CHI 2019

How to Design Voice Based Navigation for How-to Videos

Minsuk Chang
KAIST

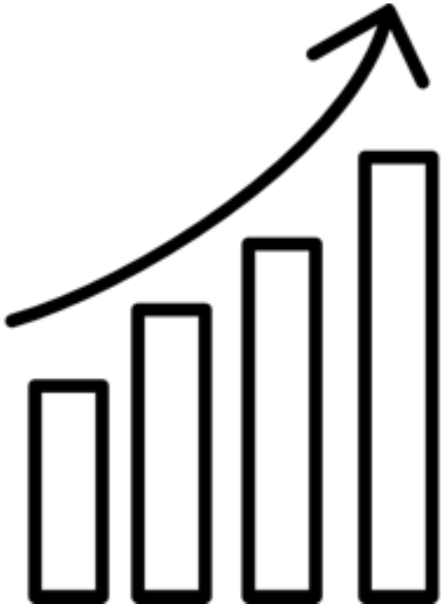
Anh Truong
Adobe Research
Stanford University

Oliver Wang
Adobe Research

Maneesh Agrawala
Stanford University

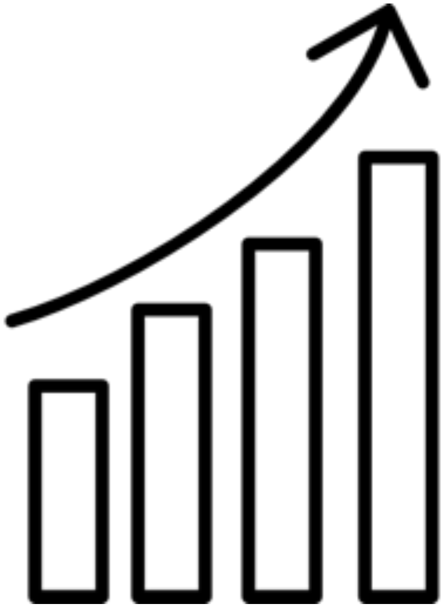
Juho Kim
KAIST

70% yearly increase in “how-to” video searches



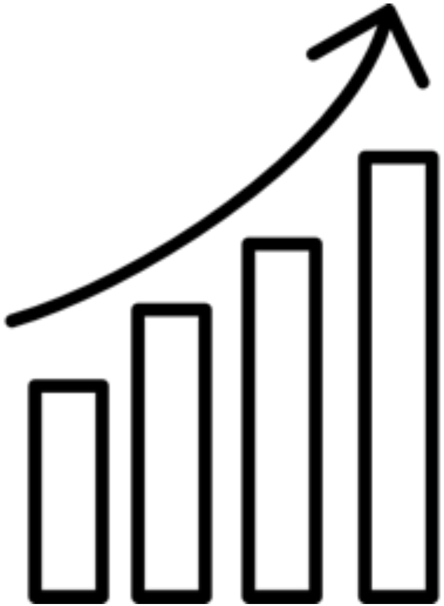
100 million hours watched

70% yearly increase in “how-to” video searches

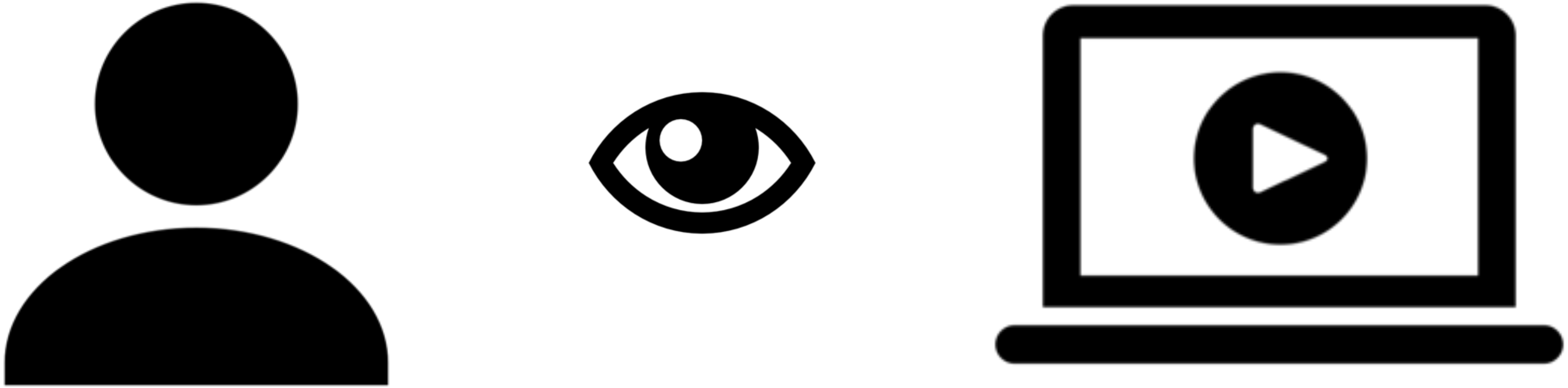


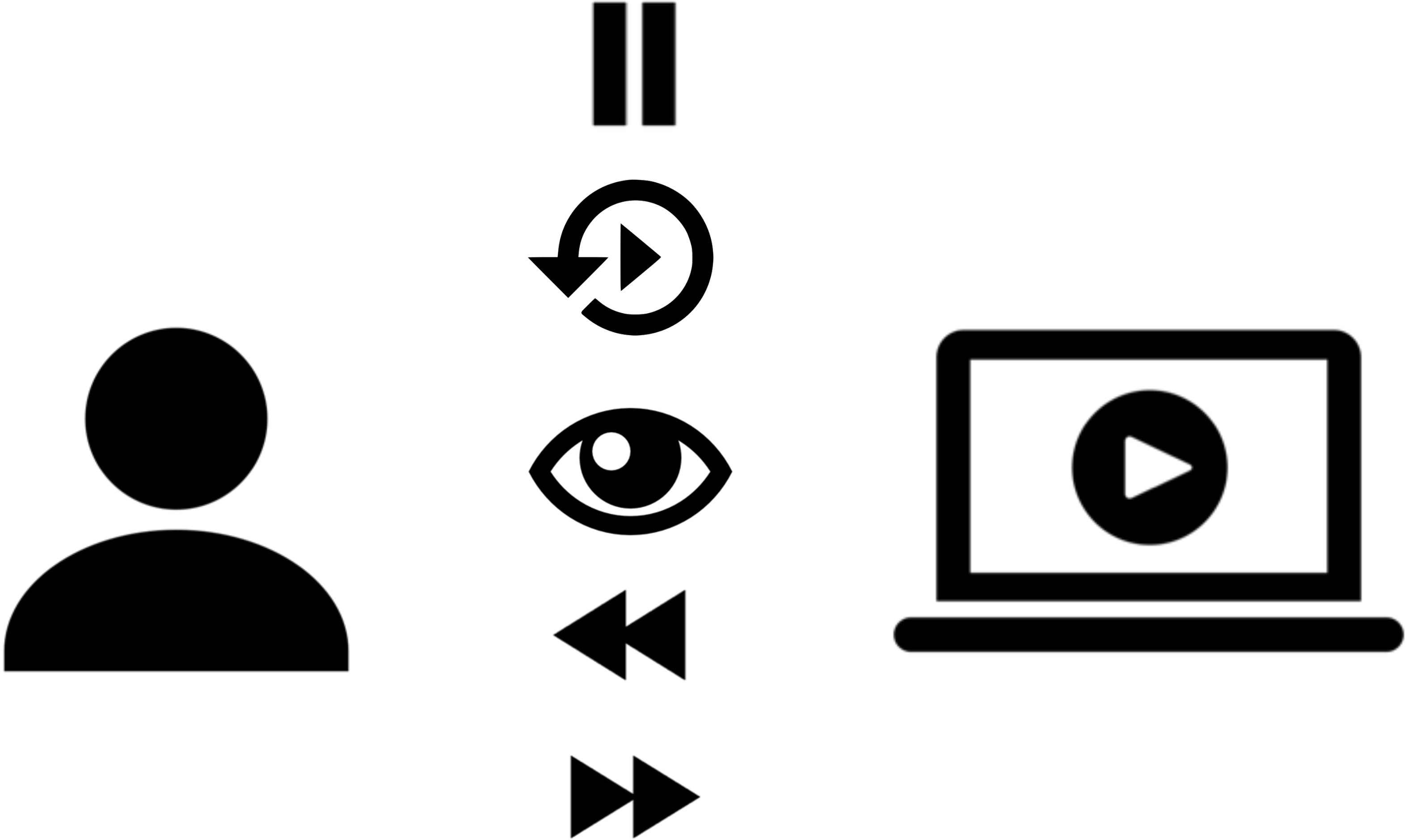
100 million hours watched **in 2015**

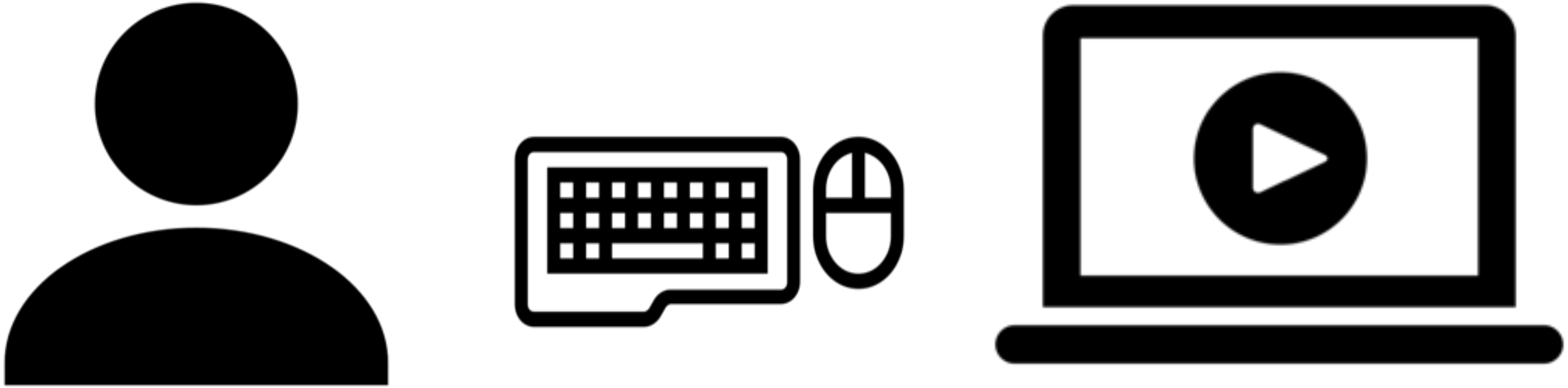
70% yearly increase in “how-to” video searches



100 million hours watched **in 2015 in North America**







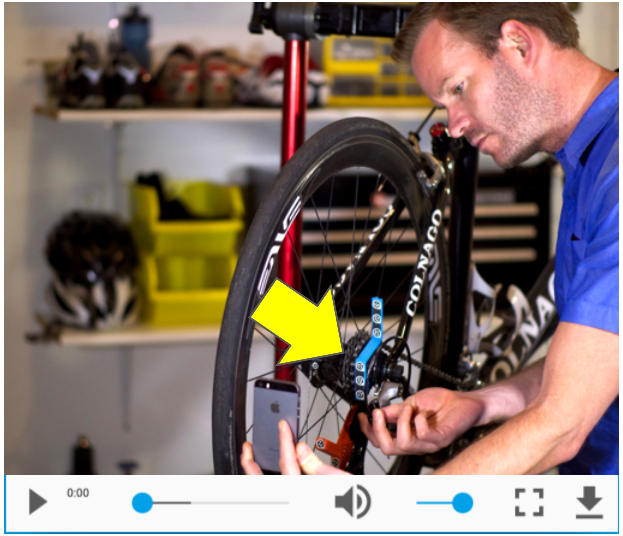
HOW TO
KNIT
A SCARF
STEP-BY-STEP
FOR TOTAL BEGINNERS
brought to you by
SHEEP & STITCH
0:00 / 27:10

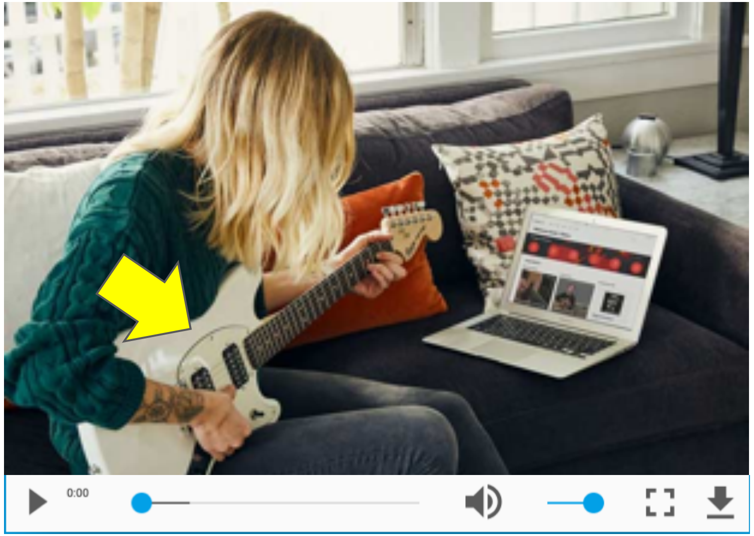
ENVE
COLNAGO
COLNAGO

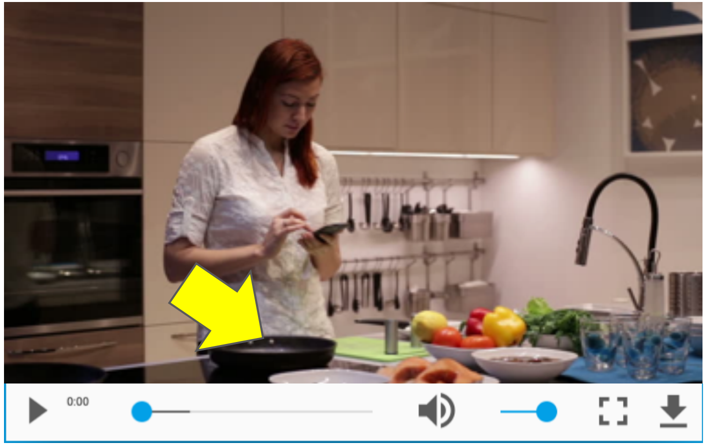
COLNAGO

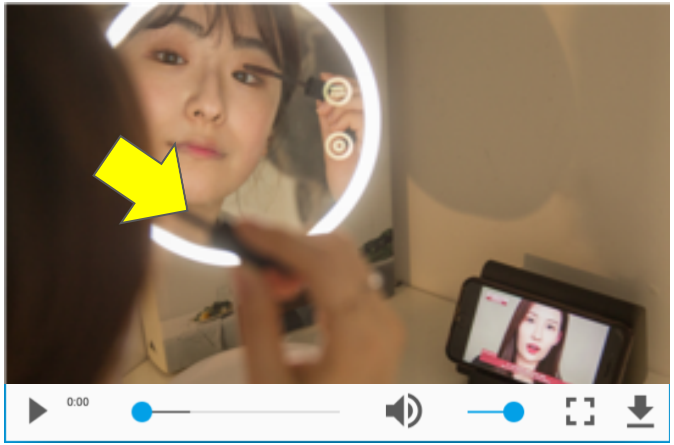
COLNAGO

COLNAGO


COLNAGO
0:00


0:00


0:00


0:00

I HEAR WHAT YOU'RE SAYING
BUT I STILL ONLY HAVE TWO HANDS

Apple Siri

What can I help you with?

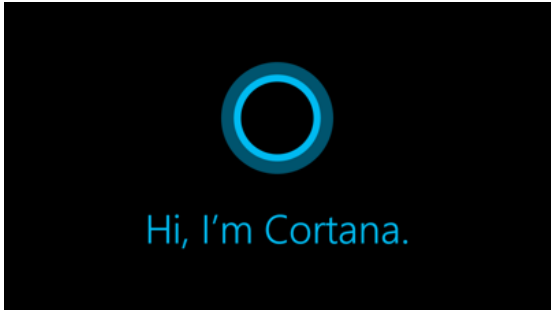
Google Assistant

Hi, how can I help?

What can you do?

Amazon Alexa

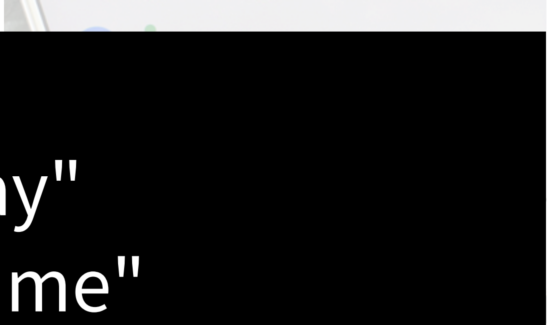
Microsoft Cortana



Hi, I'm Cortana.

Apple Siri

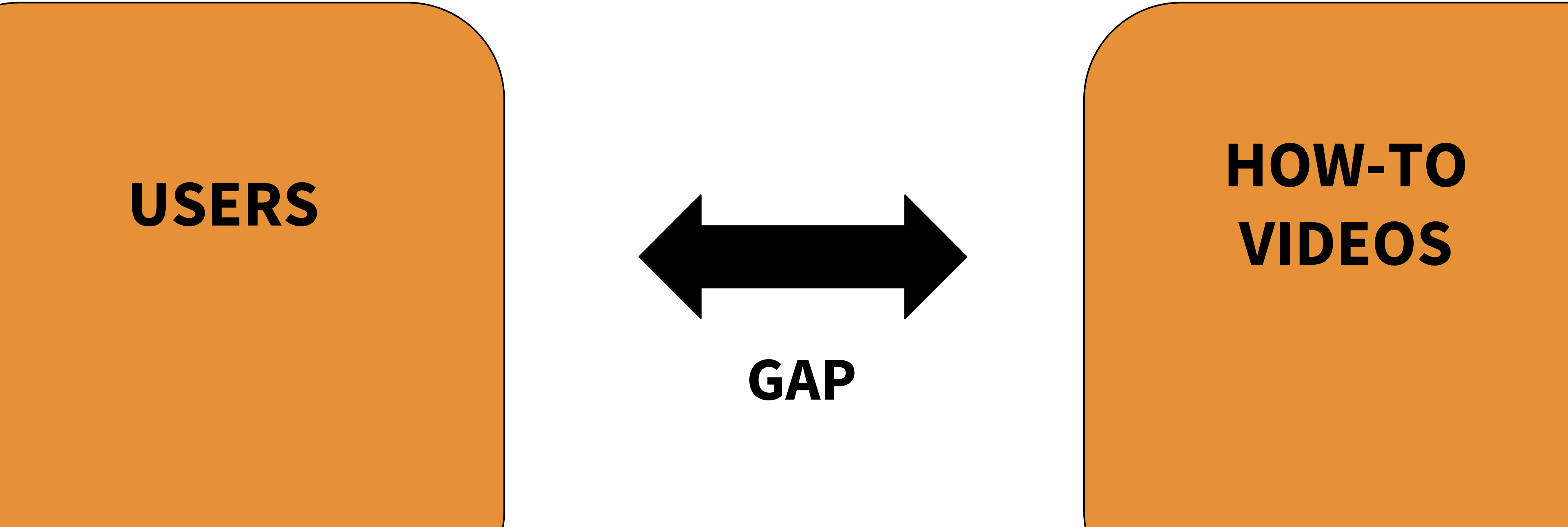
Google Assistant



"Play"
"Resume"
"Pause"
"Fast-forward"
"Rewind"

How should a useful voice interface for navigating how-to videos be designed?

USERS



USERS

GAP

HOW-TO VIDEOS

VOICE INTERFACE

USERS

HOW-TO VIDEOS

1. **What are the navigational needs for how-to videos?**

2. How are they realized with remote-control like voice interactions?

3. What would an ideal voice interface be like for how-to videos?

1. What are the navigational needs for how-to videos?

2. **How are they realized with remote-control like voice interactions?**

3. What would an ideal voice interface be like for how-to videos?

1. What are the navigational needs for How-to videos?

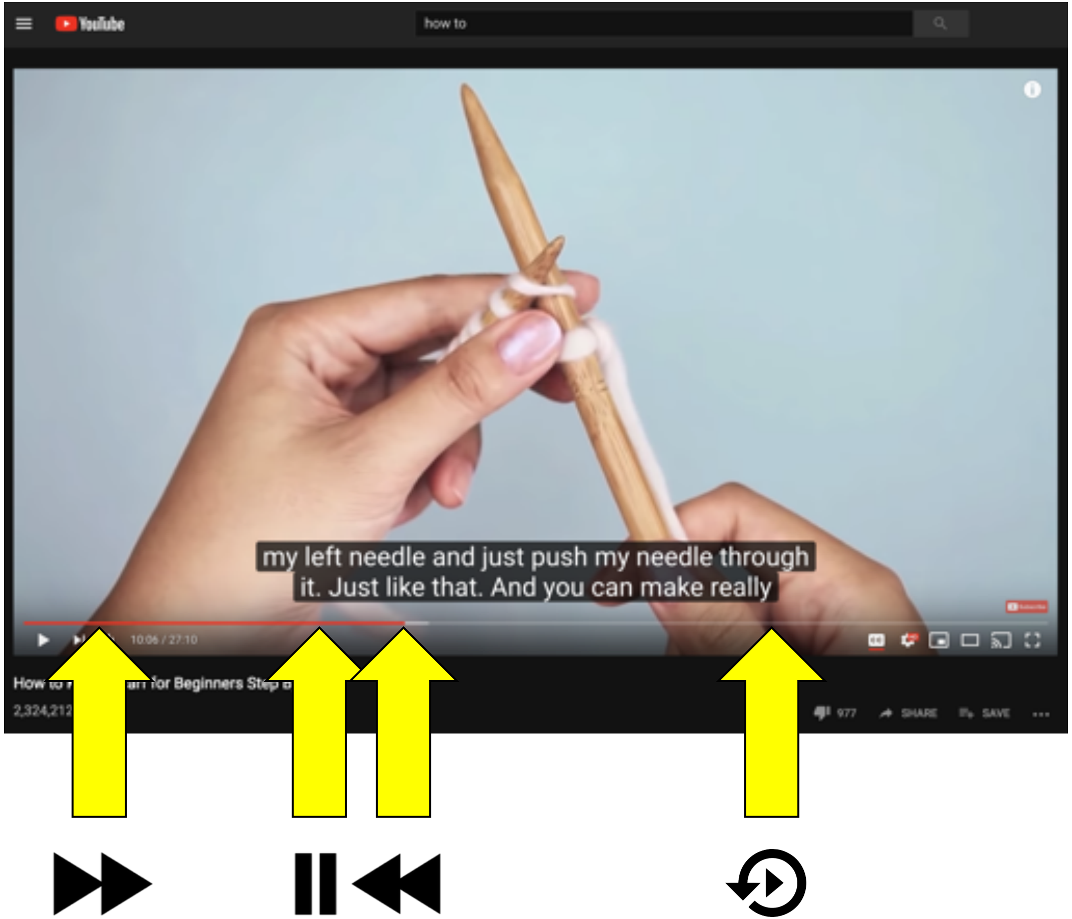
2. How are they realized with remote-control like voice interactions?

3. **What would an ideal voice interface be like for consuming How-to videos?**

1. **What are the navigational needs for how-to videos?**

2. How are they realized with remote-control like voice interactions?

3. How would an ideal voice interface be like for consuming how-to videos?



my left needle and just push my needle through
it. Just like that. And you can make really

Think-aloud Study

10

participants

10

participants

Pauses **Jumps**

Pauses (95)

Pace Control Pause (78): "I need more time"

Pauses (95)

Pace Control Pause (78): “I need more time”

Content Alignment Pause (9): “am I doing it correctly?”

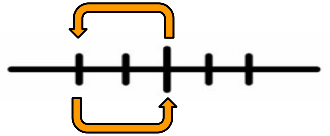
Pauses (95)

Pace Control Pause (78): “I need more time”

Content Alignment Pause (9): “am I doing it correctly?”

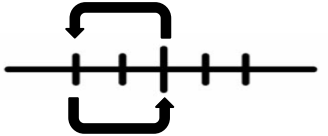
Video Control Pause (8): “I want something else”

Jumps (43)

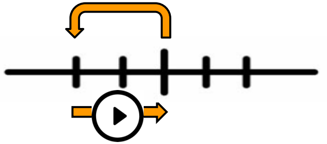


Reference Jump (5):
“I need to check something”

Jumps (43)

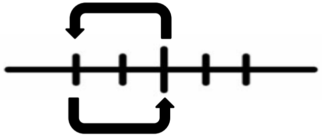


Reference Jump (5):
“I need to check something”



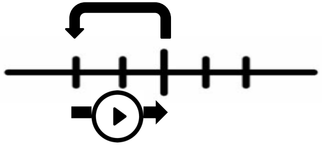
Replay Jump (24):
“I want to see something again”

Jumps (43)



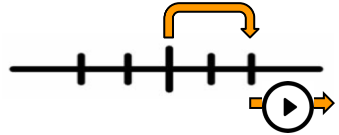
Reference Jump (5):

“I need to check something”



Replay Jump (24):

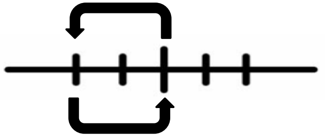
“I want to see something again”



Skip Jump (10):

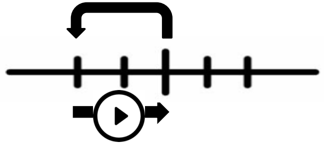
“I’m not interested in this part”

Jumps (43)



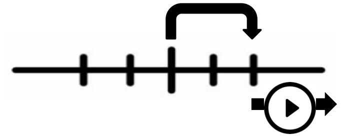
Reference Jump (5):

“I need to check something”



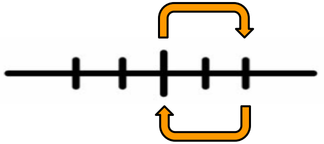
Replay Jump (24):

“I want to see something again”



Skip Jump (10):

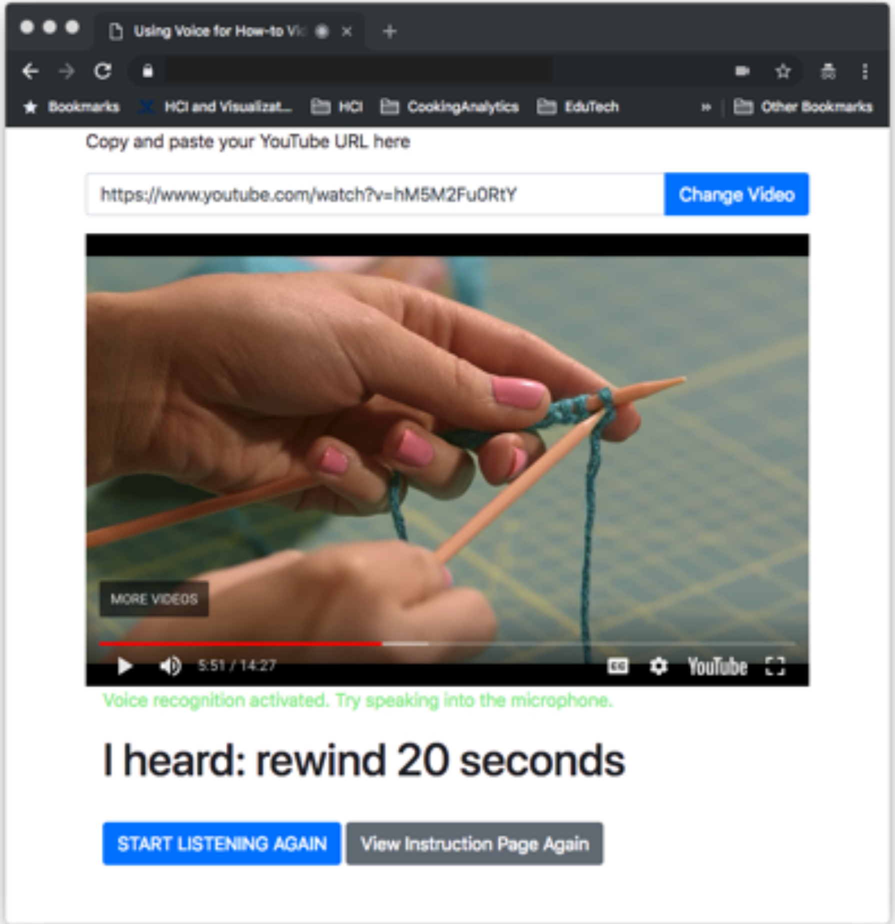
“I’m not interested in this part”



Peek Jump (4):

“I want to see what’s coming up”

1. What are the navigational needs for how-to videos?
2. **How are the navigational needs realized with remote-control like voice interactions?**
3. How would an ideal voice interface be like for consuming how-to videos?

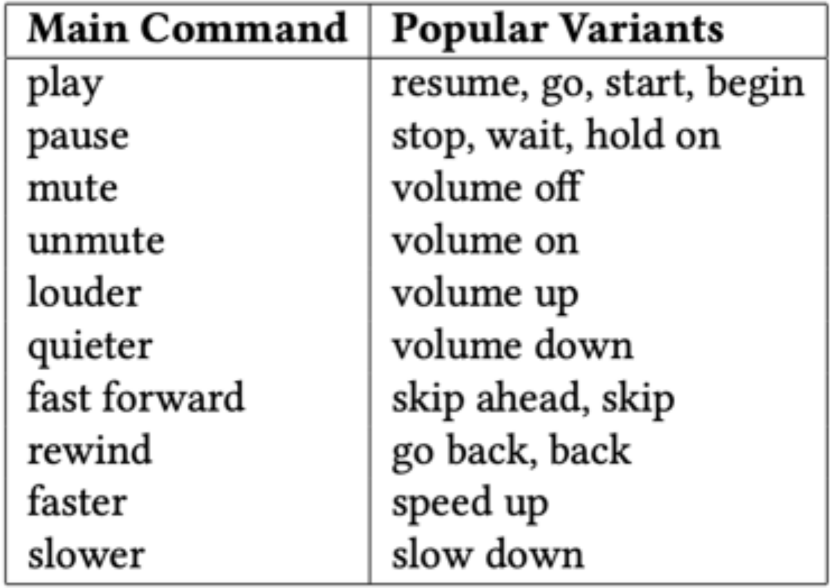


Copy and paste your YouTube URL here

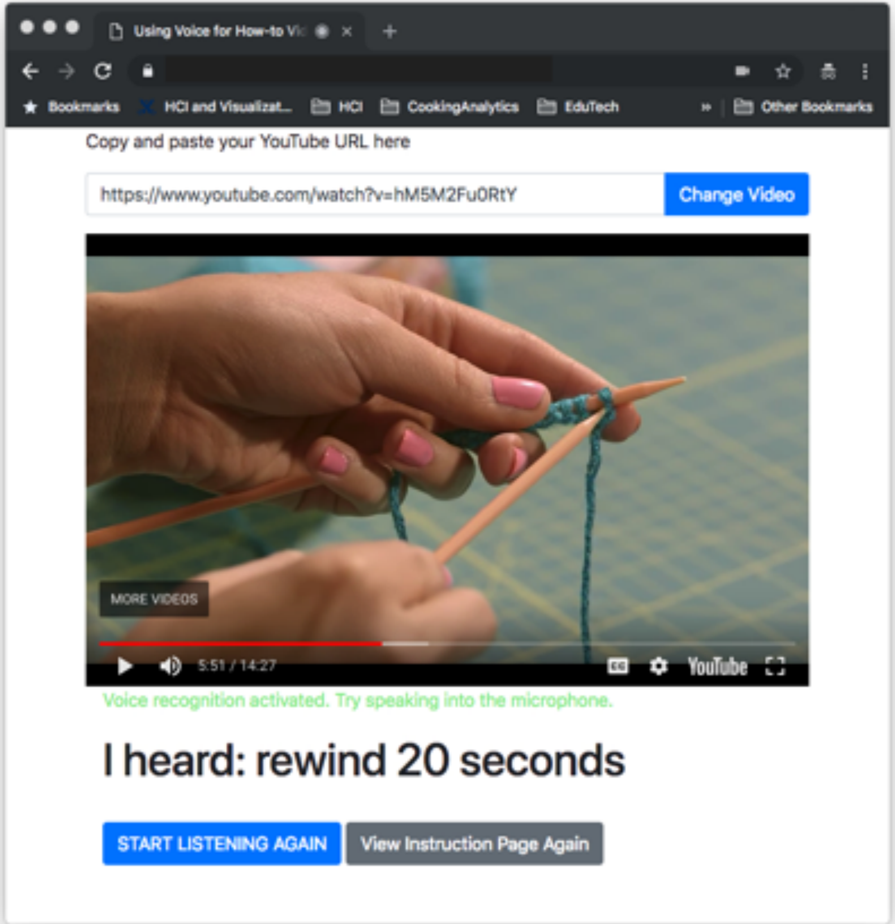
https://www.youtube.com/watch?v=hM5M2Fu0RtY

Voice recognition activated. Try speaking into the microphone.

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

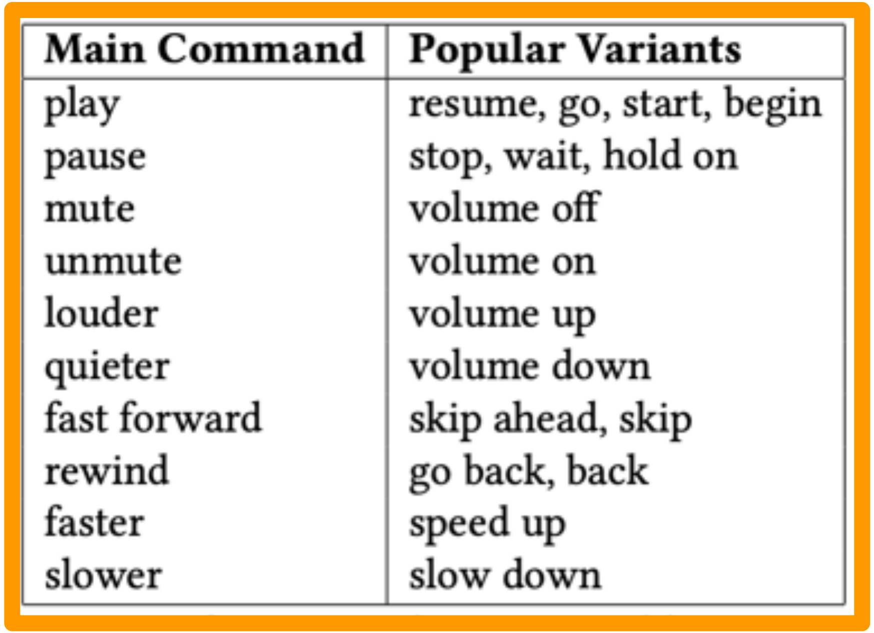


Copy and paste your YouTube URL here

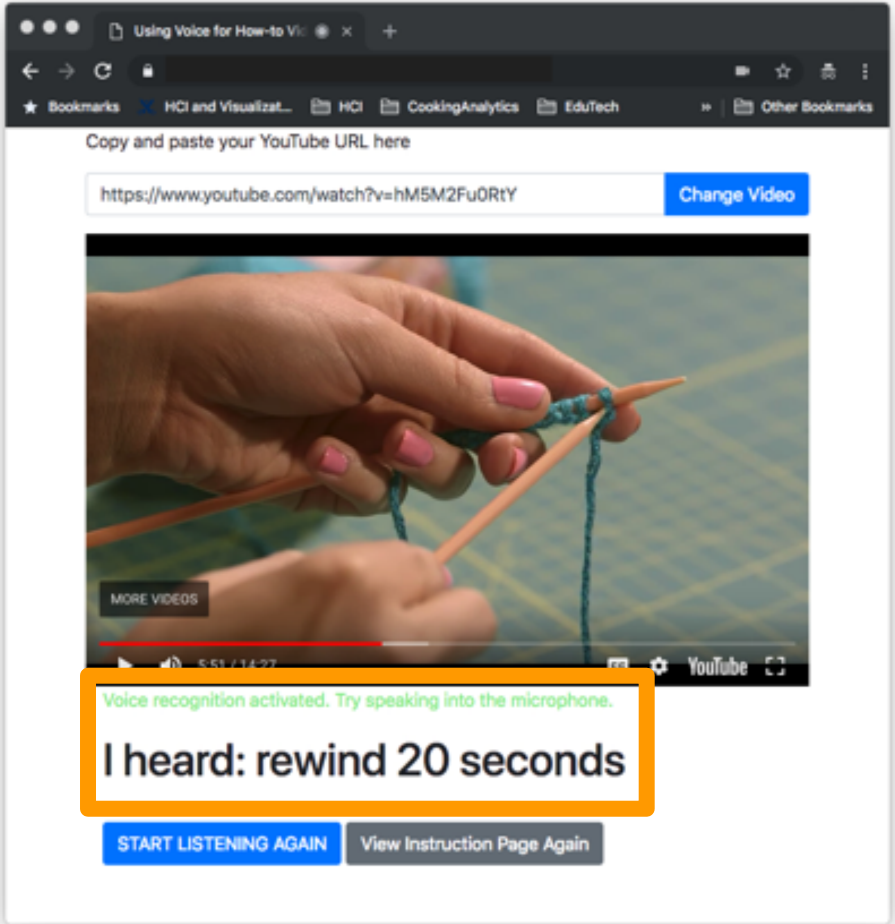
https://www.youtube.com/watch?v=hM5M2Fu0RtY

Voice recognition activated. Try speaking into the microphone.

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

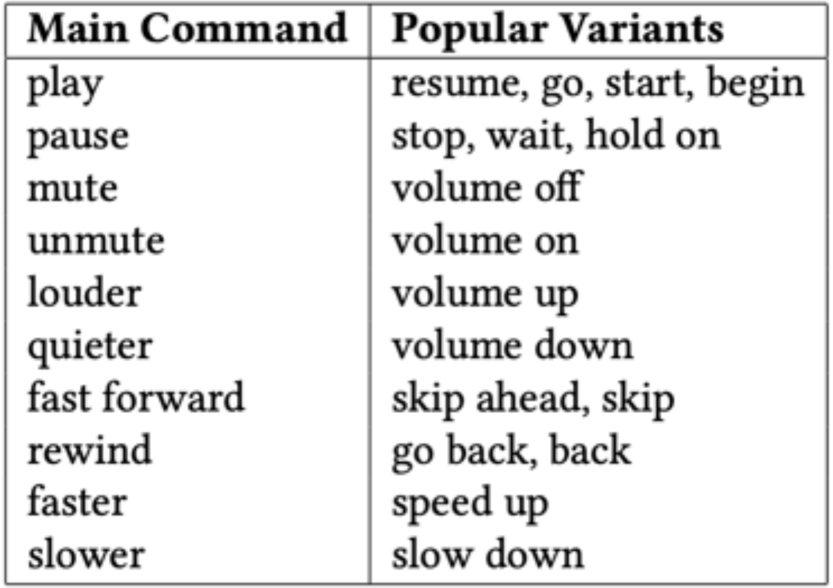


Copy and paste your YouTube URL here

https://www.youtube.com/watch?v=hM5M2Fu0RtY

Voice recognition activated. Try speaking into the microphone.

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

Think-aloud Study

Available commands were not available in advance

7 participants

4 male, 3 female

average age: 35

4 participants

all female

average age: 26

5 participants

all female

average age: 31

Pauses **Jumps**

“Pause” and “Stop” are used differently

Pace Control Pause:
“I need more time”

Content Alignment Pause :
“am I doing it correctly?”

“Pause” (24/43)
“Pause Video” (10/12)

Pace Control Pause:
“I need more time”

Content Alignment Pause :
“am I doing it correctly?”

“Pause” (24/43)
“Pause Video” (10/12)

Video Control Pause:
“I want something else”

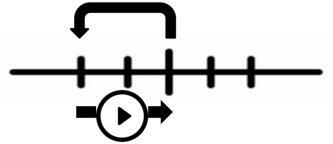
“Stop” (18/25)

STOP
=
URGENT

“Go back by a little”, “Go back by a bit”, “Stop!!”
“Stop”, “I don’t know what’s going on”

Contextual and *explicit* time references are used differently

Replay Jumps

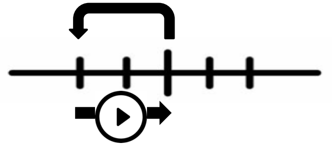


Contextual time reference:

“let me see that again”

“go back just a bit”

Replay Jumps

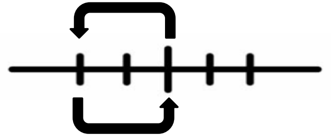


Contextual time reference:

“let me see that again”

“go back just a bit”

Reference Jumps



Explicit time reference:

“go to 2 minute mark”

“go back 30 seconds”

Replay Jumps

Contextual time reference:

“let me see that again”

“go back just a bit”

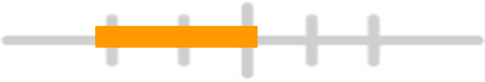
Reference Jumps

Explicit time reference:

“go to 2 minute mark”

“go back 30 seconds”

Replay Jumps

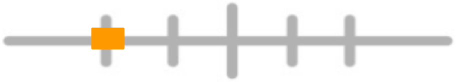


Contextual time reference:

“let me see that again”

“go back just a bit”

Reference Jumps



Explicit time reference:

“go to 2 minute mark”

“go back 30 seconds”

“Play”

“Pause”

“Skip 20 seconds”

Is this the best we can do?

1. What are the navigational needs for how-to videos?
2. How are the navigational needs realized with remote-control like voice interactions?
3. **What would an ideal voice interface be like for consuming how-to videos?**

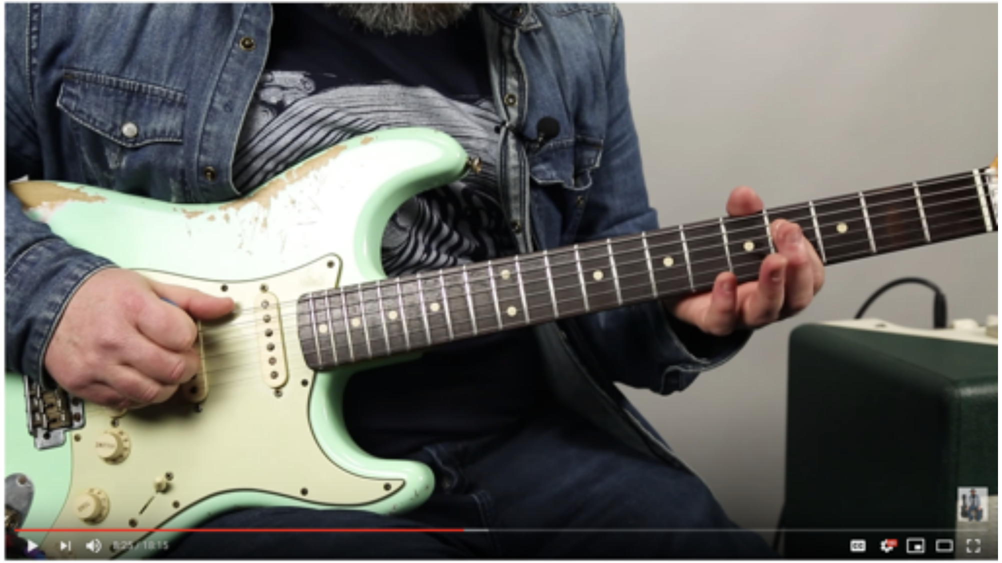
Wizard of Oz Experiment

Participant
Wizard

6 participants (3 male, 3 female) - 5 novice, 1 expert knitters

Challenges

Additional “Stop” before Further Navigation



“Go back 20 seconds” vs. ↖

Uncertainty from Unseen Content

How to “scrub” or “click-guess” in VUI?

Three Design Recommendations

Design Recommendations

Support Conversational Strategies

Support Iterative Refinements of Commands

Support Interactions with User Context

Conversational Strategies

conversational intents

“Can I see it again, 10 seconds before?”

“Can I see the last three knit?”

“Can you move it back to when she shows how it looks like from the back?”

Conversational Strategies

conversational intents

“Can I see it again, 10 seconds before?”

“Can I see the last three knit?”

“Can you move it back to when she shows how it looks like from the back?”

sequence expansion

(“rewind 30 seconds until 3 minutes”, “again”)

(“slow it down to .5 and play from 4 minutes”, “okay, from 3:55”)

Conversational Strategies

conversational intents

“Can I see it again, 10 seconds before?”

“Can I see the last three knit?”

“Can you move it back to when she shows how it looks like from the back?”

sequence expansion

(“rewind 30 seconds until 3 minutes”, “again”)

(“slow it down to .5 and play from 4 minutes”, “okay, from 3:55”)

command queues

“could we change the speed to like 1.25? I want to slow it back down when she actually starts the tutorial”

“I want to sequence commands, set rules like if there is a second needle, slow it down”

Design Recommendations

Support Conversational Strategies

Support Iterative Refinements of Commands

Support Interactions with User Context

Iterative Refinements of Commands

Users WANT TO reference to objects, actions, what speakers have said

“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)”*

“can you repeat that again? How she did multiples of four, the ***part where she said*** *multiples of four”.*

Design Recommendations

Support Conversational Strategies

Support Iterative Refinements of Commands

Support Interactions with User Context

Support In-depth Examination of Tutorial Content

Augmenting User Memory

1. Keep a pointer to the origin and provide “comeback” to the point
2. Update interval of jumps

Provide Visual Feedback

1) Thumbnails
2) Instruction milestones
3) Key frames
4) Frames with user-specified keywords

VOICE INTERACTION

USER

HOW-TO VIDEOS

VOICE INTERACTION

USER

Pace Control Pause

Content Alignment Pause

Video Control Pause

Reference Jump

Replay Jump

Skip Jump

Peek Jump

HOW-TO VIDEOS

VOICE INTERACTION

USER

Pace Control Pause
Content Alignment Pause
Video Control Pause

Reference Jump
Replay Jump
Skip Jump
Peek Jump

HOW-TO VIDEOS

In-depth Content Examination

Memory Augmentation

Visual Feedback

VOICE INTERACTION

USER

Pace Control Pause

Content Alignment Pause

Video Control Pause

Reference Jump

Replay Jump

Skip Jump

Peek Jump

Support Conversational Strategies

Support Iterative Refinements of Commands

Support Interactions with User Context

HOW-TO VIDEOS

In-depth Content Examination

Memory Augmentation

Visual Feedback

Acknowledgements

KAIST

HCI @ KAIST

Adobe®

Stanford

Project website

minsukchang.com/proj-voicevideo

USER

Pace Control Pause
Content Alignment Pause
Video Control Pause

Reference Jump
Replay Jump
Skip Jump
Peek Jump

Support Conversational Strategies

Support Iterative Refinements of Commands

Support Interactions with User Context

HOW-TO VIDEOS

In-depth content examination

Memory augmentation

Visual Feedback

"67% of the millenials agreed they can find a YouTube video on anything they want to learn"

Think with Google: I want-to-do moments: From home to beauty, May 2015

“Of smartphone users, 91% turn to their devices for ideas while completing a task”

Think with Google: I want-to-do moments: From home to beauty, May 2015