Speech Input and Output

https://www.twitch.tv/seebotschat

This week

- Introduction to Voice UIs (VUIs)
- Thinking about the design of VUIs
- Interesting examples
- Prototyping VUIs with DialogFlow
  - Leading in to Project 2

Housekeeping

- Extra time for Project 1
- Demo code for Project 1
- Two extra credit assignments posted

Understanding speech input

Why speech?

- Shifting gears a little bit from focusing on specific population groups, to a whole class of interaction
- Why?
  - Speech is a broad-ranging topic
  - Speech just works differently than other technologies

Big ideas

- Consuming speech is very different than reading on screen
- Navigating speech interfaces is very different than navigating GUIs
- Development tools are quite different too

Drawing from our everyday experiences

Let’s talk about voice interaction

1a. When has it worked well for you?
2a. When has it broken down?

1b. Can we figure out **what kinds of tasks are well-suited** for voice?
1c. **What kinds of tasks are NOT** well-suited for voice?
2b. Can we come up with a list of the **types of errors** that occur?

(and write them down)

Discussion

- menti.com, xx yy zz

When it works well (and why)

More benefits of speech

- Support completion of tasks using natural language, without knowing commands
- Commands can take many forms
  (“play music”, “play Kanye West”, “play track 3 from *Graduation*”)
- Accessible to users of different abilities
- Can interact in the background or at a distance

When it does not (and why)

More challenges of speech

- Accounting for many different entry paths
- Partial or incorrectly formed requests
- Noise and recognition errors
- Navigating output can be tricky

When to include voice

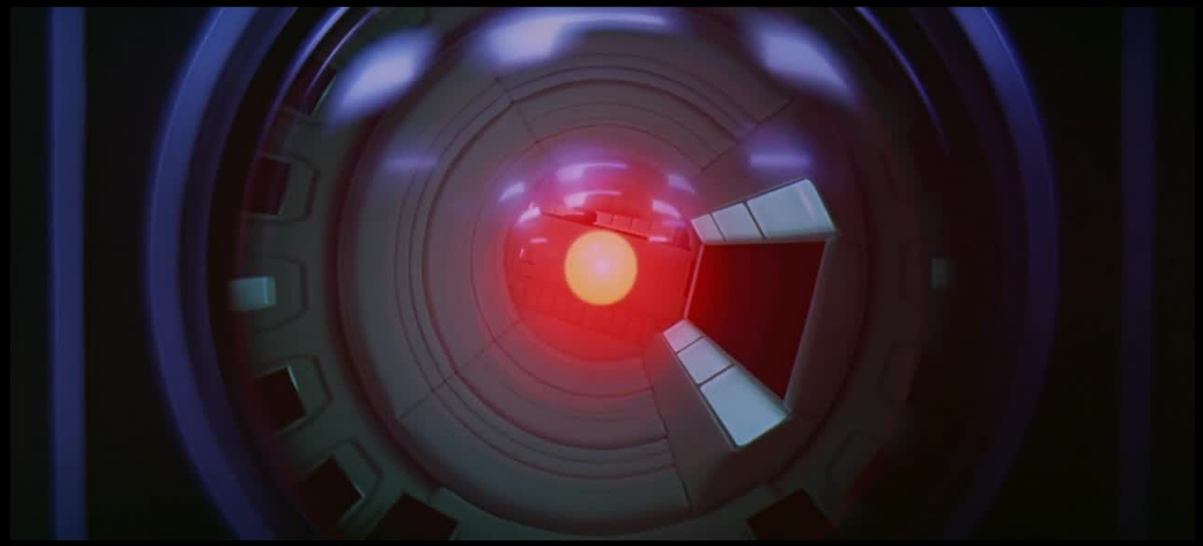
- There are certain contexts in which voice is the only (or obvious) solution
- But there are also certain kinds of tasks where voice is ideal (and others where it is not)

History of voice interaction

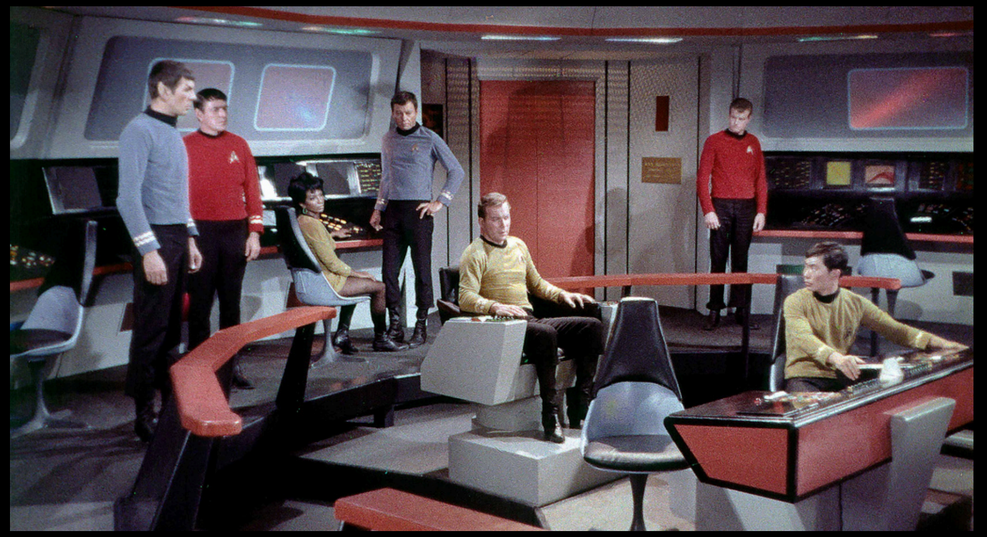
Examples from fiction

- This is an instance where we have had many fictional systems before real-world systems

Voice interfaces in fiction



HAL 9000, *2001: A Space Odyssey* (1968)



Star Trek (1966)

User expectations?

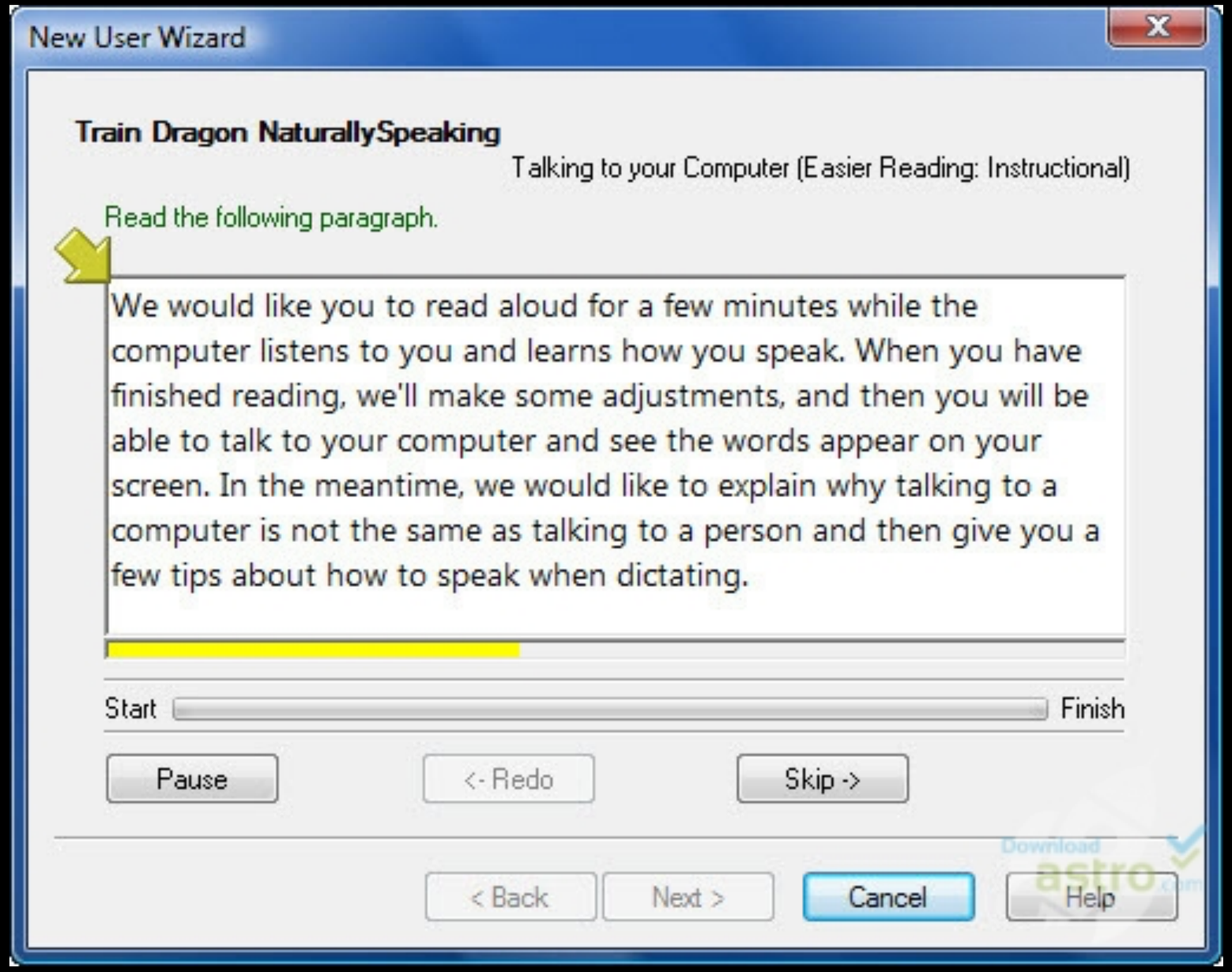
Some user expectations

- Can speak in a natural voice, at normal rate
- Can phrase commands in various ways and via natural language
- Can assume context from questions
- If information is missing, can ask follow-up
- Response is instant (or at least very fast)

A brief history of (real) voice UIs

- 1970s/80s - Interactive Voice Response systems (“press 3 for voicemail…”)
- 1990s/2000s - first commercial speech recognition systems (e.g., Dragon NaturallySpeaking)
  - Require significant per-user training
- 2010s - Siri and voice agents, **speaker-independent** speech recognition

Training speaker-dependent voice recognition

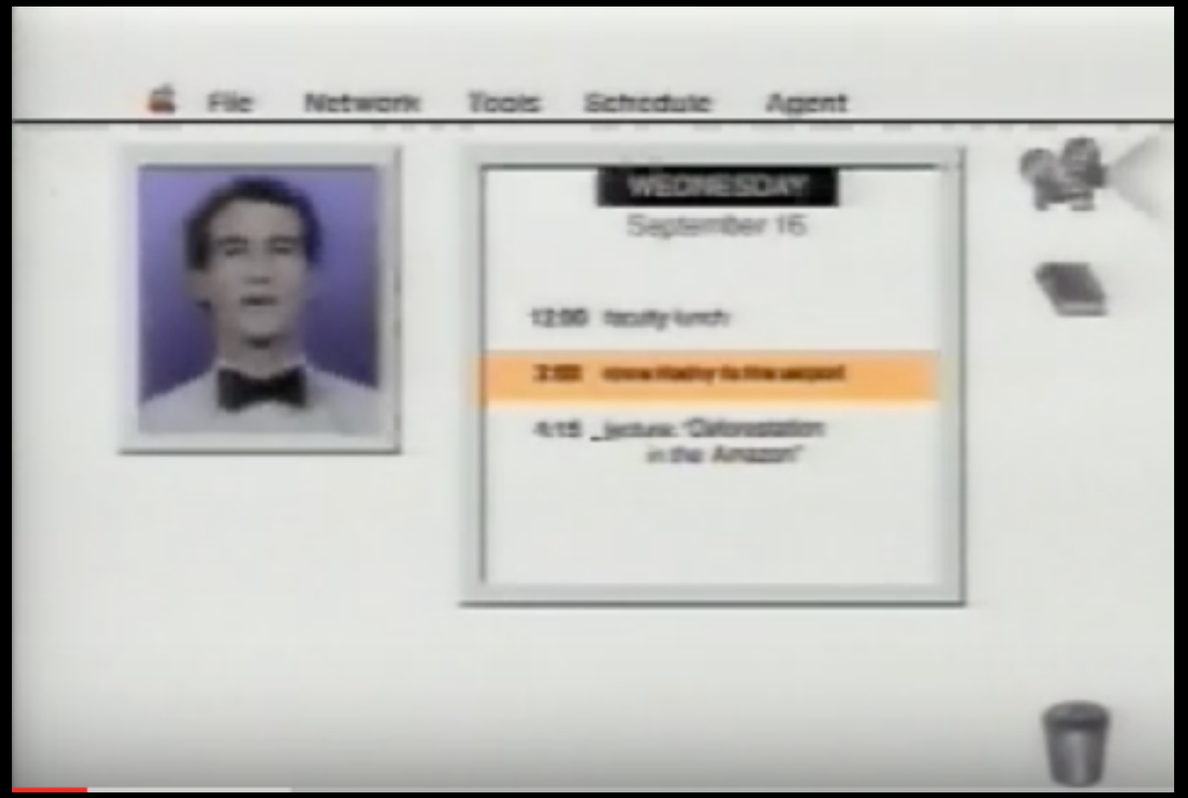


Research prototypes

- “Put that There”
- Knowledge Navigator

Put That There (1979)

Knowledge Navigator (1987)



What features did you notice?

Features in KN

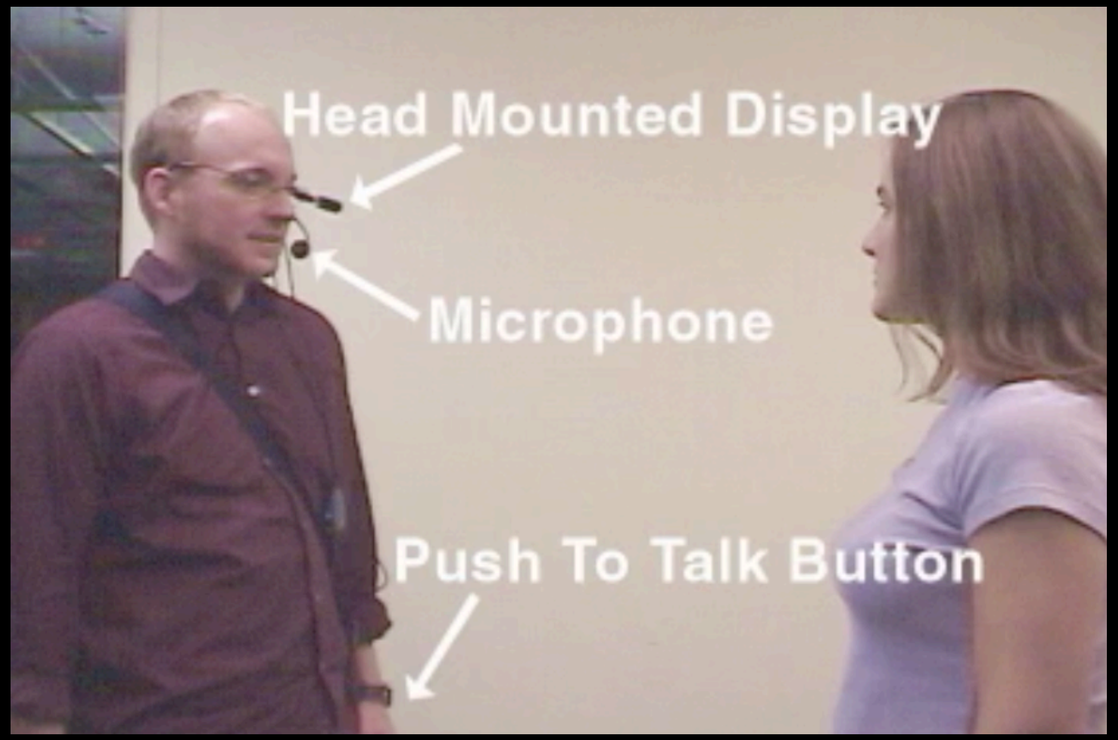
- Context (space, conversation)
- Personal assistant
- Human conversation (no delimiter)
- Summary questions
- Local variables (Jill)
- Errors, partial knowledge

Considering use context

- We should not just accept any speech recognizer, but instead tailor it to the context
- Customizing for the context
  - Kitchen vs. operating room?
  - Home vs. mobile/wearable?

Dual purpose speech

- Taking advantage of mobile context & display



Designing voice UIs

Some notes

- A good way to learn this is to look at translating from a GUI to a Voice UI
  - What do we get “for free” visually?
  - How can we simplify interactions?
  - How to create a fundamental set of actions?
- Dev tools still in an early stage
  - Opportunity to do better

Why is speech challenging?

adapted from Schnelle and Lyardet

- Speech is **one-dimensional**
- Speech is **transient**
- Speech is **invisible**
- Speech is **asymmetric**
- Flexibility vs. accuracy

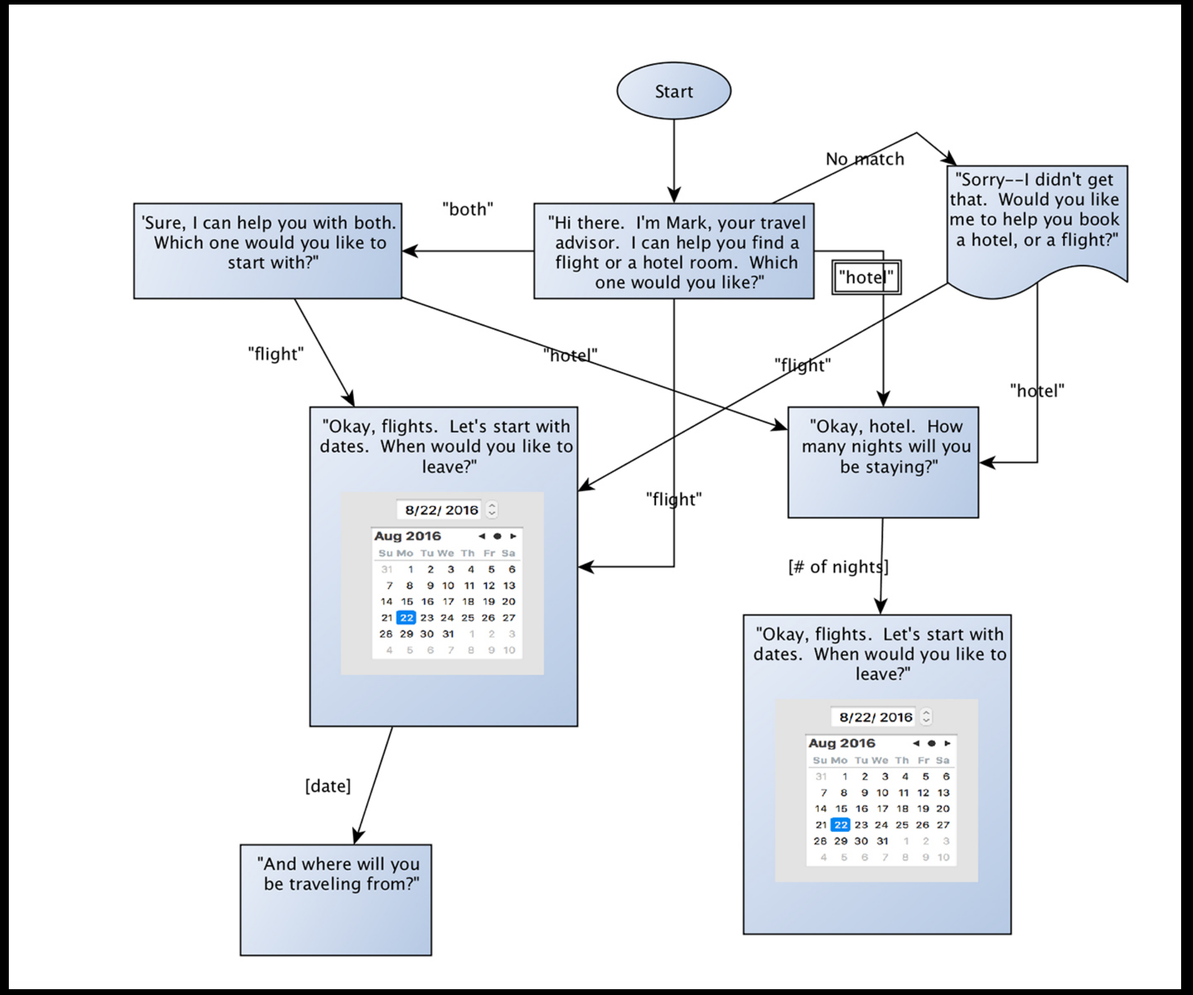
Why is speech challenging?

adapted from Schnelle and Lyardet

- **One-dimensional.** Not easy to skim speech, have to listen through it
- **Transient.** Spoken information requires short- or long-term memory
- **Invisible.** Possible actions are not clear
- **Asymmetric.** Faster to speak than type; slower to listen than read
- **Flexibility vs. accuracy.** Classic HCI tradeoff!

How to design voice UIs

- Most current systems involve some sort of **pattern matching** and **entity extraction**
- Pattern: “turn on the <name_of_light>”
- Dialog tree: model possible inputs and conversation paths



Wizard of Oz prototyping

- Person acts out what the computer would do
- Can use this as a prototyping tool, **but we have to follow the rules we ourselves set**
- Is our interaction logic robust enough for the real world?

Example WOz fail

- “What time is the **<meeting_name>**?”
- What inputs will break this?

Let’s try it

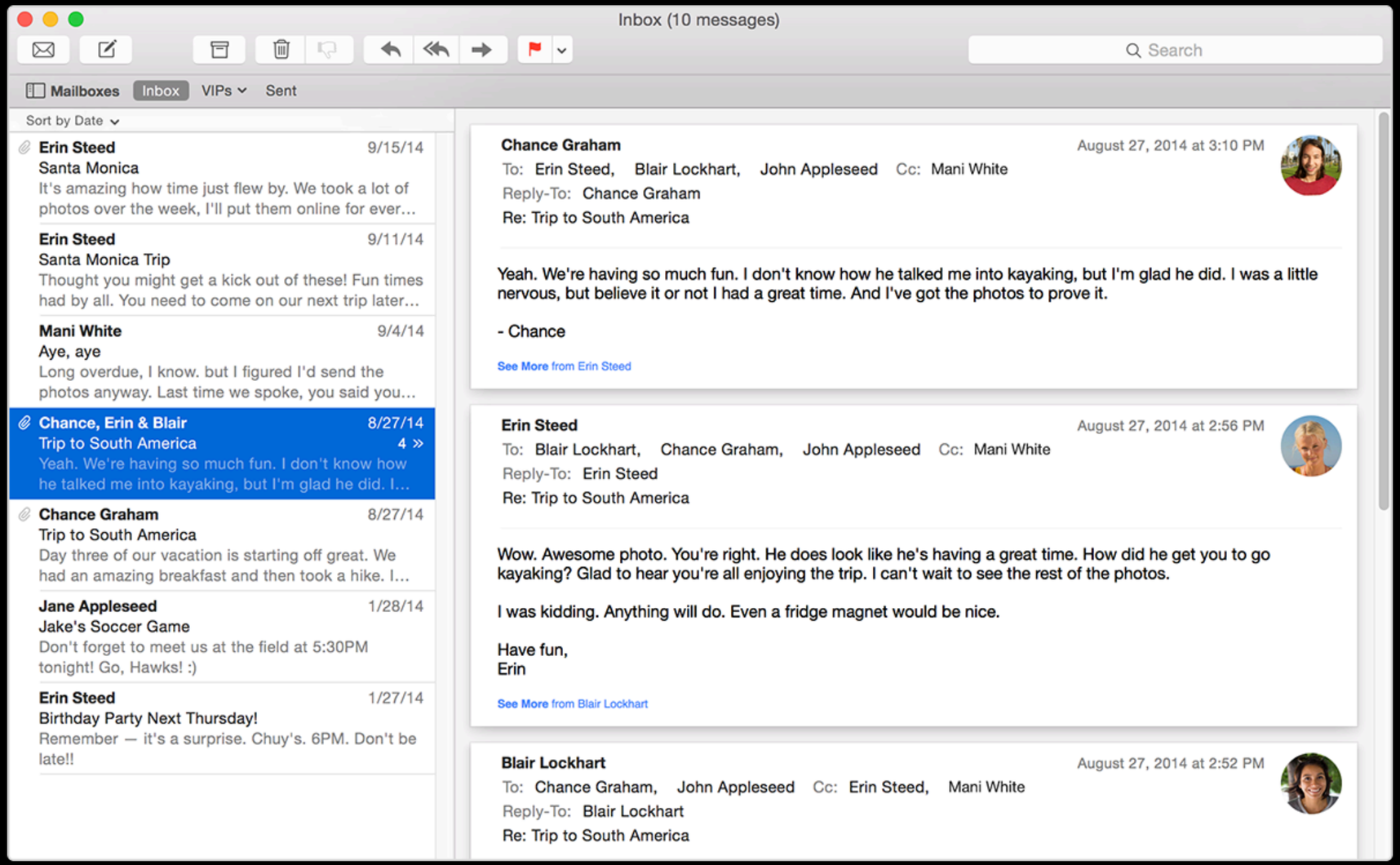
- Let’s model the dialog tree for a smart thermostat
- Assume we have working speech recognition, entity extraction
- Spend 5 minutes designing your voice UI, 5 minutes testing

Where did we go wrong?

Designing “natural” voice UIs

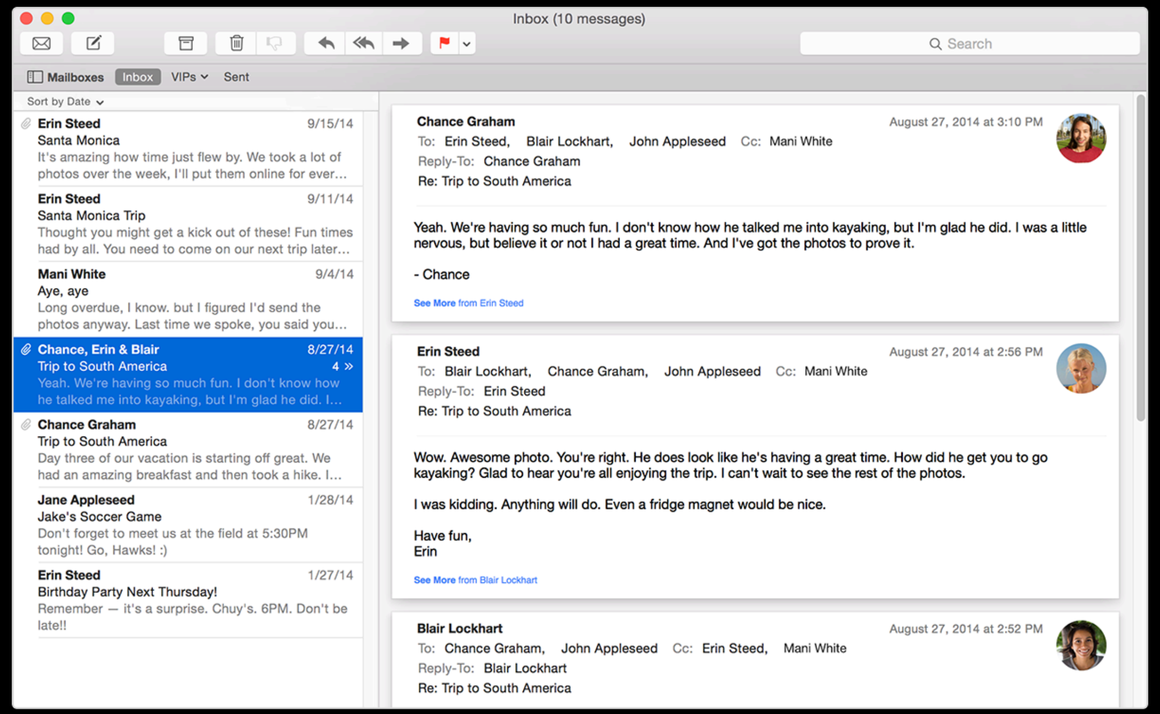
1. Identify use cases, *intents*
2. Identify use cases that we get “for free” with a graphical user interface, and forgot about
3. Work out the details of each intent (including edge cases)
4. Design output (verbosity, controls to navigate output)
5. Test and iterate

Example: an email voice client



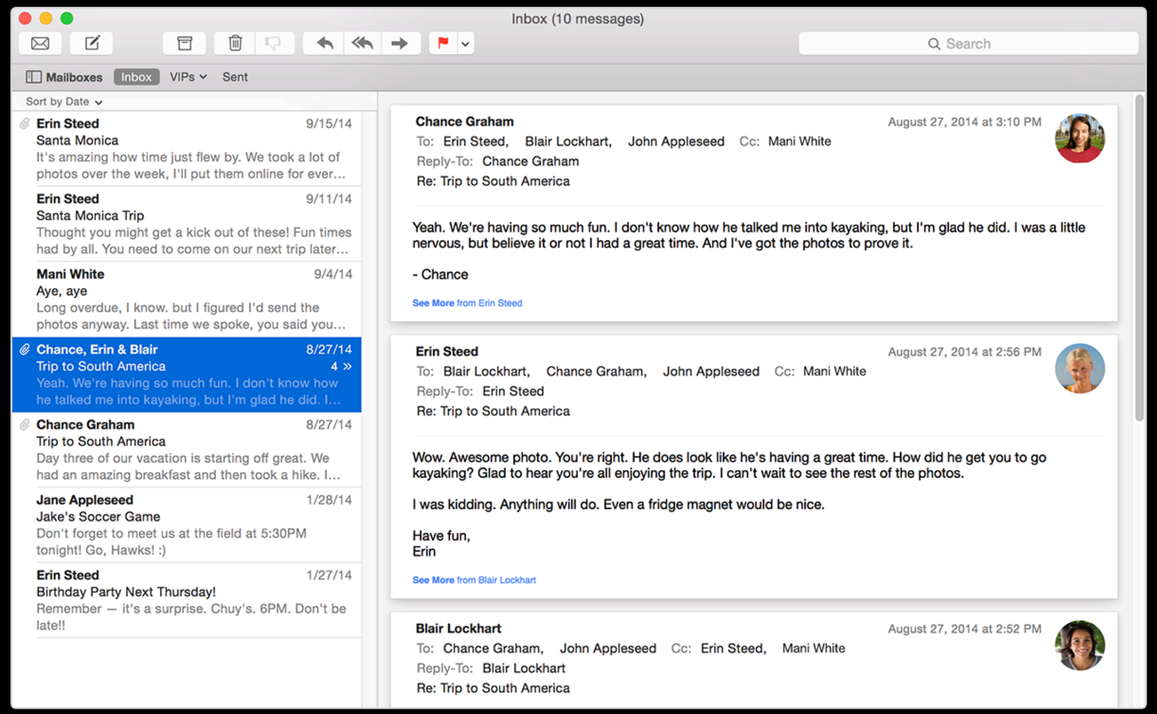
1. Identify intents

- Open/read
- Compose
- Navigate
- Delete
- Reply



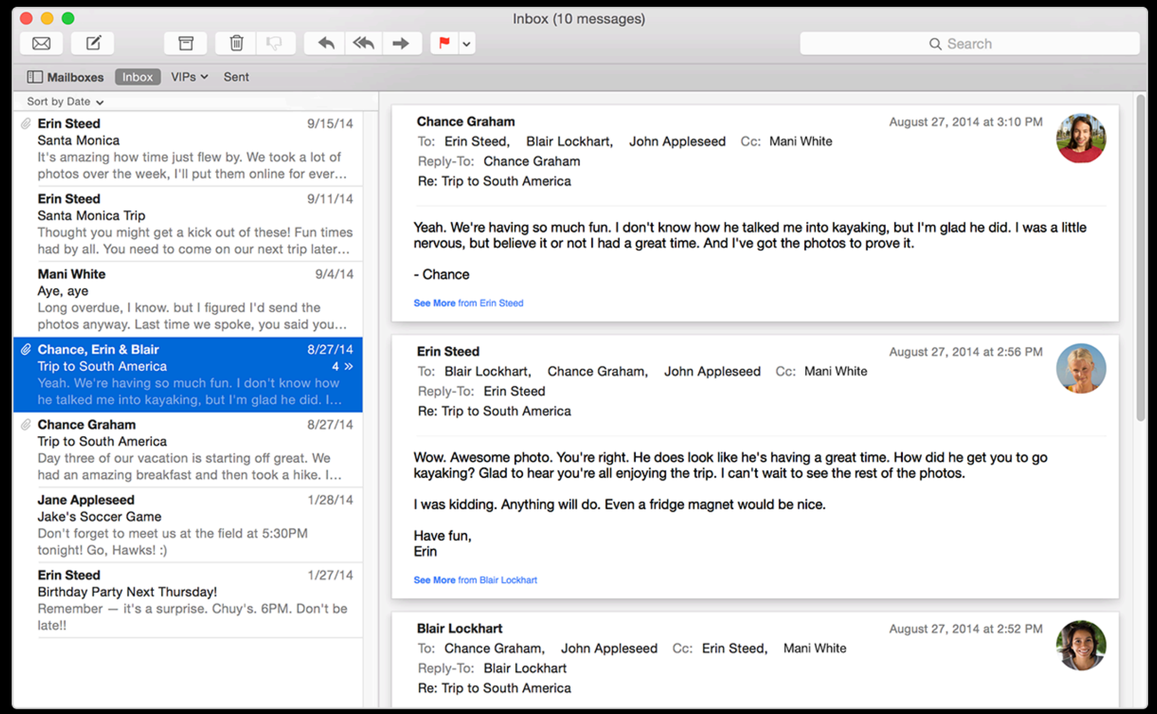
2. Identify visual tasks

- “Do I have an email from grandma?”
- “Delete all emails from grandma”



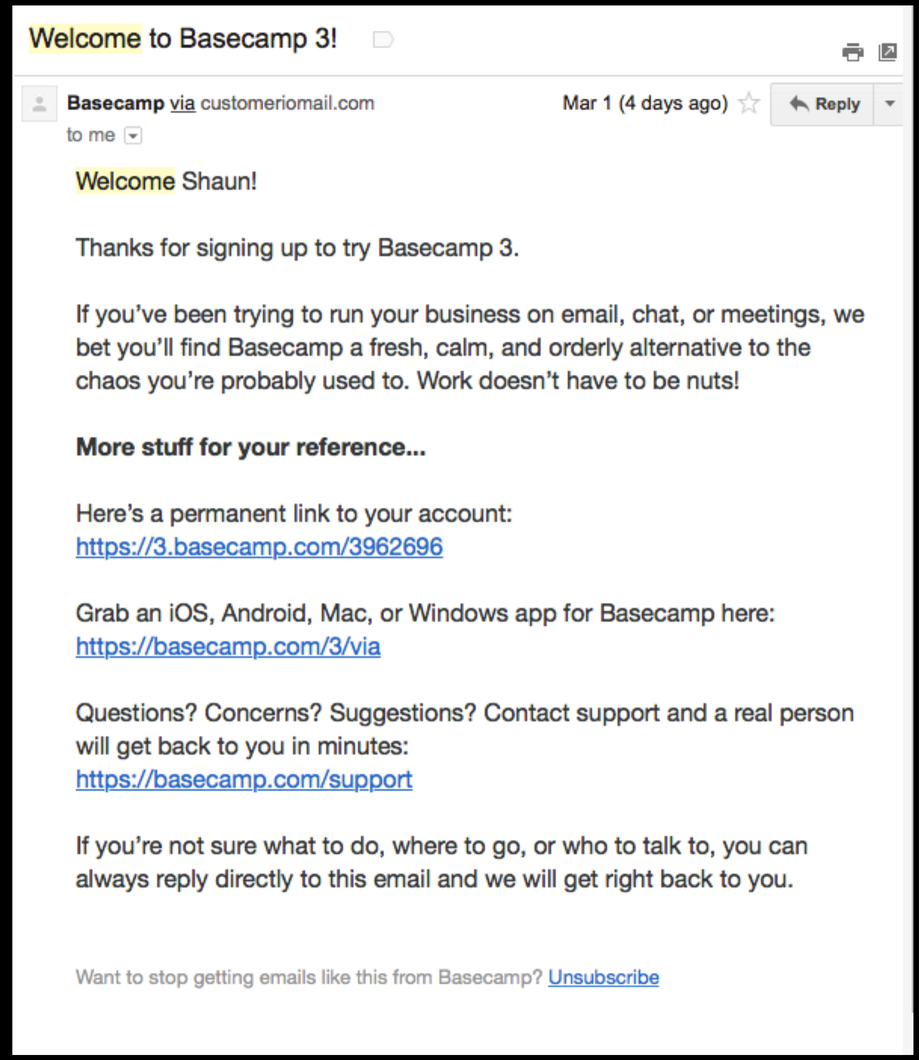
3. Work out logic for each intent

- For now, let’s design “compose”



4. Design output

- What to read?
- What to ignore?
- What navigation to provide?



Welcome to Basecamp 3!

Basecamp via customeriomail.com
to me

Mar 1 (4 days ago) Reply

Welcome Shaun!

Thanks for signing up to try Basecamp 3.

If you've been trying to run your business on email, chat, or meetings, we bet you'll find Basecamp a fresh, calm, and orderly alternative to the chaos you're probably used to. Work doesn't have to be nuts!

More stuff for your reference...

Here's a permanent link to your account:
https://3.basecamp.com/3962696

Grab an iOS, Android, Mac, or Windows app for Basecamp here:
https://basecamp.com/3/via

Questions? Concerns? Suggestions? Contact support and a real person will get back to you in minutes:
https://basecamp.com/support

If you're not sure what to do, where to go, or who to talk to, you can always reply directly to this email and we will get right back to you.

Want to stop getting emails like this from Basecamp? Unsubscribe

What we didn’t cover

- Advanced composing options (reply all, forward)
- Address book lookup
- Formatting
- Organizing messages into tags and folders
- …

Design guidelines

- Questions may be formulated in a variety of ways
  - Can use dialogs to fill in missing variables
- User must remember what she just heard (short-term), must remember what the system can do (long-term)
- Provide clear feedback:
  - System is listening, system has heard your command, system has performed an action
- Balance verbosity with clarity about what the system is expecting
- Provide help and guidance

It can get complicated…

- Using pronouns (and keeping them across queries)
  - “What is the capitol of Germany?”
  - “How many people live there?”
- Bilingual queries
  - “Siri, play Los Lobos”
- Domain-specific language changes meaning
  - “Play *Nevermind*”

Some unsolved problems

- Expressivity of synthesized speech
- Identifying speakers (although this is improving)
- Understanding user's context
- User privacy

Wednesday

Today

- Accessibility of voice interfaces
- Intro to DialogFlow
- Begin Project 2

Voice agents vs. screen readers

- How are voice agents different than screen readers?

Voice agents vs. screen readers

	Voice Agents	Screen readers
Users	Untrained	Trained
Input	Speech	Keyboard
Visual UI as fallback?	Maybe	Probably not