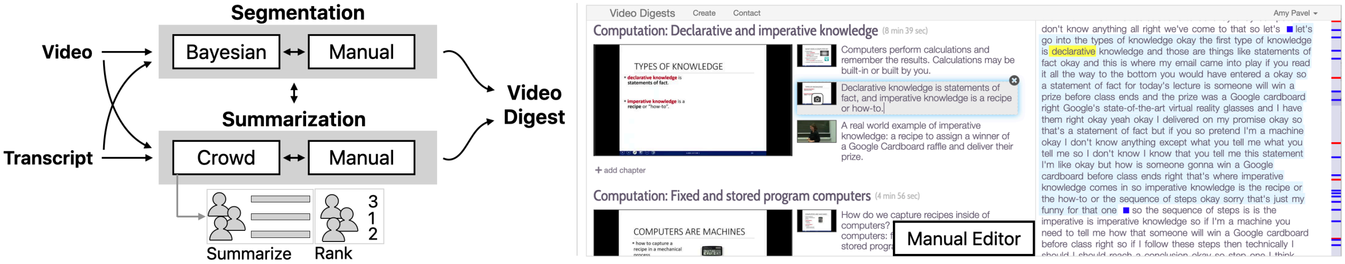
Effective communication requires communicators to address the goals, abilities, and contexts of their audiences. My research leverages the interconnectedness of communication mediums (e.g., spoken audio, multiple video sources, text transcripts) to enable people to fully participate in and better work with digital communication. Using my systems, remote content creators more effectively collaborate [1], video authors efficiently create accessible descriptions for blind viewers [2, 3], and network administrators scalably defend against personalized attacks [4]. As a systems researcher in Human-Computer Interaction, I embed machine learning technologies (e.g., NLP, Computer Vision) into new human interactions that I then deploy to test. In the near term, my research answers what intelligent systems are possible and desirable, and over the long term drives work in machine learning and computer science in directions beneficial to humans.

In this document, I characterize my work via projects where I improved digital communication for a diverse set of audience goals, abilities, and contexts: (1) enabling information-seeking in videos [5, 6, 7], (2) making visual communication non-visually accessible [2, 3, 8], and (3) augmenting conversational context in text-based communication for clarity and scale [1, 4]. My future work envisions a world in which communicators and audience members can create new human-AI systems to support their full participation in communication.

VISUALIZING INFORMATION IN VIDEOS

For decades, people communicated information online via text including articles, how-to instructions, and expert blogs; today, informative videos including explainer videos, how-to demonstrations, video essays, and vlogs increasingly eclipse their text counterparts. While videos provide rich audio and visuals, their timeline- or transcript-based navigation lacks the *domain-specific structure* of text documents (*e.g.*, headings, paragraphs) that cues key information, enabling people to skim for key points or browse for content of interest. I contributed interactive systems that combine the benefits of video and structured text through my *interactive video abstractions*.

A *Video Digest* [5] is an expert-inspired abstraction that enables students to search, browse, and skim lecture videos by linking videos to structured text representations with titled chapters and summarized sections. Creating a video digest by hand is time consuming – experts manually scrub back and forth along the video timeline, playing and pausing the video to identify each segment boundary, then summarize each section by watching the video to recall the content. As lecture videos convey most information via speech, I introduced interactive *transcript-based segmentation* and *transcript-based summarization* to help authors efficiently create video digests. To achieve high-quality digests, the system first segments the video into sections by applying Bayesian topic segmentation over a transcript time-aligned to the audio at word level (using HMM/Viterbi-based forced alignment), then uses a crowd-powered summarize-and-rank method to create section summaries. To group sections into chapters, the system re-applies topic segmentation over the crowd-sourced summaries. Authors can then refine the digest using the aligned transcript of the video in our manual authoring tool (Figure 1). In a study (N=192), people using manual and automated digests recalled more key points of the video than those using the video or



Computation: Declarative and imperative knowledge (8 min 39 sec)

TYPES OF KNOWLEDGE
- declarative knowledge is statements of fact.
- imperative knowledge is a recipe or "how-to".

Computers perform calculations and remember the results. Calculations may be built-in or built by you.

Declarative knowledge is statements of fact, and imperative knowledge is a recipe or how-to.

A real world example of imperative knowledge: a recipe to assign a winner of a Google Cardboard raffle and deliver their prize.

Computation: Fixed and stored program computers (4 min 56 sec)

COMPUTERS ARE MACHINES
- how to capture a recipe in a mechanical process

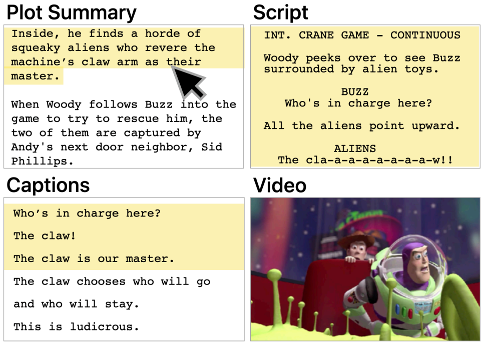
How do we capture recipes inside of computers? computers: f stored progra

don't know anything all right we've come to that so let's ■ let's go into the types of knowledge okay the first type of knowledge is declarative knowledge and those are things like statements of fact okay and this is where my email came into play if you read it all the way to the bottom you would have entered a okay so a statement of fact for today's lecture is someone will win a prize before class ends and the prize was a Google cardboard right Google's state-of-the-art virtual reality glasses and I have them right okay yeah okay I delivered on my promise okay so that's a statement of fact but if you so pretend I'm a machine okay I don't know anything except what you tell me what you tell me so I don't know I know that you tell me this statement I'm like okay but how is someone gonna win a Google cardboard before class ends right that's where imperative knowledge comes in so imperative knowledge is the recipe or the how-to or the sequence of steps okay sorry that's just my funny for that one ■ so the sequence of steps is is the imperative is imperative knowledge so if I'm a machine you need to tell me how that someone will win a Google cardboard before class right so if I follow these steps then technically I should I should reach a conclusion okay so step one I think

Figure 1: Video Digests system diagram (left) and manual transcript-based editing interface (right). Video Digests takes in a video and its transcript, and provides computational support for segmenting and summarizing, authors can manually refine its outputs using the editing interface.

aligned text transcript alone. We deployed video digests in a large introductory CS course where students scribed their lectures, and we released our authoring tool to instructors.

I extended interactive video abstractions to support searching and browsing in film – a domain where, unlike lecture videos, expert-authored texts already exist. In formative interviews, film professionals, including producers and scholars, revealed that they often searched for moments, settings, actions, or characters in film in order to communicate ideas and inform future art. To make tedious search tasks trivial, I created *SceneSkim* [6]. SceneSkim aligns existing documents including plot summaries (to revisit particular moments), scripts (to search and skim objects, actions, settings and characters) and captions (to search sound effects and dialogue) to each other and to the film (Figure 2). SceneSkim aligns caption words and their predicted phonemes to features in the audio, aligns the caption words to the script dialogue words using Needleman-Wunsch, and aligns script segments to plot summary sentences using a dynamic programming approach that maximizes similarity of important words in matched segments. SceneSkim enables searches across many movies at once and indicates *alignment scent* to show if a line of text likely appears in the source media. Film studies scholars successfully generated and answered novel film questions with SceneSkim. We aligned texts for 819 films to answer existing questions in film literature, in less than 5 minutes of search time each (*e.g.*, did red lights flash whenever a couple was on screen across 3 movies? – no, refuting prior work). The work sparked discussions with Pixar, film instructors, and university librarians that indicated the utility of the work; it also appeared in press including Engadget. Going forward, I will further study ML-assisted search tools and extend interactive video abstractions to new media (*e.g.*, 360° video [7]).


Plot Summary

Inside, he finds a horde of squeaky aliens who revere the machine's claw arm as their master.

When Woody follows Buzz into the game to try to rescue him, the two of them are captured by Andy's next door neighbor, Sid Phillips.

Script

INT. CRANE GAME - CONTINUOUS

Woody peeks over to see Buzz surrounded by alien toys.

BUZZ
Who's in charge here?

All the aliens point upward.

ALIENS
The cla-a-a-a-a-a-a-a-w!!

Captions

Who's in charge here?

The claw!

The claw is our master.

The claw chooses who will go

and who will stay.

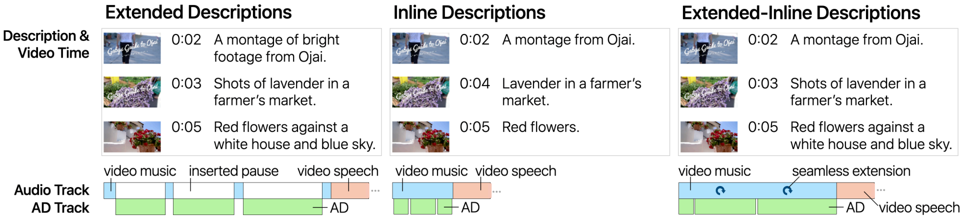
This is ludicrous.

Video

Figure 2: Users navigate in films using summaries, scripts, and captions, and across films with *search* (not pictured).

MAKING VISUAL COMMUNICATION NON-VISUALLY ACCESSIBLE

The visual content that makes videos engaging can render videos inaccessible to blind and visually impaired audience members, restricting access to a vast array of communication.


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.

Inline Descriptions

0:02 A montage from Ojai.
0:04 Lavender in a farmer's market.
0:05 Red flowers.

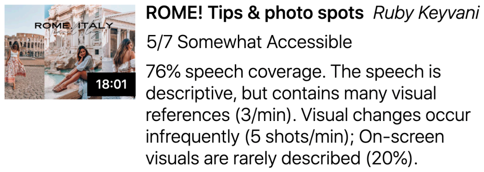
Extended-Inline Descriptions

0:02 A montage from Ojai.
0:03 Shots of lavender in a farmer's market.
0:05 Red flowers against a white house and blue sky.

Figure 3: Authors write descriptions, then Rescribe edits the text (top) and underlying audio (bottom) to make extended, inline, or extended-inline descriptions.

To make videos accessible, professionals create audio descriptions (AD), or narrations of the visual content in the video. In formative interviews, professional audio describers revealed that ADs are challenging to create as it requires iteratively editing descriptions to fit within limited time between video dialogue [2]. We created a tool called *Rescribe* that integrates ML and user interaction to support experts and novices in crafting high-quality ADs. Rescribe first uses a CNN to classify video "gaps" that include primarily music or silence. Authors describe their video and record their narration; then, Rescribe *jointly optimizes the text and media* to fit the description in the gaps by paraphrasing description sentences, adjusting placement, and subtly looping the underlying audio track (to fit in more text content). Before optimization, Rescribe generates candidate paraphrases for each description by dropping subtree combinations from the dependency parse tree. Rescribe scores each text paraphrase candidate based on domain-specific properties (*e.g.*, relevance to script, preference for nouns/verbs, audio length) and GPT-2 perplexity to avoid highly improbable (often grammatically incorrect) sentences. Authors can iteratively refine their descriptions working in concert with the tool. Rescribe reduced the amount of descriptions overlapping with speech from 45% to 0% in the same amount of time. Blind participants preferred Rescribe-edited descriptions, and most preferred our *extended-inline descriptions* (that optimally extends underlying audio using beat detection and MFCC-based beat similarity) that Rescribe uniquely enables. Experts, initially skeptical of automation, wanted to use the tool as it automated the tedious part of their job. The system was built in collaboration with AD professionals. I aim to continue close corporate collaborations to facilitate broad deployment of such systems. I presented Rescribe at UIST 2020 where it was the only paper highlighted in both the Opening CSCW/UIST Plenary and the Closing UIST Keynote.

Beyond the traditional approach of making videos accessible by describing them, I have explored new ML-driven interfaces that leverage connections between visual content and narration to make video search and viewing tasks non-visually accessible by: (1) surfacing *inherently accessible videos* during video search through novel automated video accessibility metrics [8], and (2) encouraging authors to make media *accessible at capture-time* through automated element-level feedback [3]. The systems we built to support search and author feedback were informed by three studies with blind participants [8] and well-established community guidelines [3]. Both systems resulted in recently accepted CHI 2021 papers, and are now being deployed by the two students that I mentored on these projects. Apart from video, people use images, memes, and GIFs to share information (including during time-sensitive emergencies [9]), or reactions in conversations on social media. To inform future automated work, I've collaborated on qualitative work to understand what blind people would like to know about memes, GIFs [10], and images on Twitter [11]. Reusable content like memes and GIFs provide an opportunity to scale human-written descriptions (*e.g.*, match a GIF to a set of base GIFs), or to adapt descriptions to new contexts (*e.g.*, automatically describing changes to a base meme image).



ROME! Tips & photo spots *Ruby Keyvani*

5/7 Somewhat Accessible

76% speech coverage. The speech is descriptive, but contains many visual references (3/min). Visual changes occur infrequently (5 shots/min); On-screen visuals are rarely described (20%).

Figure 4: A video search result augmented with our accessibility metrics.

AMPLIFYING CONVERSATIONAL TEXT COMMUNICATION

In addition to adapting informative communication to audience goals, and abilities, I've explored how to augment the clarity and scale of conversational text communication through computational support.

First, I have explored adding tone and reactions to text communication. Giving and receiving criticism is an essential but delicate part of creative work. Formative interviews with professionals who exchange feedback on videos revealed they preferred to give in-person critiques, but they gave emailed text critiques due to scheduling constraints [1]. Text critiques are time-consuming to give and challenging to understand due to a lack of shared video context. I built *VidCrit*, a tool to offer *asynchronous video feedback* that let reviewers speak their comments and add context by drawing on the video or scrubbing on the video timeline. VidCrit segments the transcribed comments and links them to the source video time. Critiquers added twice as many critiques with VidCrit as they did in text. Critique recipients liked the ease of finding context for critiques and gleaning immediate reactions.

Second, I have scaled expert text conversations to new contexts. Expert "scambaiters" converse with scammers online to waste scammer time and elicit information, then share successful scripts (*e.g.*, on 419eater.com). To scale expert work, I led a project creating a conversational agent that *adapts expert scripts to new contexts* of each unique scam. The agent benefits from the performance of a large language model and expert scripts by conditioning dialogue generation on *semantic exemplars*, or the extracted semantic frames of a retrieved expert response. We submitted this work to NAACL 2021 [4], finding that the model performed better than state-of-the-art retrieval-based dialogue methods and generative models alone. The conversational agent won the DARPA program-wide competition for anti-scam defence, resulting in 2 years of project funding. We are continuing to develop the system for deployment within agencies and companies defending against scam and spear-phishing attacks.

RESEARCH AGENDA

My research goal is to empower people to fully participate in communication. Here, I discuss future research questions I am excited to explore:

How can we democratize the creation of human-AI communication tools? I have created bespoke systems to detect and expose links between multiple communication mediums in order to support information-seeking goals, and make communication accessible. But, my work alone does not address all communication domains or user needs. To make these tools more equitable, I will democratize their creation through: (1) an alignment toolkit that surfaces primitives that let novices create interfaces like SceneSkim [6] and VidCrit [1] in their own domains, and (2) interactive systems where users specify through examples and feedback how they want a system to behave. For example, rather than selecting between the three given ADs in Rescribe [2], users would customize automated descriptions to their interests by providing examples, guiding questions, or feedback. Shrinking the gap between the creators and users of AI-driven tools is necessary to an equitable future of technology.

How can we create representative data labels to improve human-AI interactions? Visual labeling is often performed as a non-expert, context-agnostic task. In practice, communication is best interpreted by different people for different audiences (*e.g.*, crowdsourced translations for multiple subtitle languages). Video Digests [5] and Rescribe [2] showed that people with different backgrounds provide different perspectives in abstracting video content that are valuable to different audiences (*e.g.*, an expert may have *expert blind spots*, a community outsider might miss terminology). In a study comparing domain expert and domain novice ADs, the accuracy and detail of a description was limited by the describer's *domain* expertise (their experience with the topic), and *craft* expertise (their experience creating ADs) [12]. I will explore collaborative social systems to collect a diversity of labels for visual data, understand biases (and expertise) of the labelers, and build interactions to flexibly combine and consume the results for applications including descriptions for blind users, expert search, and education.

What interactions enable communication for people with cognitive and motor impairments? In-person communication relies on non-verbal cues to regulate turn-taking, and it is often challenging for people using a gaze- or switch-based AAC device to participate – especially during group conversation [13]. We are exploring algorithms that give AAC users more control in conversations by referencing audio from earlier in the conversation, and by controlling turn-taking through movement of a robotic "sidekick." While multiple modes of communication can be helpful for all users, people with cognitive impairments can particularly benefit from multiple ways of understanding information. In the future, I will explore how best to support users with impairments in consuming informational media with AI-driven interactions.

What can applications in accessibility teach us about human-AI interactions at large? Accessible technology is uniquely challenging because (1) it has a high potential value (*e.g.*, provide a way for a user to access information that is otherwise unavailable to them), and (2) the experience of disability is personal and intersectional, so no single technology is a panacea. I will continue working with people with disabilities to understand their real-world problems, thus informing new, beneficial technology.

I am a systems HCI researcher enabling accessible and effective human-to-human communication through computational tools. I augment interactive systems with AI to achieve more than humans or computers could alone, and I base the development of the systems on qualitative research with domain experts. I have co-authored papers with over 40 researchers at 10 institutions, and collaborated with people in diverse fields including NLP, Graphics, Computer Vision, Robotics, and Humanities. I aim to continue such collaborations in the future, in order to amplify expert work through computational assistance.

REFERENCES

[1] **Amy Pavel**, Dan B Goldman, Björn Hartmann, and Maneesh Agrawala. “VidCrit: video-based asynchronous video review”. In: UIST 2016.

[2] **Amy Pavel**, Gabriel Reyes, and Jeffrey P Bigham. “Rescribe: Authoring and Automatically Editing Audio Descriptions”. In: UIST 2020.

[3] Yi-Hao Peng, Joon Jang, Jeffrey P Bigham, and **Amy Pavel**. “Say It All: Feedback for Non-Visually Accessible Presentations”. In: Conditionally accepted to CHI 2021.

[4] Prakhar Gupta, Jeffrey P Bigham, Yulia Tsvetkov, and **Amy Pavel**. “Controlling Dialogue Generation with Semantic Exemplars”. In: Submission to NAACL 2021.

[5] **Amy Pavel**, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. “Video Digests: a browsable, skimmable format for informational lecture videos.” In: UIST 2014.

[6] **Amy Pavel**, Dan B Goldman, Björn Hartmann, and Maneesh Agrawala. “Sceneskim: Searching and browsing movies using synchronized captions, scripts and plot summaries”. In: UIST 2015.

[7] **Amy Pavel**, Björn Hartmann, and Maneesh Agrawala. “Shot orientation controls for interactive cinematography with 360 video”. In: *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. UIST 2017.

[8] Bruce Liu, Patrick Carrington, Anthony Chen, and **Amy Pavel**. “What Makes a Video Non-Visually Accessible?” In: Conditionally accepted to CHI 2021.

[9] Cole Gleason, Stephanie Valencia, Lynn Kirabo, Jason Wu, Anhong Guo, Elizabeth Jeanne Carter, Jeffrey Bigham, Cynthia Bennett, and **Amy Pavel**. “Disability and the COVID-19 Pandemic: Using Twitter to Understand Accessibility during Rapid Societal Transition”. In: ASSETS 2020.

[10] Cole Gleason, **Amy Pavel**, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. “Making GIFs Accessible”. In: ASSETS 2020.

[11] Cole Gleason, **Amy Pavel**, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. “Twitter A11y: A Browser Extension to Make Twitter Images Accessible”. In: CHI 2020.

[12] Kimberly Do, Cole Gleason, Jeffrey P Bigham, and **Amy Pavel**. “How Does Domain Expertise Impact Audio Description?” In: Preparation.

[13] Stephanie Valencia, **Amy Pavel**, Jared Santa Maria, Seunga Yu, Jeffrey P Bigham, and Henny Admoni. “Conversational Agency in Augmentative and Alternative Communication”. In: CHI 2020.