How to Design Voice Based Navigation for How-To Videos

Minsuk Chang
School of Computing
KAIST
minsuk@kaist.ac.kr

Anh Truong
Adobe Research
Stanford University
anhlt92@cs.stanford.edu

Oliver Wang
Adobe Research
owang@adobe.com

Maneesh Agrawala
Stanford University
maneesh@cs.stanford.edu

Juho Kim
School of Computing
KAIST
juhokim@kaist.ac.kr

ABSTRACT

When watching how-to videos related to physical tasks, users' hands are often occupied by the task, making voice input a natural fit. To better understand the design space of voice interactions for how-to video navigation, we conducted three think-aloud studies using: 1) a traditional video interface, 2) a research probe providing a voice controlled video interface, and 3) a wizard-of-oz interface. From the studies, we distill seven navigation objectives and their underlying intents: pace control pause, content alignment pause, video control pause, reference jump, replay jump, skip jump, and peek jump. Our analysis found that users' navigation objectives and intents affect the choice of referent type and referencing approach in command utterances. Based on our findings, we recommend to 1) support conversational strategies like sequence expansions and command queues, 2) allow users to identify and refine their navigation objectives explicitly, and 3) support the seven interaction intents.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in interaction design**;

KEYWORDS

How-to Videos, Video Tutorials, Voice User Interface, Conversational Interaction, Video Navigation

ACM Reference Format:
Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to Design Voice Based Navigation for How-To Videos. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3290605.3300931

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300931

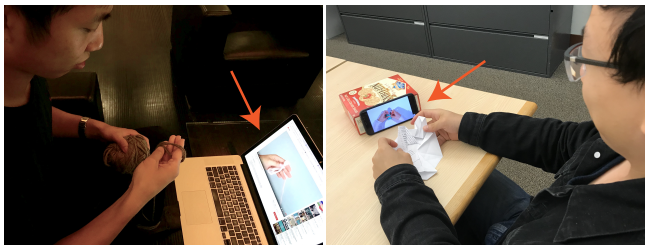

Figure 1: People use how-to videos to accomplish a variety of physical tasks. The person in the left photo is using the video on his laptop to learn how to stitch while the person on the right is attempting to fold an origami turtle by following the video on his phone. Since both of these are hands-on tasks, it would be difficult for these people to navigate the videos using traditional click or touch based input modalities.

1 INTRODUCTION

People are increasingly turning to online how-to videos as guides for learning new skills or accomplishing unfamiliar tasks. YouTube searches for how-to videos are growing 70% year to year and as of 2015, 67% of the millennials believe they can find anything they want to learn on YouTube [20].

Instead of watching these how-to videos passively, viewers actively control the video to pause, replay, and skip forward or backwards while following along with the video instructions. We identified that these active control moments arise when 1) the viewer is presented with non-tutorial content (e.g., chit-chat and general introductions), 2) the viewer fails to match the pace, content, or context of the video (e.g., needing more time to complete a step, not understanding the instruction, or needing to compare the outcome in the video with the viewer's own result), and 3) the viewer is only interested in a specific part of the tutorial (e.g. a particular method or step). Using traditional mouse and keyboard interfaces, viewers commonly navigate tutorial videos by either sequentially scrubbing through the video timeline to examine the preview thumbnails or click-guessing on the

timeline until the resulting video frame matches the desired point of interest.

However, many how-to videos teach *physical* tasks (e.g., playing an instrument, applying makeup, etc.) that involve interaction with real world objects. When following along with such videos, viewers need their hands both to execute the task and to control the video. Since they can't do both at once, viewers must alternate between the two operations. In having to alternate between these two operations, viewers incur a costly context switch: they must stop concentrating on the task itself to instead concentrate on controlling the video. Controlling the video alone can be both difficult and tedious using traditional timeline-based interfaces. This work specifically focuses on exploring navigation solutions for how-to videos for *physical* tasks.

Voice-based user interfaces (e.g., Apple Siri, Amazon Alexa, and Google Assistant) are becoming increasingly ubiquitous in commercial devices and provide a potential alternative for controlling how-to videos of physical tasks. Current voice-based video navigation systems (e.g., those in the web accessibility initiative [29], or virtual assistants with displays like Amazon Echo Show) support basic operations for video navigation such as pause, play, rewind, or fast forward (e.g., by 20 seconds). While these systems provide some help in the context of how-to videos, they are not specifically designed for this domain. Rather, they directly translate of low-level remote control operations (pause, play, etc.) into voice commands. An important question is whether this low-level remote-control-like interface is suitable for voice-driven video navigation interfaces for how-to videos, and if not, how should a useful voice interface for navigating how-to videos be designed?

In this work, we explore the mental models users have when navigating how-to videos, and report how the navigation objectives affect the command utterances. In particular, we answer the following research questions:

RQ1 What are the different types of navigation objectives and user intentions when "actively controlling" how-to videos?

RQ2 Do these objectives and intents affect users' linguistic choices for voice commands? If so, what linguistic characteristics are related to the objectives and intents?

RQ3 What are the challenges and opportunities for designing voice navigation interactions for how-to videos?

To this end, we report findings from three user studies. We conducted a think-aloud study of 20 participants using a YouTube interface to accomplish a how-to task, a 16-participant think-aloud study using a custom-built voice based video navigation research probe, and a Wizard-of-Oz study with 6 participants. From these studies, we distill a list of design recommendations for how to build a navigation system for how-to videos with voice and visual feedback.

In summary, our contributions include:

- A range of user objectives and contexts for navigating how-to videos. They are pace control, content alignment, video control, reference, replay, skip, and peek.
- An analysis of how these objectives and contexts, when applied to voice interactions for how-to video navigation, affect linguistic characteristics of user command utterances.
- A set of design recommendations for voice interaction with how-to videos.

2 RELATED WORK

Our work extends previous research in the domains of browsing and navigation for conventional video interfaces, interaction techniques for how-to videos, and the design of voice user interfaces.

Video Browsing and Navigation

Previous works have investigated interaction techniques for navigating videos beyond simple timeline interfaces. For example, Dragicevic et al. [7] found that direct manipulation of video content (via dragging interactions) is more suitable than direct manipulation of a timeline interface for visual content search tasks. Swift [18] and Swifter [19] improved scrubbing interfaces by presenting pre-cached thumbnails on the timeline. Hajri et al. [1] use personal browsing history visualizations to enable users to more effectively spot previously watched videos.

Similar to how our studies investigate voice UI patterns, Li et al. [17] investigated digital video browsing strategies using traditional mouse-based graphical user interfaces. They found the most frequently used features were time compression, pause removal, and navigation using shot boundaries. Crockford et al. [5] found that VCR-like control sets, consisting of low-level pause/play operations, both enhanced and limited users' browsing capabilities, and that users employ different playback speeds for different content.

A study on video browsing strategies reported that in-video object identification and video understanding tasks require different cognitive processes [6]. Object identification requires localized attention, whereas video understanding requires global attention.

We extend this line of video navigation research and specifically investigate problems for users of how-to videos for physical tasks. Our goal is to understand users' browsing and navigation objectives from the perspective of voice interaction and linguistic characteristics.

Interaction Techniques for How-To Videos

Interactions with tutorials has been a popular research subject in the HCI community. Web tutorials serve a variety of needs from expanding skill sets to experiencing experts' practices [13].

MixT [4] automatically generates step-by-step mixed media tutorials from user demonstrations and Duploblock [8] infers and tracks the assembly process of a snap-together block model in real-time. Panopticon [9] displays multiple sub-sequences in parallel to present a rapid overview of the entire sequence.

For software tutorials, Nguyen et al. [22] found that users complete tasks more effectively by interacting with the software through direct manipulation of the tutorial video than with conventional video players. Pause-and-play [23] detected important events in the video and linked them with corresponding events in the target application for software tutorials. FollowUs [14] captured video demonstrations of users as they perform a tutorial so that subsequent users can use the original tutorial, or choose from a library of captured community demonstrations of each tutorial step. Similarly, Wang et al. [31] showed that at-scale analysis of community-generated videos and command logs can provide workflow recommendations and tutorials for complex software.

Specific to educational videos, LectureScape [11] utilized large scale user interaction traces to augment a conventional interface, while ToolScape [10] utilized storyboard summaries and an interactive timeline to enable learners to quickly scan, filter, and review multiple videos without having to play them.

We extend this line of research by investigating how voice interaction could be designed to expand users' capabilities and assist in accomplishing physical tasks.

Designing Voice User Interfaces

Most recent work on voice interfaces is done in an "assistant" context. For example, Porcheron et al. [24] studied how voice-based virtual assistants are made accountable to and embedded into conversational settings such as dinner tables. Myers et al. [21] also reported that while natural language processing errors occur the most, other types of errors frustrate users more, and users often take a guessing approach when voice interfaces fail. Moreover, nearly one quarter of all user-assistant exchanges were initiated from implicit conversational cues rather than from plain questions [28]. Users frequently use a diverse set of imprecise temporal expressions in both communication and planning, and have a variety of expectations about time inputs for virtual assistants [25].

Researchers have also successfully implemented voice user interfaces for specific user facing tasks. PixelTone [15] enabled users to edit photos using both speech and direct manipulation. ImageSpirit [3] enabled users to verbally refine image search results using the automatically extracted labels. Apparition [16] enabled users to sketch their interface, describe verbally where crowd workers and sketch recognition algorithms translate the input into user interface elements, add animations, and provide Wizard-of-Oz functionality.

Using voice to navigate how-to videos introduces challenges that have not yet been explored. We expect our investigation to lead the design of more usable voice interactions for how-to videos.

3 EXPERIMENT OVERVIEW

We designed three experiments to address the following research questions: RQ1. What are the different types of navigation objectives and user intentions when "actively controlling" how-to videos? RQ2. What are the linguistic characteristics of user command utterances with respect to their navigation objectives and intents? RQ3. What are challenges and opportunities for designing voice navigation interactions for how-to videos?

First, we conducted an experiment to understand how users currently interact with how-to videos for physical tasks. Specifically, to understand different types of navigation objectives and user intentions (RQ1), we examined when users broke away from their tasks and actively control the video using a conventional mouse-based video navigation interface. To minimize workflow interruptions and capture habitual behaviors, we designed this study as a think-aloud experiment and instructed our participants to vocalize the motivation behind their actions. For the analysis, we extracted the sequence of commands from the user study recordings. At each user interaction point, we examined the subsequent user action and commands to enumerate and determine the different types of pauses and jumps. We identified three types of pauses (pace control, content alignment, and video control) and four types of jumps (reference, replay, skip, and peek).

The second experiment studied how user behavior changes when using simple voice commands rather than a mouse interface. In particular, we examined when this type of simple voice command based interface fails. We focused on linguistic characteristics of users' voice commands in relation to users' intents(RQ2), our analysis identified possible design recommendations to alleviate such failures and support voice-based video navigation. We phrased this experiment as a think-aloud study with a simple voice-based video navigation interface that we built (Figure 2). Similar to study 1, we extracted the command sequences from the user study recordings. For each of the seven interaction types, we

counted the frequency of user utterances to make a dictionary of common command intents.

Finally, we conducted a wizard-of-oz study where we allow users to express complex voice commands. The goal of this study was to better understand the challenges and opportunities of designing voice user interactions for navigating how-to videos when users are not limited to basic commands (RQ3). From this study, we gathered more "conversational" control actions, identified user intentions behind each control action, and captured what users would "ideally" like to do in a voice navigation system with no technical constraints. For the analysis, we used open coding with thematic analysis. Two authors independently extracted themes from the study recordings and independently brainstormed hypotheses and frames of explanation. Through rounds of discussions, they agreed that the "challenges and opportunities" framework is most explanatory.

4 STUDY 1 - UNDERSTANDING HOW PEOPLE CURRENTLY NAVIGATE HOW-TO VIDEOS

In the first study, we asked participants to perform a physical task by watching a how-to video with a conventional mouse interface and the standard YouTube video player. Our primary observation points were when and how users pause and jump during the session.

In this experiment, we focused on two specific how-to tasks: learning to play a song on an instrument and learning to apply makeup. We recruited 20 participants on usertesting.com, an online user study platform. We recorded participant screens and think-aloud audio for all sessions. The music experiment consisted of 10 participants (8 male, 2 female, average age: 40, min: 21, max: 71) who regularly watch tutorial videos to learn how to play songs on their musical instruments. The makeup experiment consisted of 10 participants (all female, average age: 33, min: 21, max: 56) who regularly watch makeup tutorial videos.

We instructed participants to select a video of their choice from YouTube that consisted of an unfamiliar song or an unfamiliar makeup style, respectively. They then had to follow along with the video tutorial and describe their thought process out loud. We specifically asked participants to explain what they were trying to achieve whenever they controlled the video (i.e., pause, play, rewind, etc.).

Findings

Participants picked tutorial videos with average lengths of 4 minutes 40 seconds (0:29 min, 12:08 max) for music tutorials, and 7 minutes 51 seconds (3:39 min, 11:04 max) for makeup tutorials. The average session length was 15 minutes 32 seconds (10:15 min, 25:55 max) for learning a song, and 20 minutes 21 seconds (10:42 min, 40:58 max) for learning a makeup look. Participants spent on average 3 times the length of the video following it.

Task (# of participants)	Music (10)	Makeup (10)
Total minutes of tutorial	46:47	78:34
Total # of pauses	33	65
Total # of jumps	27	16
Per minute pauses	0.42	1.38
Per minute jumps	0.47	0.21

Table 1: Frequency of user pauses and jumps in a traditional interface

General Impressions. For the task of learning a song, we observed that participants tended to either try to get a rough understanding of the entire song or to focus on learning a specific part. Two participants pointed out that this is due to an inherent characteristic of the task where it would take days or even weeks of practice for them to fully learn an entire song.

We also found that content of the video as well as task characteristics affected the distribution of interactions. For example, as seen in Table 1, the music session users jumped backward or forward over twice as often per minute as those in the makeup task. We observed this is because for makeup how-to videos, the distinction of where each step begins and ends is much more apparent as each step builds on top of the previous step. This allows natural pauses in between steps, giving space for users to catch up. In fact, per minute of viewing, users paused three times as often in the makeup task as the music task. Music how-to videos usually do not contain explicit steps, making the beginning and end points of a navigation unit ambiguous and more user dependent. Also each makeup how-to video was specific to certain aspects of makeup in general, e.g., eye makeup or contour makeup, whereas music tutorials usually try to tackle the entire song in one video.

Types of Pause Interactions. In this experiment, we observed 98 total pauses across both tasks (Table 1). From these tasks, we observed three different types of pause interaction.

Pace Control Pause. The most common type of pause was a pause to gain more time (78 of 98 pauses: 29 of 33 in music, 49 of 65 in makeup). This happens when the user understands the video content but fails to match the pace of the video. With this pause, the user is trying to finish one step before moving onto the next. Unlike other types of pauses, in a pace control pause, the user is usually detached from the video, while concentrating on the physical task. Once users are caught up to the video using pace control pauses, they often end the pause by pressing play and without performing any other type of video navigation.

Content Alignment Pause. The second type of pause is a pause to compare what's in the video with what's in the hands of the user. This pause precedes the user checking to make sure that their state is similar to that of the video. For example, after the pause, users say *"I'm just trying to see if this is what he—the video instructor—has done."* or *"I need to see if this is it."* while making the comparison between what's in the video and what's in the hands of the user. Users often observe the paused video frame several times during these pauses. Content alignment pauses made up 9 out of 98 total pauses observed: 2 of 33 in music, 7 of 65 in makeup. Out of the 9 pauses of this type, 7 pauses were at the end of a step, right after the next step has begun, where the information in the still frame has not made a full transition yet. During the content alignment pauses, the user attention is split between the video and the physical task.

Video Control Pause. The final type of pause we observed is a pause for further video control. Reference jumps In this case, the user pauses the video and searches for the next navigation target point on the timeline by either guess-clicking, or scrubbing and examining the thumbnails. In this use case, the user's attention is entirely in the video. Video control pauses occurred in 8 of 98 total pauses observed: 1 of 33 in music, 7 of 65 in makeup. Video control pauses are always followed by a jump interaction described in detail in the next section.

Types of Jumping Interactions. In this experiment, we observed 43 total jump interactions from both tasks (Table 1). These jumps are broadly split into forward and backward jumps, and we break down the different user motivations that we observed. Users carried out *jumps* by pressing right or left arrow keys on the keyboard, or by clicking on the point of interest on the timeline, or by scrubbing the timeline.

Reference Jump. The first type of jump we observed is a *reference jump*. We observed 5 (out of 43) reference jumps (3 in music, 2 in makeup). In this case, the user jumps backwards in the video to remind themselves of something they saw in the past. Users typically only need to see a still image of the video for this jump.Usually a forward jump back to the original position is followed by a reference jump to continue where they left off.

Replay Jump. A *replay jump* is a different form of backward jump, where the user wants to re-watch a segment of the video again. We observed 24 (out of 43) replay jumps (21 in music, 3 in makeup). This jump happens when the user needs to get a better understanding, clarify a possible mistake, or to assure that the current understanding is correct. This jump if often followed by a play or a pause interaction.

Skip Jump. A *skip jump* is a type of forward jump where the user wants to skip content that is less interesting, like the introduction of the channel or the personal life of the YouTuber. We observed 10 (out of 43) replay jumps (2 in music, 8 in makeup). When the goal is to skip introductory content, the target is almost always "the beginning of the actual tutorial". Since the user cannot tell where exactly "the actual tutorial" begins, skip jumps happen in multiples. This forward jump often is followed by another skip jump or a play interaction.

Peek Jump. The second type of forward jump is a *peek jump*, where the user wants to skip ahead to see what the user should expect after performing one or a number of steps. We observed 4 (out of 43) replay jumps (1 in music, 3 in makeup). This happens when users want to check the intermediate or the final result in order to prepare and also check if the user is on the right track. The goal is not to skip the current step, but rather to help by anticipating future steps. A peek jump is often followed by a jump back to the original position in the video.

Other Interactions. Users sometimes paused the video to get the surprise introduction of an additional tool or material like a guitar capo or an unconventional makeup tool. We observed this 3 times (1 of 33 in music, 2 of 65 in makeup). Users also sometimes let the video play while concentrating on the physical task without paying much attention, but still listening to it. We observed this 26 times (10 in music, 16 in makeup).

5 STUDY 2 - UNDERSTANDING HOW PEOPLE NAVIGATE HOW-TO VIDEOS USING A BASIC VOICE UI

Results of our first study show that people often stop and jump within the videos, which requires frequent context switches. To understand what differences might be observed in users' thoughts and preferences of *voice* interactions in navigating how-to videos, we built a voice-enabled video player as a research probe. This research probe served as a "tools for design and understanding" [30] not a prototype interface to suggest new interaction techniques. We used our research probe as an apparatus to observe and elicit similarities and differences in user behavior in relation to the different types of pauses and jumps observed with a traditional mouse-based interface.

With our research probe, the user can play, pause, jump backward and forward by specifying the jump interval in seconds or minutes, speed up, slow down, mute, unmute, increase volume, and decrease volume. We used a grammar consisting of popular variations of the above commands (Table 2). We piloted the research probe and enumerated the list of command variants, and iterated until no new variant was observed. The interface also indicates when it is listening 2 and transcribes commands in real time to provide visual

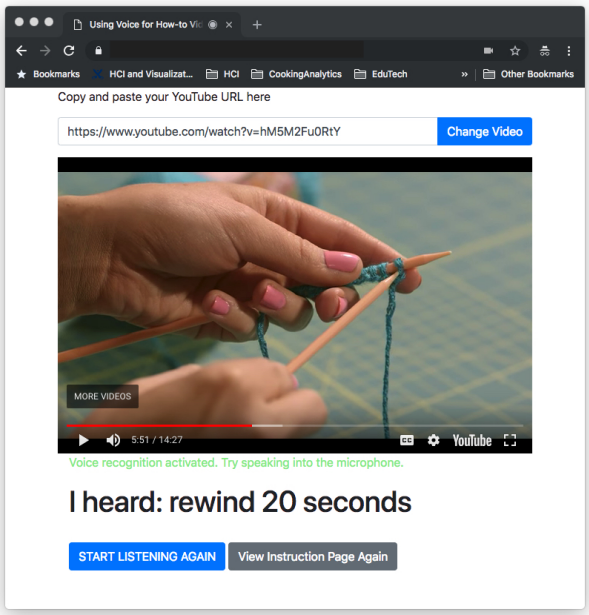

Figure 2: Our research probe allows users to select a YouTube video of their choice and control it with voice. The interface also indicates when it is listening and transcribes commands in real time to provide visual feedback to the user. For example, "I heard: rewind 20 seconds".

Main Command	Popular Variants
play	resume, go, start, begin
pause	stop, wait, hold on
mute	volume off
unmute	volume on
louder	volume up
quieter	volume down
fast forward	skip ahead, skip
rewind	go back, back
faster	speed up
slower	slow down

Table 2: List of commands supported by our system

feedback to the user. Transcription uses HTML5 Web Speech API. Figure 2 shows the research probe interface.

We conducted a study that mirrored the previous one, except that participants were asked to use our voice-based research probe instead of YouTube, and a third task, knitting, was added to cover a more diverse set of tasks. We recruited 16 participants in total. The music task consisted of 7 (4 male and 3 female, average age of 35) participants, the makeup task consisted of 4 (all female, average age of 26) participants, and the knitting task consisted of 5 (all female, average age of 31) participants. None of the participants had participated in our previous experiment.

Features	Tags
Referent Type	Time, action, object
Referencing Styles	Specify an interval, specify a point, contextual description, content description, no referent
Reference Direction	Backward (rewind), forward (fast forward)
Causes	Mismatch in pace, mismatch in content, mismatch on context

Table 3: Features used for analyzing referencing utterances. Since we are interested in how users make references to what and why, user utterances are annotated with one tag from each of the features.

To minimize priming participants to use specific utterance choices, they were not instructed on what voice commands were available when communicating with the interface. When the system did not support a specific command, participants were instructed to think aloud and carry on.

Similar to the first study, we annotated each interaction and their occurrence counts, focusing on the three types of pauses and four types of jumps we have identified. Repeated utterances due to speech recognition failure were counted as only once. To further break down the composition of each command utterances, we annotated how users made references to navigation targets when they performed jump interactions using the features presented in Table 3.

Findings

The average length of the tutorial videos participants picked was 9 minutes 2 seconds (4:56 min, 15:29 max) for music tutorials, 13 minutes 47 seconds (10:09 min, 17:45 max) for the makeup tutorials, and 14 minutes 44 seconds (6:21 min, 23:10 max) for the knitting tutorials. The average session length was 25 minutes 39 seconds (17:29 min, 41:23 max) for learning a song, 34 minutes 54 seconds (26:24 min, 43:44 max) for learning a makeup look, and 26 minutes 5 seconds (16:22 min, 33:05 max) for learning a new knitting pattern.

Types of Pause Interactions. We observed that the command "stop" is used mostly for **video control pauses** (18 out of 25 "stop"s) where the command was followed by a jump. In contrast, "stop video" was used mostly for **content alignment pauses**, where the command was followed by a play command (13 out of 15 "stop video"s).

We also observed that users use the word "stop" to indicate a higher sense of urgency, or a need to navigate to a very specific point in the video. Here are some of the example use cases we've observed:

(1) "Go back by a little", "Go back by a bit", "Stop"
(2) "Stop", "I want to stop at this specific note (piano)"
(3) "Stop", "I'm missing something here"
(4) "Stop", "I don't know what's going on"

Participant 3 in the makeup experiment is an illustrative example. This participant used "stop video" and "play video" ten times each throughout the session to follow along the instruction. But when she needed to quickly go back and check how mascara was done in a hurry before moving on to the next step, she used "stop!" followed by "I need to go back to mascara now, I'm doing something different".

We found that both "pause" and "pause video" were frequently used for **content alignment jumps** and **pace control jumps**, "pause" was used 24 times out of 43, while "pause video" was used 10 times out of 12 for these jumps.

Types of Jump Interactions. Two frequently used commands for backward jumps were "go back" and "rewind". In this experiment, we observed 23 **replay jumps** and 28 **reference jumps**. We noticed for replay jumps, users use less concrete commands than for reference jumps, such as, "start from beginning", "let me see that again". "go back **about** 30 seconds", "go back **just a bit**", "go back **by little**", "go back **to the beginning**".

However, for **reference jumps**, users tend to be more specific, and repeat multiple times to find the exact target, using commands like "go back 30 seconds" and "go to 2 minute mark". Users also repeat concrete backward commands to find a specific desired position. Also, some users said "go back to where I paused" to go back to the original position in the video before the backward jump, which indicates the user is expecting the system to remember this position when performing jumps.

We observed forward jumps that refer both to contextual details, as well as concrete units of time. Examples are "skip about 30 seconds", "skip to the lesson". "skip to the makeup (as opposed to cleansing)", and "fast forward a bit", and "skip to next step." We could not observe any different linguistic pattern between **skip jumps** and **peek jumps**.

One reason for this might be because users do not know the target position of the peek or the skip because they are both in the future (later in the video). In contrast, backward jump targets are usually those users have already seen once, which enables users to refer to their memory for more specific descriptions.

General Impressions. Participants found the concept of using voice to navigate how-to videos useful for their tasks. From the music experiment, P3 noted *"it's an interesting experience having to stop and play video without taking my hands off my guitar, it's wonderful."* and P4 also noted *"'this is a very powerful tool, especially if you're doing something with your hands."* From the makeup experiment, P3 reported *"I really like that I can get my products ready without touching the video"*. From the knitting experiment, P1 commented *"I love being able to use voice to control the video while I'm knitting so I don't have to stop from knitting."*.

Task (# of participants)	Music (7)	Makeup (3)	Knitting (6)
Total minutes of tutorial	63:25	41:21	88:28
Total # of pauses	46	32	28
Total # of jumps	43	22	21
Per minute pauses	0.73	0.78	0.31
Per minute jumps	0.68	0.54	0.24

Table 4: Frequency of user pauses and jumps in voice-enabled interface

We also noticed users would "stop" or "pause" the video before jumps a lot more often while using voice user interfaces. Jumps with specific references like "go back 20 seconds" is dependent on both the current position and the target, and without the pause the current position would keep changing, resulting inconveniences to adjust the interval or make multiple subsequent jumps. With the mouse interactions, in contrast, users are only specifying the target position and not the origin.

6 STUDY 3 - UNDERSTANDING EXPECTATIONS OF VOICE UI FOR HOW-TO VIDEOS

From the previous study, we learned that users' navigation intents affect their linguistic choices for command utterances. We also observed that commonly supported voice commands are limited to simple words, that it can be difficult for users to express their intents with a restrictive command space, and that it is difficult for systems to understand the intents. For example, different backward jump intents for "stop" and "pause" can only be understood in context of other commands before and after the stop, specifically analyzing preceding and succeeding commands and user goals, which is impractical in application settings where users need systems to understand the user intents in real time.

To inform how to disambiguate voice commands and corresponding user intents for navigating how-to videos, we conducted a Wizard-of-Oz experiment to learn how users would naturally converse for video navigation in the absence of these constraints. Participants were instructed to find a knitting video appropriate to their expertise level, and follow the video while performing the task.

We invited 6 participants (3 male, 3 female), 5 of whom were complete novices in knitting, and 1 of whom was a hobbyist with self-reported expertise level of intermediate. A researcher was sitting next to the participant as the wizard

Challenges	Opportunities
Problems from interacting with video	Visual feedback strategies
Problems from interacting with voice	Conversational strategies
Problems from interacting with wizard	Wizard strategies

Table 5: Resulting code book for analysis of Wizard of Oz Study(Study 3)

video controller, watching the tutorial video with the participant. The participant could only control the video by talking to the wizard video controller. Users were encouraged to converse without any technical barriers in mind. We also conducted semi-structured interviews at the end of each the study to further understand noticeable behaviors exhibited during the sessions. The average duration of the video tutorial used was 13 minutes 18 seconds (7:08 min, 14:48 max). The average duration of the sessions was 32 minutes 38 seconds (19:48 min, 40:32 max). Each participant was rewarded with a 15 USD giftcard.

We follow the recommendations for thematic analysis [2], and iteratively analyzed the interview data and the conversation logs three times in total with an interval of at least one day between sessions to enhance validity. Authors on our research team watched and open coded all screen recordings and think-aloud sessions. Then, the identified codes were reconstructed to the codes most relevant to our research questions through discussions. The codes were constructed around two themes: challenges, and opportunities of voice user interface in navigating how-to videos (Table 5).

Voice based interactions between users and systems can be seen as a type of conversation. To understand user strategies from their command utterances, we analyzed dialogue data between the user and the wizard using the turn-taking framework in conversational analysis [26].

Findings

Challenge 1 - Characteristics of How-to Videos. Because of the sequential nature of the video (there is the concept of an **unknown future**), users often make a guess to navigate forward in the video, or they have to watch less relevant or less interesting segments. One illustrative example was when P2 asked the wizard *"could we change the speed to like 1.25? I want to slow it back down when she actually starts"*. Also, in the interview, P1 noted *"If I don't know what's coming up, I'm very uncomfortable skipping. If there's an outline, I would, but otherwise I don't know how much to skip or how much to speed it up by."* and P4 commented *"If I knew where I was going, I feel like I would progress better"* . From this we can conclude that it was difficult for users to anticipate what is coming up, and dealing with this uncertainty is an important design issue.

Challenge 2 - Voice Inherent Problems. When participants used a specific time interval for jumps, it often required multiple adjustments to navigate to the target even when the participant had a good sense of where the target was. In this case, **command parsing delays** become an important user interface limitation. P4 explained "saying go back by how much creates a delay between the point where I started saying the command (the point where I started saying the command) and when I finish the sentence and for you (wizard) to understand it. So I would have to say, for example, go back 30 seconds, and then go back 5 more."

7 DESIGN RECOMMENDATION

Based on our findings and understanding from the three studies, we propose the following recommendations for designing voice based navigation for how-to videos.

Support Conversational Strategies

Support sequence expansions and command queues as both are strategies users often use. For example, supporting users to perform a single command multiple times in a row by recognizing "again" following "go back 5 seconds", and supporting users to place multiple commands in one utterance like "go to 2 minutes and 4 second mark and pause" would be useful.

Support Iterative Refinements of Commands

Users often need multiple tries to find the intended navigation target. It is because a) what users remember can be different from the part they are looking for or vice versa, b) sometimes users don't remember, and c) sometimes users remember but don't know the exact vocabulary like the names of knitting techniques and tools. Good examples are support for descriptive commands and keyword search in transcripts.

Support Interactions with User Context

Designing voice commands for how-to videos is not about supporting a single command, but understanding the higher level user intent behind the utterance is crucial. We identified all seven interaction intents (pace control pause, content alignment pause, video control pause, reference jump, replay jump, skip jump, and peek jump) that can be supported. One possible solution in distinguishing them is to set up the command vocabulary such that each intent has its unique keyword. For each of the intents, specific design recommendations are as follows:

Pace Control Pause & Content Alignment Pause. This is the pause for users to gain more time to finish the step. Keep a record of the point of pause for future references. Allow the user to easily compare the progress or the state of the user

and those of the video by supporting various examination features like zoom or taking screenshots.

Video Control Pause. This is the pause where the user has an intention to navigate to other places in the video. Keep a pointer to the origin and provide "comeback" to this point, as it will often happen after jumps.

Reference Jump. Provide "memory". Augment users' memory to enable more accurate references by using features like markers and object annotations. Also, as reference jumps often happen in multiples, make the subsequent search processes easier, by suggesting updates or narrowing down of the interval of jumps.

Replay Jump. Support replay by allowing users to set a loop interval and the number of iterations.

Skip jump. Provide a visual summary of the remaining sections of the video for users to skip around. Approaches using instruction milestones, key frames, or frames containing user-specified keywords are all suitable.

Peek Jump. Provide a "comeback" feature to the origin position of the jump.

8 DISCUSSION

Our study and interview results highlight the challenges, opportunities, and user expectations of using voice interfaces for video tutorials. We first discuss the user challenges of adapting to a VUI from a GUI when learning physical tasks with video tutorials. We then discuss how our research methodology of designing a series of experiments in progression can be extended to designing VUI for other applications and domains.

Transitioning from GUI to VUI

Mouse vs Voice. We found voice interfaces require an initial pause while issuing subsequent commands. For example, when using voice input in Study 2, users issued a pause command before every rewind command. In contrast, when using the traditional mouse interface, users never paused the video before skipping to another point. We think this is due to the time it takes for the user to speak the voice command and for the system to process it. Also, the target is directly specified with mouse (or touch) but with voice the target is often specified relative to the current position of the video. For example, if the user does not pause the video before jumping, the original reference keeps moving, and the interval they had thought of will not get them to the point they intended. As a result, the larger consequence is that voice-based interactions require more steps to achieve the same objective (i.e., pause + jump) than mouse-based interactions do (i.e., click).

Uncertainty from Unseen Content. When trying to navigate a video tutorial using voice, users make more concrete references to the past, whereas users have challenges describing later part of the video. For traditional video interfaces, scrubbing and clicking around are often used a solution to quickly peeking into the future. However, for voice interfaces, such a solution does not exist yet. Handling this uncertainty is an important design issue which would improve the usability of voice interactions for videos.

Recognition of Speech Input and Command Learnability. While the concept of using voice to navigate how-to videos is generally welcomed, participants also reported well-known problems of voice user interfaces. Speech recognition does not always work as expected, especially if users have accents or are in a noisy environment. In Study 2, nine participants also reported difficulty in figuring out the available commands. All participants showed frustration when the system did not respond to their command. Usability of VUI suffers due to relatively poor recognition, poor learnability and discoverability of available commands, and lack of feedback.

User Expectations and Opportunities

Multimodal Reference Strategies. Users often wanted to make references to the objects and speaker in the video. In Study 3, when users were making multiple corrections to navigate to a specific target point in the video, users have the advantage of utilizing the paused video frame as additional references, often employing **object references**. P1 explained "I look at the frame and the state of the objects that appear to see if it's before or after (the point I want to jump to)". Also, users often made **transcript references**, referring to things that the tutor has said. For example, P3 commanded the system "can you repeat that again? How she did multiples of four, the part where she said multiples of four". We believe voice assistants with a visual display could utilize this finding, as the referent needs to be visual or textual.

Conversational Strategies. Users often employ conversational strategies such as latent conversational intents and sequence expansions. Participants employed a lot of latent **conversational intents** frequently used in human-human conversations [26]. For example, participants said "Can I see it again, 10 seconds before?", "Can I see the last three knit?", and "Can you move it back to when she shows how it looks like from the back?". While a semantic answer to all of those questions would be a yes or a no, we contextually understand that these are requests, not participants asking for permission. Also, "I want to go back to the first time she does this with the second needle" by P6 is not a remark, but a command.

Participants often used **sequence expansion**, also heavily used in human-to-human conversations. For example, P4 said ("rewind 30 seconds until 3 minutes", "again") and ("slow

it down to .5 and play from 4 minutes", "okay, from 3:55"). Users expected the wizard to have a memory of previous commands, and believed the wizard has the shared context.

Another strategy participants often used was including **command queues** in a single utterance. For example, P2 said "could we change the speed to like 1.25? I want to slow it back down when she actually starts the tutorial" in the beginning of the video in the introductions segment. This requires multiple levels of understanding. The system would need to understand the first command to change the playback speed, detect when the tutorial starts, and remember to change the playback speed to normal. This is a powerful strategy that gives users more room to concentrate on the tasks by queuing multiple commands. P3 explicitly mentioned in the interview that "I want to sequence commands, set rules like if there is a second needle, slow it down." These techniques are applicable to generic voice interaction design, and existing services such as Siri and Google Assistant already support parsing "can I" questions as commands. However, all other conversational strategies described above is not supported.

Wizard Strategies. Users want "smarter" voice interactions that resemble a conversation with another human; the conversational agent that has complete knowledge of the viewing experience and can track progress. In Study 3, there are strategies participants used by relying on the "wizard" being another human. P6 requested "scrub the entire video" during the experiment and P4 noted in the interview "recognizing the repetition of commands like how you (wizard) did would be useful. If the system learned what I mean when I just say go back, and not having the description afterwards would be best".

Progression of Experiment Designs

In order to understand a user-centric design of voice interfaces for video tutorials, we carefully designed the three studies posing users in three scenarios in progression. Starting from how users use the current interface without voice interaction, to a basic adoption of voice interaction, to a Wizard-of-Oz interface with "ideal" voice interactions. We were able to create a taxonomy of current interactions, classify user intents in video navigation, and understand user challenges and opportunities for eliciting design recommendations.

We believe this progression of experiment design is generalizable to understanding how to design voice interactions for new applications or other domains like driving and exercising. For example, when understanding how to design voice interactions while driving, the same progression of studies could be just as effective. Understanding the current practices and needs of voice interactions while driving, and then using a design probe using a voice interface probe to understand opportunities and challenges, and then carrying out a Wizard-of-Oz study to elicit ideal interaction scenarios.

9 LIMITATIONS AND CONCLUSION

One limitation of this work is that our design implications are observational with respect to 41 participants across three tasks. It is possible that other behaviors will emerge in tasks that have substantially different characteristics. Future work that analyzes voice commands at scale might be able to detect additional patterns.

Similar to how large-scale clickstream analysis can aid instructors to better understand learners' hurdles in watching online educational videos and reason about unsatisfactory learning outcomes [27], and to improve our understanding of interaction peaks and dropouts [12], we believe an at-scale analysis of voice interaction traces has potential to further our understanding on how to design better voice user interface. An an initial step, a live deployment of a voice-based interface for navigating how-to videos would be a worthwhile effort.

We also acknowledge there are other possible perspectives that we did not touch upon in the analysis. For example, how navigation behavior and intents differ for novices and experts, and for first time videos and revisiting videos.

There are also practical issues related to implementing voice user interfaces that we do not address in this work. While speech recognition is rapidly improving, it is still far from perfect, and as observed in our experiments, speech recognition failures and delays cause user frustrations.

An additional technical challenge is related to audio source separation. In practical settings, audio coming from the video and possibly from the task itself may interfere with the user's voice commands, which would result in even poorer command recognition. While wireless headphones and earbuds are becoming more popular, there may be some situations where the user cannot use a dedicated headset.

Additionally, many ambiguities in users' voice command utterances that we discovered can be resolved by designing a system that understands and adapts to the *intent* of user and the *content* of the video. We believe determining these two variables in the wild is an interesting research challenge.

In conclusion, we present the first set of experiments that explicitly target voice based user interactions for navigating how-to videos of physical tasks. We examined how different user navigation objectives and intentions affect their word choices in voice command utterances, and reported a lexicon of types of interactions and the motivating factors behind these commands. Despite the limitations listed above, we believe that our experiments will be informative for researchers and practitioners who design voice-based video navigation systems, which have the potential to play a large role in how learning systems of the future operate.

10 ACKNOWLEDGEMENT

We thank our study participants for their time and feedback. We also thank Adobe Research for their support in this research.

REFERENCES

[1] Abir Al-Hajri, Gregor Miller, Matthew Fong, and Sidney S Fels. 2014. Visualization of personal history for video navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1187–1196.

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[3] Ming-Ming Cheng, Shuai Zheng, Wen-Yan Lin, Vibhav Vineet, Paul Sturgess, Nigel Crook, Niloy J Mitra, and Philip Torr. 2014. ImageSpirit: Verbal guided image parsing. *ACM Transactions on Graphics (TOG)* 34, 1 (2014), 3.

[4] Pei-Yu Chi, Sally Ahn, Amanda Ren, Mira Dontcheva, Wilmot Li, and Björn Hartmann. 2012. MixT: automatic generation of step-by-step mixed media tutorials. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 93–102.

[5] Chris Crockford and Harry Agius. 2006. An empirical investigation into user navigation of digital video using the VCR-like control set. *International Journal of Human-Computer Studies* 64, 4 (2006), 340–355.

[6] Wei Ding and Gary Marchionini. 1998. *A study on video browsing strategies*. Technical Report.

[7] Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video browsing by direct manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 237–246.

[8] Ankit Gupta, Dieter Fox, Brian Curless, and Michael Cohen. 2012. DuploTrack: a real-time system for authoring and guiding duplo block assembly. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 389–402.

[9] Dan Jackson, James Nicholson, Gerrit Stoeckigt, Rebecca Wrobel, Anja Thieme, and Patrick Olivier. 2013. Panopticon: A parallel video overview system. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. ACM, 123–130.

[10] Juho Kim. 2013. Toolscape: enhancing the learning experience of how-to videos. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2707–2712.

[11] Juho Kim, Philip J Guo, Carrie J Cai, Shang-Wen Daniel Li, Krzysztof Z Gajos, and Robert C Miller. 2014. Data-driven interaction techniques for improving navigation of educational videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 563–572.

[12] Juho Kim, Philip J Guo, Daniel T Seaton, Piotr Mitros, Krzysztof Z Gajos, and Robert C Miller. 2014. Understanding in-video dropouts and interaction peaks in online lecture videos. In *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 31–40.

[13] Ben Lafreniere, Andrea Bunt, Matthew Lount, and Michael Terry. 2013. Understanding the Roles and Uses of Web Tutorials. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[14] Benjamin Lafreniere, Tovi Grossman, and George Fitzmaurice. 2013. Community enhanced tutorials: improving tutorials with multiple demonstrations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1779–1788.

[15] Gierad P Laput, Mira Dontcheva, Gregg Wilensky, Walter Chang, Aseem Agarwala, Jason Linder, and Eytan Adar. 2013. Pixeltone: A multimodal interface for image editing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2185–2194.

[16] Walter S Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P Bigham, and Michael S Bernstein. 2015. Apparition: Crowdsourced user interfaces that come to life as you sketch them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1925–1934.

[17] Francis C Li, Anoop Gupta, Elizabeth Sanocki, Li-wei He, and Yong Rui. 2000. Browsing digital video. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 169–176.

[18] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2012. Swift: reducing the effects of latency in online video scrubbing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 637–646.

[19] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2013. Swifter: improved online video scrubbing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1159–1168.

[20] David Mogensen. 2015. I want-to-do moments: From home to beauty. *Think with Google* (2015).

[21] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 6.

[22] Cuong Nguyen and Feng Liu. 2015. Making software tutorial video responsive. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1565–1568.

[23] Suporn Pongnumkul, Mira Dontcheva, Wilmot Li, Jue Wang, Lubomir Bourdev, Shai Avidan, and Michael F Cohen. 2011. Pause-and-play: automatically linking screencast video tutorials with applications. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 135–144.

[24] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 640.

[25] Xin Rong, Adam Fourney, Robin N Brewer, Meredith Ringel Morris, and Paul N Bennett. 2017. Managing uncertainty in time expressions for virtual assistants. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 568–579.

[26] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*. Elsevier, 7–55.

[27] Tanmay Sinha, Patrick Jermann, Nan Li, and Pierre Dillenbourg. 2014. Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*. 3–14.

[28] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 208.

[29] w3c. 2018. Web Accessibility Initiative. https://www.w3.org/WAI/.

[30] Jayne Wallace, John McCarthy, Peter C Wright, and Patrick Olivier. 2013. Making design probes work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3441–3450.

[31] Xu Wang, Benjamin Lafreniere, and Tovi Grossman. 2018. Leveraging community-generated videos and command logs to classify and recommend software workflows. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 285.