Describing Videos for Accessibility
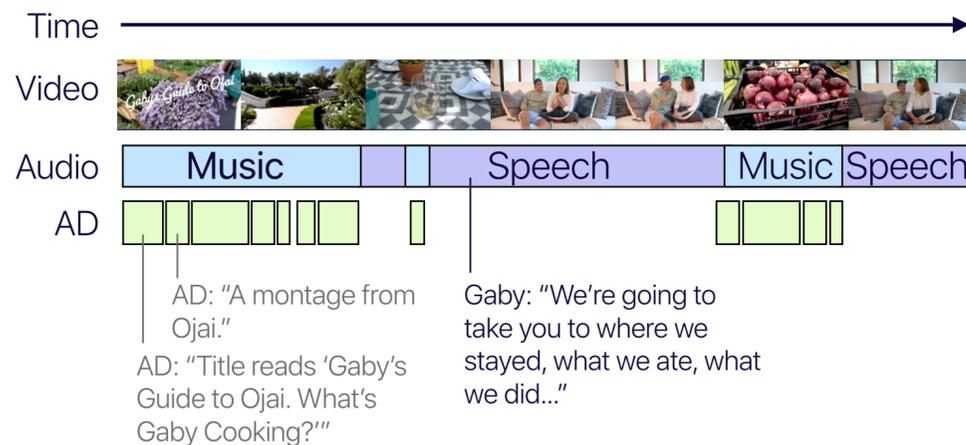
Amy Pavel | *Apple, CMU*

People increasingly use video rather than text to communicate information. Many videos are **inaccessible** for blind and visually impaired audiences.

Video — Visual — Audio
inaccessible information — accessible information

Audio Descriptions

A narration track that describes important visual content, and avoids overlap with important audio content.

Time
Video
Audio — Music — Speech — Music — Speech
AD

AD: "A montage from Ojai."
AD: "Title reads 'Gaby's Guide to Ojai. What's Gaby Cooking?'"
Gaby: "We're going to take you to where we stayed, what we ate, what we did…"

Interviews with AD Professionals "*Time is everything*"
They must iteratively adjust their description content and timing to fit within gaps — a tedious task and time consuming task that takes ~10x video length.

Audio Descriptions with Rescribe

Authors describe the visual content and record their descriptions. We optimize the descriptions to fit by **jointly editing text and underlying audio.**

Extended Descriptions

0:02 A montage of bright footage from Ojai.

0:03 Shots of lavender in a farmer's market.

0:05 Red flowers against a white house and blue sky.

Insert pauses

Inline Descriptions

0:02 A montage from Ojai.

0:04 Shots in a farmers market.

0:06 Red flowers.

Edit text, Adjust timing

Extended-Inline Descriptions (new)

0:02 A montage from Ojai.

0:04 Shots of lavender in a farmers market.

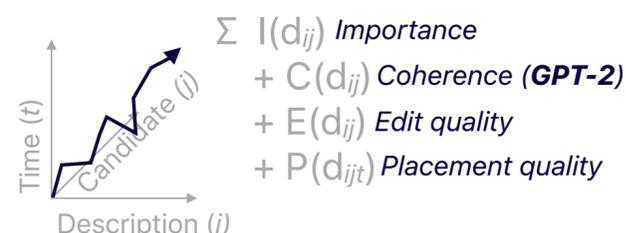0:06 Red flowers against a white house and blue sky.

Edit text, Adjust timing
Lengthen music

Algorithms

Generate description alternatives
"Gaby and Thomas are back in their living room sitting on the couch."
"Gaby and Thomas are back on the couch."
"Gaby are back." …

Optimize text and placement

$$\Sigma \ I(d_{ij}) \ \textit{Importance}$$
$$+ \ C(d_{ij}) \ \textit{Coherence} \ (\textbf{GPT-2})$$
$$+ \ E(d_{ij}) \ \textit{Edit quality}$$
$$+ \ P(d_{ijt}) \ \textit{Placement quality}$$

Time (t)
Candidate (i)
Description (i)

Evaluations

AD Novices efficiently created AD with fewer errors using Rescribe. 45% (± 7%) overlapping —> 0%

Blind users most preferred the Rescribe edited inline and *new extended-inline descriptions*.

AD Experts liked that Rescribe preserved agency while automating the tedious editing.

Future

Accessibility Prediction for Search

5/7 Somewhat Accessible
73% of video contains speech. The speech is descriptive, but has many visual references (3 per minute). Visual changes occur infrequently (5 shots per minute); few of the on-screen objects are described (20%).

Feedback for Capture-Time

Automating Description Generation