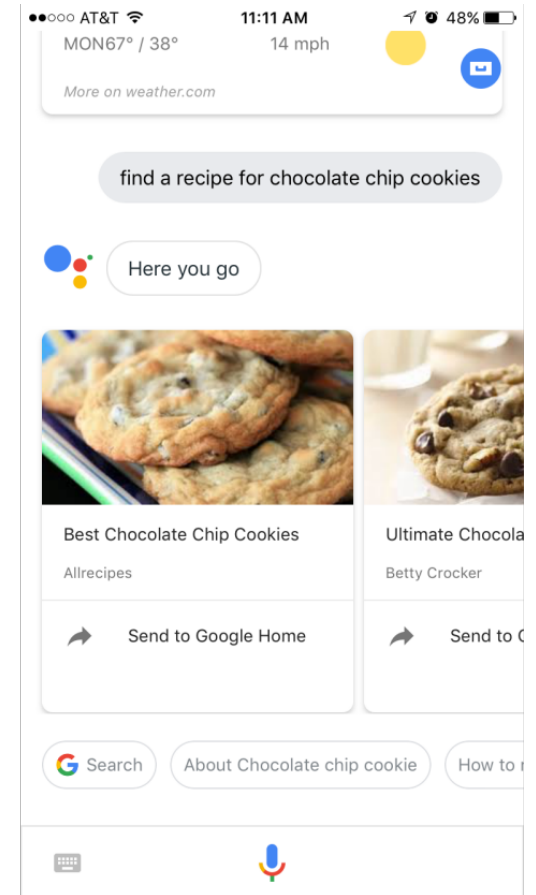
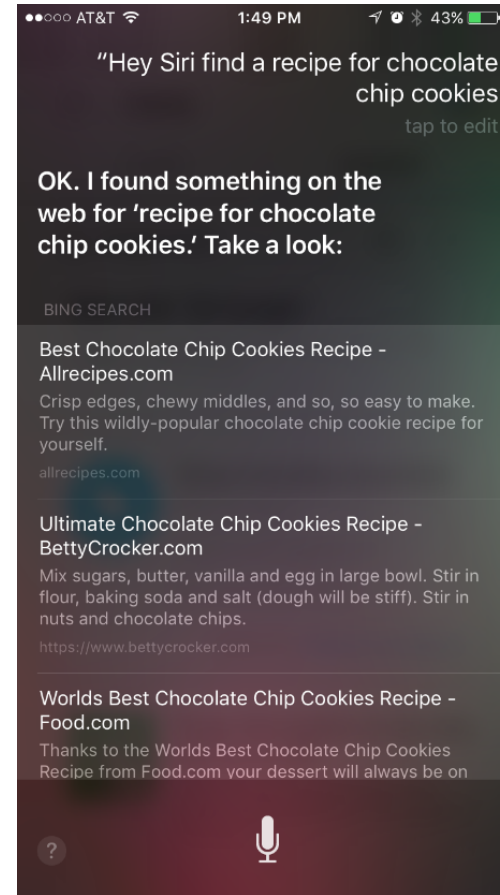
- Voice User Interfaces (VUIs) allow the user to interact with a system through voice or speech commands
  - primary advantage: hands-free, possibly eyes-free interaction
- Voice User Interfaces or Conversational User Interfaces?
  - *"which mimics a conversation with humans"*
  - "conversational" applies to both text-based chatbots and VUIs
- Contemporary VUIs can be divided in:
  - screen-first systems
  - voice-only systems
  - voice-first systems

# Screen-First Devices

- Most of contemporary voice interaction happens on screen-first devices
  - smartphones, mainly
- Impressive speech recognition and language processing features
  - but overall experience is fragmented
- Main limitations
  - missing functionality
  - poor use of screen space while speaking
  - missing affordances

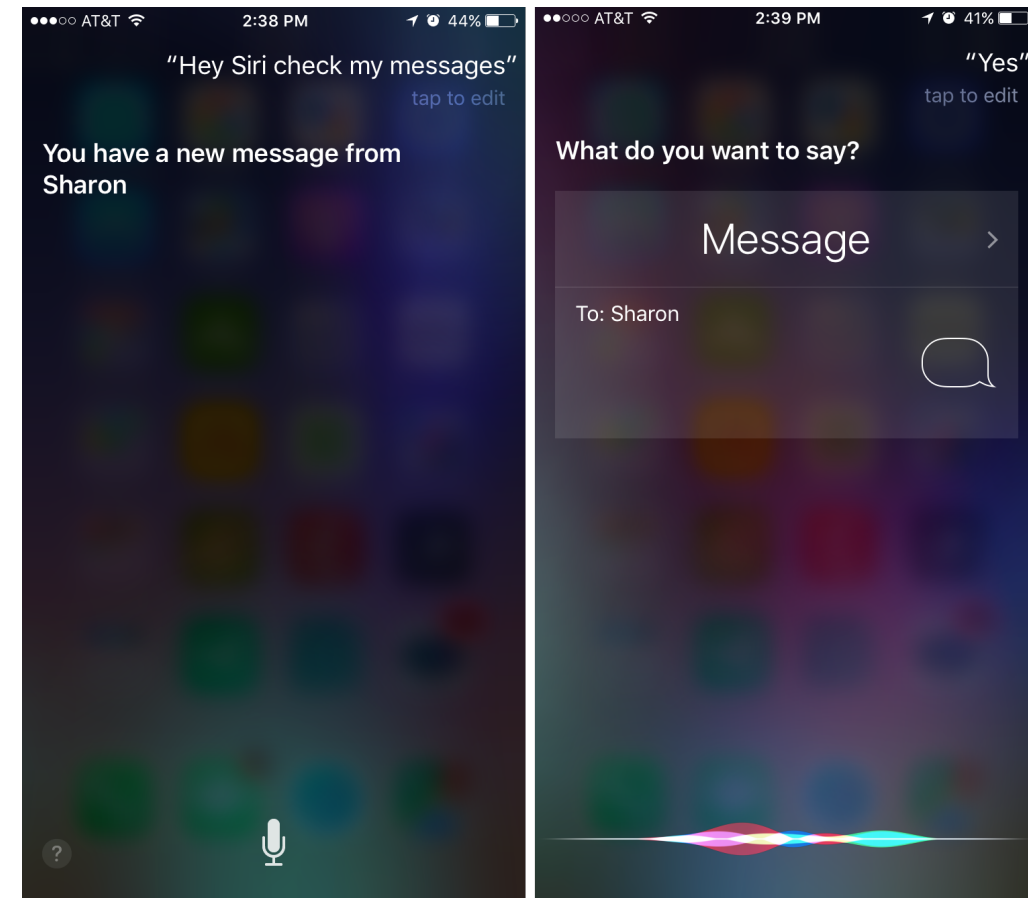
# Missing Functionality and Affordances

- Users can start a task via voice, but subsequent steps require them to use the touchscreen
- Visual affordances are missing (or poor)
  - Siri omits several visual affordances (e.g., it does not show that people can edit a text message before sending it)
  - Google Assistant is better in this



# Poor Screen Space Use

- Tasks with some support for multi-step voice input exhibit a screen design:
  - totally different from the "normal" GUI version
  - which limits the information available to the user



# Voice-Only Devices

- No visual display at all
  - like the Amazon Echo
  - audio is for input **and** output (plus some "feedback lights")
  - hands-free operation
- Quite good accuracy in speech recognition
  - if you do not mix different languages in a sentence
  - auditory signals are the only used cues (no visual affordances)

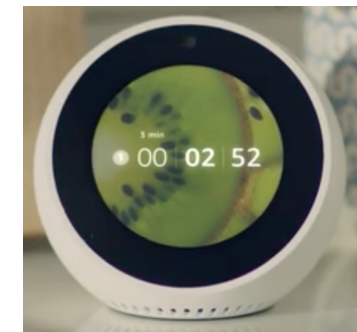


# Voice-Only Devices: Limitations

- They are quite prolix in the answers
- You have to know what to say!
- Some operations are "challenging", e.g.,
  - once a timer is set up, the user can only *ask* how much time is left
  - getting a weekly weather forecast is a... memory test
- Some actions are not allowed nor expected, e.g.,
  - you cannot insert your wifi password, vocally
  - you cannot hear about all the available (and installable) skills

# Voice-First Devices

- Voice-only devices... with a screen
- A system which primarily accept user input via voice commands, and **may** augment audio output with visual information
  - no differences from the "voice" perspective
  - GUI is less capable than the one in screen-first devices
- Typically, the display is a touch screen
  - but it rarely provides buttons or menus
  - the focus is still on voice



# Designing Conversational Agents

... and their UI



# Designing Conversational UI

- Voice interaction between people and devices is analogous to learning a foreign languages
  - both for users and designers/developers
- Easily learnt through **immersion**
  - voice-first devices have an advantage in this
- Successful examples on voice-first devices:
  - sequential numbering of search results
  - randomly show new speech commands
  - voice-accessible interactive (visual) content
- Beware: people often have unrealistic expectations
  - they think a VUI as a "natural conversation partner"



# Designing Conversational UI

- To design a VUI, you firstly need to have a clear picture of
  - who is communicating, i.e., who are your users
  - what they are communicating about, what they will ask about, i.e., what their needs are
- Then, you can write some **sample dialogs** and sketch a **diagram of the conversation flow**
  - both convey the flow that the user will actually experience
  - you can also informally experiment with and evaluate different strategies
    - e.g., is it better to confirm a user's request with an implicit confirmation or an explicit one?
- Focus on the **spoken conversation** before considering any visual element
  - imagine to work with a voice-only device

# Basic Conversational Frames

- **Controlling:** specifying a goal with means of achieving it
  - "Play Radio DeeJay from TuneIn"
- **Delegating:** asking for an outcome without specifying how to achieve it
  - "Play some jazz music"
- **Guiding:** discussing the means of achieving a goal
  - "I want to hear some music, how should I do it?"
- **Collaborating:** mutually deciding on goals between both participants
  - "What should we do?"

Currently adopted by contemporary VUIs

# Guidelines

- By Microsoft Research
  - <https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/>
- Saleema Amershi et al. Guidelines for Human-AI Interaction. ACM CHI 2019
  - <https://doi.org/10.1145/3290605.3300233>

## Guidelines for Human-AI Interaction

The Guidelines for Human-AI Interaction will help you create AI systems and features that are human-centered. We hope you use them throughout your design process – as you evaluate existing ideas, brainstorm new ones, and collaborate with the multiple perspectives involved in creating AI.

These guidelines synthesize more than 20 years of thinking and research in human-AI interaction. Learn more: <https://aka.ms/aiguidelines>.

**INITIALLY**

- 1 **INITIALLY**: Make clear what the system can do. Help the user understand what the AI system is capable of doing.
- 2 **INITIALLY**: Make clear how well the system can do what it can do. Help the user understand how often the AI system may make mistakes.

**DURING INTERACTION**

- 3 **DURING INTERACTION**: Time services based on context. Time when to act or interrupt based on the user's current task and environment.
- 4 **DURING INTERACTION**: Show contextually relevant information. Display information relevant to the user's current task and environment.
- 5 **DURING INTERACTION**: Match relevant social norms. Ensure the experience is delivered in a way that users would expect, given their social and cultural context.
- 6 **DURING INTERACTION**: Mitigate social biases. Ensure the AI system's language and behaviors do not reinforce undesirable and unfair stereotypes and biases.

**WHEN WRONG**

- 7 **WHEN WRONG**: Support efficient invocation. Make it easy to track or request the AI system's services when needed.
- 8 **WHEN WRONG**: Support efficient dismissal. Make it easy to dismiss or ignore unhelpful system services.
- 9 **WHEN WRONG**: Support efficient correction. Make it easy to self-correct, or request when the AI system is wrong.
- 10 **WHEN WRONG**: Scope services when in doubt. Engage in investigation or gracefully degrade the AI system's services when uncertain about a user's goals.
- 11 **WHEN WRONG**: Make clear why the system did what it did. Enable the user to access an explanation of why the AI system behaved as it did.

**OVER TIME**

- 12 **OVER TIME**: Remember recent interactions. Maintain short-term memory and allow the user to make efficient references to that memory.
- 13 **OVER TIME**: Learn from user behavior. Personalize the user's experience by learning from their actions over time.
- 14 **OVER TIME**: Update and adapt cautiously. Limit disruption changes when updating and adapting the AI system's behaviors.
- 15 **OVER TIME**: Encourage granular feedback. Enable the user to provide feedback indicating their preferences during regular interaction with the AI system.
- 16 **OVER TIME**: Convey the consequences of user actions. Immediately update or convey how user actions will impact future behaviors of the AI system.
- 17 **OVER TIME**: Provide global controls. Allow the user to globally customize what the AI system monitors and how it behaves.
- 18 **OVER TIME**: Notify users about changes. Inform the user when the AI system adds or updates its capabilities.

Microsoft

# A Very Simple Example

Weather Web App: let's "chat" about the weather

# Conversational Platforms

- Natural language understanding platforms
  - for developers, mainly
  - typically cloud-based
- To design and integrate voice user interfaces into mobile apps, web applications, devices, ...
- Focus on simplicity and abstraction
  - no knowledge of NLP required

# Conversational Platforms

- Two main families:
  1. Extension of a product
    - they need an existing product (software and/or hardware) to work
    - e.g., Actions on Google or Skills for Amazon Echo
  2. Standalone services
    - a series of facilities to create a wide range of conversational interfaces in one platform, *typically* integrated in "suites" of cloud services
    - e.g., Dialogflow, IBM Watson, wit.ai, ...

# Snips



- "Create a Private by Design voice assistant that runs on the edge"
  - <https://snips.ai>
- France-based startup, founded in 2013, acquired by Sonos in 2019
- Run on the edge, not in the cloud
  - Raspbian, Android, iOS, macOS, and most Linux flavors
  - the setup of the NLP component is online
- Free for makers and for building prototypes
- 6 fully supported languages, mostly uses Node.js



# DialogFlow



- "Build natural and rich conversational experiences"
  - <https://dialogflow.com>
- California-based startup, founded in 2010, acquired by Google in 2016
  - previously known as api.ai
- Free to use for simple usage
- One-click integration with several services
  - Telegram, Facebook Messenger, Cortana, Google Assistant, ...
- Multiple languages support
  - English, Dutch, Italian, Chinese, ...
- REST API and various (official) SDKs
  - Java, C#, Python, PHP, Go, and Node.js

# DialogFlow: Definitions

- Each application (an agent) will have different **entities** and **intents**
- Intent
  - a mapping between what a user says and what action should be taken by the agent
- Typically, an intent is composed by:
  - What a user says
  - An action
  - A response
- Different out-of-the-box intents can be enabled on DialogFlow

# DialogFlow: Definitions

- Entities
  - represent *concepts*
  - serve for extracting parameter values from natural language inputs
  - should be created only for concepts that require actionable data
- Many pre-existing entities are available on the platform

# Weather App Prototype

- Base implementation:
  - <https://github.com/luigidr/dialogflow-weather>
- HTML+CSS+JS and Python
- Uses the Dialogflow v2 library

# References and More Information (in English)

- *Multimodal Interaction* – slides and video lectures:
  - <https://elite.polito.it/files/courses/02JSKOV/2021/slide/10-multimodal.pdf>
  - <https://www.youtube.com/watch?v=AfVJiE1weGU>
  - <https://www.youtube.com/watch?v=wFP8g1AqDIQ>
- *Voice User Interfaces* – slides and video lecture:
  - <https://elite.polito.it/files/courses/02JSKOV/2019/slide/10-vui.pdf>
  - <https://youtu.be/bibKxK2Ok2U>

# References and More Information (in English)

- *Voice User Interfaces on the Web* – slides and video lectures:
  - <https://elite.polito.it/files/courses/02JSKOV/2019/slide/11-vui-web.pdf>
  - <https://youtu.be/RiGeYFzZxuE>
  - <https://youtu.be/mHWt63jH-ml>
  - <https://youtu.be/YilcJhpQJFk>
  - <https://youtu.be/VU5z-ALZJvo>

# 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/>

