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

✗ Fail to grasp the context of voice commands

✗ Do not integrate well or at all with third-party applications

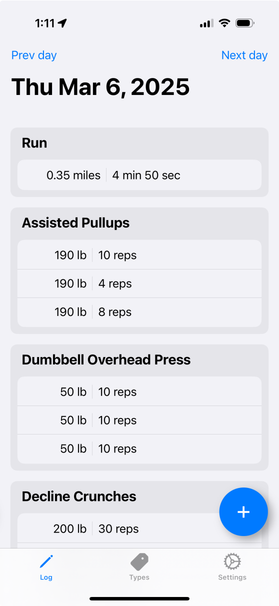
✗ Fail to transcribe words accurately that humans have no problem given a situational context (called ”hot words” in speech recognition research)

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?



Thu Mar 6, 2025

Run

0.35 miles | 4 min 50 sec

Assisted Pullups

190 lb | 10 reps

190 lb | 4 reps

190 lb | 8 reps

Dumbbell Overhead Press

50 lb | 10 reps

50 lb | 10 reps

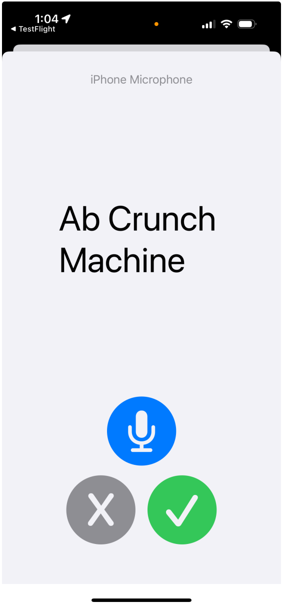
50 lb | 10 reps

Decline Crunches

200 lb | 30 reps

Prototype case: Gym Working Logging App

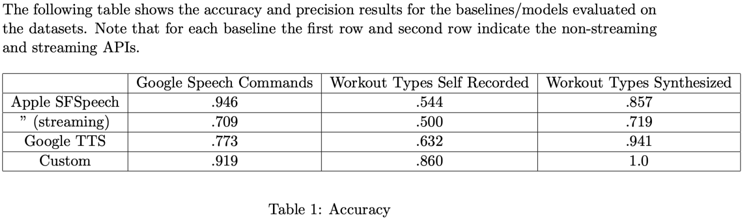
Instead of using a platform or SaaS API from Apple, Google, etc., use a tailored automatic speech recognition pipeline to control quality and viability.



Ab Crunch Machine

Prototype case: Gym Working Logging App

The following research results demonstrate the need for better contextualizing tools for speech recognition.



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