Future of Voice Agents

Christopher Hong

April 9, 2025

Observation, Values and Mission

- The advent of large language models have enabled a new bridge between everyday people and computing
- AI should allow everyday people to harness the power of computation
- People should be in control of their computing, not vice versa

Big Tech Voice Assistants (Alexa, Gemini, Siri)...

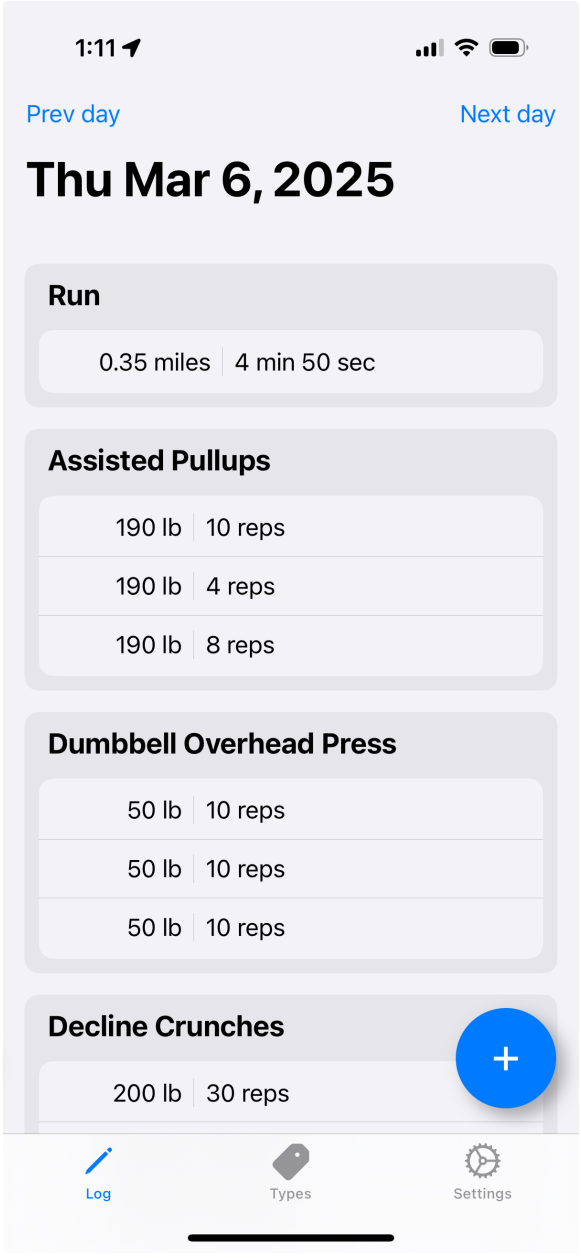
- ✓ Do well at some basic pre-programmed tasks
- ✗ Don't grasp the context of voice commands
- ✗ Don't integrate well or at all with third-party applications
- ✗ Fail to transcribe words accurately that humans have no problem given a situational context (ex. "hot words" in speech recognition research)
- ✗ Don't compose appropriate action for more complex user requests

Solution

- Bridge user-intent to application-specific code using (large) language models
- Imbue automatic speech recognition with contextual capability for short-form transcription
- Give application developers the tools to contextualize voice assistance

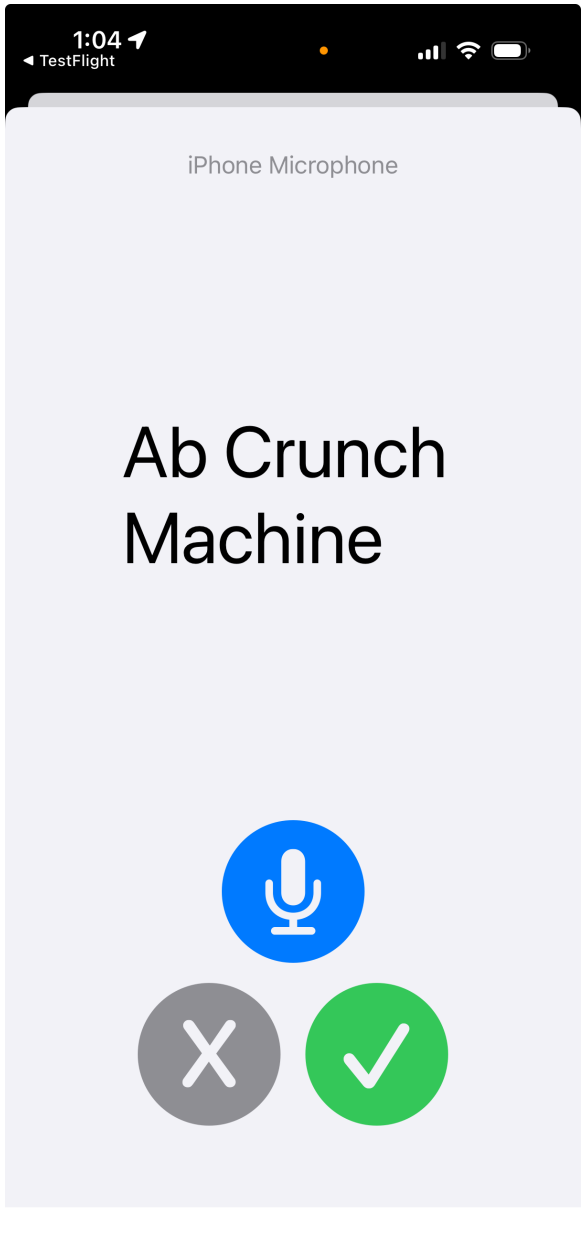
Prototype case: Gym Working Logging App

Suppose, in a workout logging app, the user wants to switch to recording a different type (eg. "Ab Crunch Machine"). How can the user invoke a voice assistant to specify it quickly?



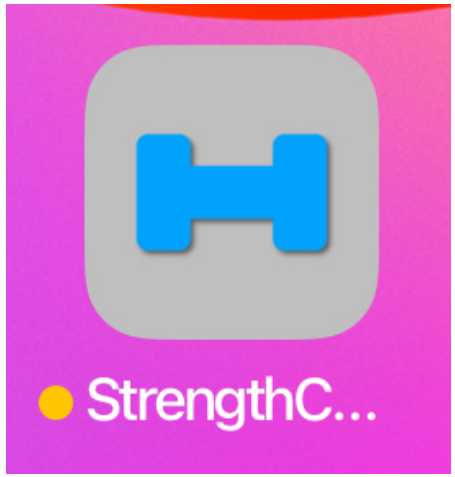
Prototype case: Gym Working Logging App

Instead of using a platform or SaaS API from Apple, Google, etc., use a tailored automatic speech recognition pipeline to control quality and viability.



Prototype case: Gym Working Logging App

The iOS app is a polished minimum-viable product that will be launched soon from the time of creating this presentation.



Prototype case: Gym Working Logging App

Research results below compares "Custom" to SaaS APIs demonstrate the need for better contextualizing tools

The following table shows the accuracy and precision results for the baselines/models evaluated on the datasets. Note that for each baseline the first row and second row indicate the non-streaming and streaming APIs.

	Google Speech Commands	Workout Types Self Recorded	Workout Types Synthesized
Apple SFSpeech	.946	.544	.857
" (streaming)	.709	.500	.719
Google TTS	.773	.632	.941
Custom	.919	.860	1.0

Table 1: Accuracy

Prototype case: Gym Working Logging App

For a whitepaper that delves further in the technical aspects and justification, visit

https://cs.brown.edu/people/ycheng79/csci1952qs23/Top_Project_3_Christopher%20Hong_Low%20Latency%20Streaming%20Speech%20Selection.pdf

Other Potential Markets

- A Desktop app for personal trainers to quickly organize routines for their clients
- Voice control for people with upper-body physical impairments
- Field operators (in various industries) who need quick access to proprietary information

A Disruptive, General Purpose Technology

- Context aware short-form speech recognition is potentially a *disruptive* technology, in the sense used in The Innovator's Dilemma
- Big Tech's strategy with voice assistants is to be useful to a wide swath of customers and gravitate them towards their own product ecosystem
- However, this leads voice assistants to be useful to nobody
- The right approach is to meet the needs of niche applications first
- In the end, the invested technology and capabilities can grow into mass-market general purpose technology of connecting voice to code