UNIVERSITY OF COPENHAGEN

A thesis presented to the Faculty of Science in partial fulfillment of the requirements for the degree

Master of Science in IT & Cognition

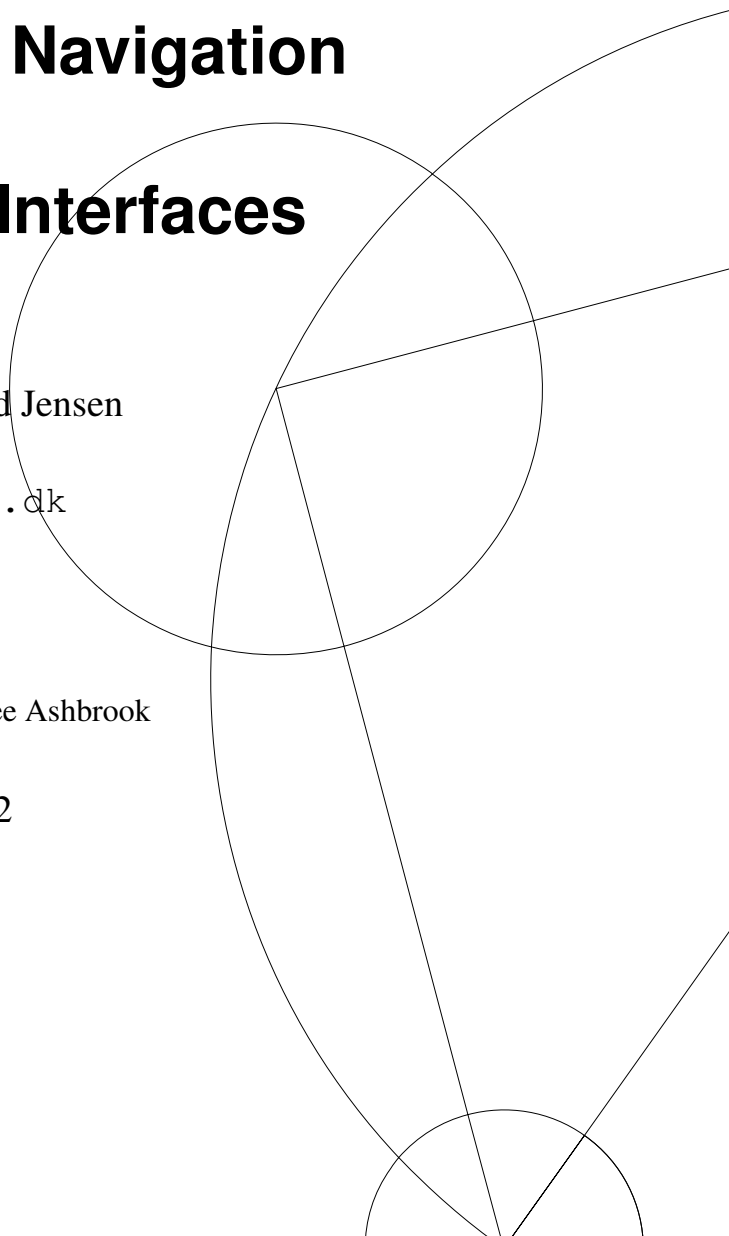
Investigating Audio Icons For Content-Based Navigation In Voice User Interfaces

Jonas Kjeldmand Jensen

jkj@di.ku.dk

Supervisor: Daniel Lee Ashbrook

June 2022



ABSTRACT

Voice user interfaces (like Alexa, Google Assistant, and Siri) are increasingly becoming more ubiquitous. They provide an alternative by controlling how-to videos of physical tasks via voice. While VUIs provide some help in the context of navigating how-to videos, no specific interaction features have successfully been designed and implemented for them. Instead, they directly inherit low-level remote-control features (e.g., pause, play, rewind, fast-forward) into voice commands.

An important question is whether this basic VCR-like control interface is suitable for navigating instructional videos via voice. If not, how do we design more fitting voice interfaces for navigating how-to videos?

In this study, we present using audio icons as an augmented content-based navigational feature for how-to videos controlled via voice. To test our ideas, we conducted a wizard-of-oz study on three conditions. Our results suggest that using augmented audio icons in VUIs for navigating how-to videos improved usability and reduced task completion time. Our analysis found that users' navigation strategy is largely affected by the mental model they employ to complete instructional tasks. Based on our findings, we recommend (1) supporting higher-level user intents, (2) allowing users to make bookmarks, (3) enabling global commands, (4) allowing for context recognition, and (5) supporting progressive command refinement.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

INTRODUCTION

1

INTRODUCTION

Our way of interacting with technology is rapidly evolving. New interaction modalities are increasingly challenging the mouse and keyboard's long reign as the main input devices. In recent years, Voice User Interfaces (VUIs) have seen significant improvement in the underlying technology and the use cases that are now being undertaken via voice. The technology is fast becoming ubiquitous, built into smartphones, car infotainment systems, and stand-alone devices in our homes. Our use and interaction with these devices seem only to increase. However, albeit recent advancements, the broad adoption of VUIs in the mainstream is still limited. The inherent invisible nature of VUIs can have the undesired effect of challenging the user's ability to discover the full breadth of capabilities (and limitations) that speech systems encompass (Klein 2015; Pearl 2017). When discoverability is challenged, learnability can be compromised (Furqan et al. 2017). Learnability is the users' ability to quickly learn how to use a new system for maximum productivity without having any previous training in using the interaction system (Corbett and Weber 2016). In overcoming the challenge of discoverability of voice user interfaces, different methods have been applied to improve the discovery features of commercially available VUIs. These methods include: tutorials, companion apps, or documentation user manuals, and some researchers have designed discovery tools to assist users in learning as they go (Corbett and Weber 2016; Huyghe et al. 2014; Kirschthaler et al. 2020; Zhong et al. 2014).

From cooking recipes to make-up tutorials to software programming how-to videos. People increasingly turn to online video instructions as guides for learning new skills or getting acquainted

with unfamiliar tasks (Chang 2019). With the increased interest in how-to online video resources for learning, the number of available instructional how-to videos has likewise followed this trend and seen significant growth. However, the increased availability of instructional material online does not automatically translate into better learning opportunities or better accessibility to these instructional materials.

Whenever users consume these videos, they are active viewers. They seek to interact with the video by controlling how it is presented to them. They pause, replay, and skip ahead and back while following along with the contents of the video. The need for navigating a how-to video can arise when the user, e.g., fails to follow the pace of the video, does not understand the instruction, or if the video includes content not relevant to the task (Chang, Huh, et al. 2021). A current often used strategy when interacting with how-to videos via mouse and keyboard is content-based navigation, e.g., when a user peeks the video's content by hovering or scrubbing the cursor over the timeline to see a small pre-cached thumbnail of what is happening at that specific time point. A central theme for this dissertation is to attempt to emulate this navigational interaction technique in a VUI context.

One recurring theme for many how-to videos is that they require manipulations of real-world physical objects. This means that users are expected to control the video via conventional direct manipulation keyboard techniques while attending to the physical task they are trying to accomplish with the help of the instructional video. This will lead to costly context switches whenever the user switches between video and real-world. Making for a clunky and awkward interaction experience demands a heavy cognitive load on the user while keeping track of both the task and screen progress. Current popular voice interfaces like Siri, Alexa, or Google Assistant seem to enable an interaction mode, which will allow the separation of the two activities — resulting in less cognitive stress and a better user experience.

While importing basic video navigation operations like pause, play, rewind, or fast-forward does provide an intuitive initial first step, naively importing these low-level remote-control commands as the crux for voice navigation does not seem to take advantage of what this new interface potentially

encompasses fully. Therefore, investigating how to better design voice interfaces for navigating how-to videos is pertinent.

2

RESEARCH QUESTIONS

2.1 PROBLEM STATEMENT

The conventional way people engage in watching how-to videos on Youtube will typically include navigating the footage via a mouse and keyboard as input. They will scrub through the video to have it match whatever they are trying to learn. Many how-to videos teach physical tasks that involve interaction with natural world objects. When following such videos, viewers need to have their hands free to execute the job and control the video while switching their vision between the task and the video. Navigating the video through traditional timeline-based interactions can be inefficient and time-consuming. This work explicitly explores how video navigation with VUIs can assist users in learning physical lessons from tutorial videos.

Voice-based user interfaces provide a potential alternative for controlling how we consume how-to videos for physical tasks. Current voice-based navigation systems (such as those found in smartphone accessibility features) support basic operations for video navigation such as pause, play, rewind, or fast-forward. While these systems do provide some help in the context of how-to videos, they are not purposely designed for this domain. Instead, they translate low-level remote-control operations (play, pause, etc.) into voice commands. An essential question of this study is whether this low-level remote-control-like interface is suitable for voice-driven video navigation interfaces for how-to videos and how we can design good voice interfaces for navigating how-to videos.

While instructional how-to videos are especially effective for conveying actions challenging to describe with words or static images alone, following their speed for physical tasks can be challenging. The linear nature of the video forces viewers to watch the procedure at the pace of the video (which might be either too fast or too slow) or scrub the timeline to jump ahead or go back to an earlier step. While existing voice systems can accommodate basic operations for video navigation (i.e., pause, play, next), they are not specifically designed with voice-first navigation in mind. This work will explore voice navigation solutions for how-to videos for physical tasks.

The challenge of designing a navigation system for how-to videos with voice and visual feedback has previously been explored in the literature (Chang, Huh, et al. 2021; Chang, Wang, et al. 2019). However, research on voice-driven step-by-step how-to navigation systems is limited. To bridge the gulf of execution for voice navigation systems in complex task scenarios, this study will investigate whether an interaction strategy of using audio icons (nonverbal audio signifiers) benefits users regarding usability and cognitive load when navigating how-to videos.

2.1.1 *Research Questions*

To investigate this area of research, the following research questions have been established to learn how we can make use of audio icons as a mode to improve voice navigation with voice user interfaces for how-to videos. With these questions we seek to investigate the potential benefits of using sounds as signifiers to enhance the VUI experience in complex task scenarios over existing practices.

- Are audio icons able to assist in teaching users to navigate complex voice-only tasks better?
- What are the challenges and opportunities for designing voice navigation interaction for voice-only step-by-step how-to guides?

.

Part II

BACKGROUND

3

BACKGROUND

This section delineates all the concepts necessary to read this study. First, the section tours the current challenges existing in VUI research. Next, we elaborate on the concept of discoverability and how it relates to VUI research. Then we focus on auditory icons as modes for communicating information to the user.

3.1 NAVIGATING INTERFACES USING VOICE

Previous studies of how people interact and navigate how-to videos via voice show that users tend to make more specific references to past events, as opposed to how users have more difficulty describing later, unseen parts of the video (Chang, Wang, et al. 2019). In a conventional keyboard and cursor user interface, clicking and scrubbing over the video timeline are often times the solution users opt for in order to get a quick glimpse into the future parts of the tutorial video (Matejka et al. 2013).

For this study, we consider design aspects inherent to audio interfaces, like sound's invisible and asymmetric nature. In 2016, (Corbett and Weber 2016) discussed the challenges of making VUI features adequately discoverable. Their conclusion back then was that voice interfaces have always struggled with learnability and discoverability issues, even after introducing the new generation of voice systems like Siri, Alexa, and Cortana. They further alluded to how users will often experience inconsistency in their user experiences across devices and form incorrect mental models of the VUIs

Table 1: Challenges in VUI (Ma and A. Liu 2020).

Challenge Categories	Detail
Technical	Speech synthesis quality
	Speech recognition performance
	Flexibility and accuracy
Audio Inherent	One-dimensionality
	Transience
	Invisibility
	Asymmetry

concerning what they can and cannot do. Radlinski and Craswell's VUI design heuristics delineate how users expect a voice-first system to be quite advanced, whereas, in reality, their current capabilities are quite simple and rudimentary (Radlinski and Craswell 2017).

3.2 DISCOVERABILITY OF VOICE USER INTERFACES

One of the most complex design challenges of VUIs is teaching users "What to say." This holds especially true for technologies like voice, as it inherently does not have any visual cues on how to use the system (Abdolrahmani et al. 2018). To do this, they must first discover and learn the system's capabilities and limitations (Corbett and Weber 2016; Kirschthaler et al. 2020). Discoverability was first described as a fundamental usability issue for speech technologies in the mid-90s (Yankelovich et al. 1995; Yankelovich 1996). The difficulties described range from users struggling to discover and learn commands, forgetting utterances or using them at incorrect points during the interaction, especially during the initial interactions (Kirschthaler et al. 2020). Analogously, speech-only interfaces deal with similar discoverability issues as a command-line interface (Harris 2005). The functionalities are hidden, and the boundaries for what actions can and cannot be performed are invisible (Yankelovich

et al. 1995). In extension, applying conventional design guidelines established from GUIs to designing VUIs is not feasible, as there are no visual affordances to indicate the options available to the VUI user.

Discoverability describes a user's ability to find and execute an interface's features and commands (Norman 2002). Suppose a product possesses a high degree of discoverability. In that case, it will display more-ease-of use for its intended end-users, as they clearly understand which actions and commands are possible to carry out by using the product. Voice user interfaces have long been scrutinized for their underlying lack of techniques to assist users in discovering and learning a voice-first system's functionalities and attributes (Corbett and Weber 2016). If the user can navigate the VUI effortlessly to get what they want, know how to start over, or get the correct help when needed, they will feel comfortable with the whole interaction. The outcome is a greater willingness to explore and try new requests, thus heightening their chance of reaching their goals (Norman 1998). The relative lack of discoverability and learnability aspects of VUIs might very well be due to the nature of the medium, as it is more challenging to provide the user with hints and cues compared to what is possible via a graphical user interface. As a result, the discoverability shortcomings have even been accused of being the main driver for the absence of wide adoption of voice assistants (Bentley et al. 2018; Kirschthaler et al. 2020). Users do not know what to say. One example of this perceived lack of transparency is the difficulty users experience when trying to discover what voice inputs the voice assistant supports. This challenge becomes even more pervasive when it is the first time a user interacts with an unacquainted VUI (Beneteau et al. 2020; Bentley et al. 2018; Corbett and Weber 2016). A Consequence of the lack of discoverability is a poor user experience due to an incomplete mental model of the capabilities of the VUI.

This ultimately leads to usability issues and the potential abandonment of the technology altogether (C. Myers et al. 2018). A further implication is that VUI users are coerced into staying with the same specific types of uses, as they are unaware of what other actions the system can perform (Bentley et al. 2018). This often limits use to relatively simple tasks like checking the weather or setting

a timer. Ideally, VUIs as a different mode of interaction from GUIs should encompass new and yet-to-be-conceived possibilities. Types of uses are often recited to include hands-free and eye-free uses as times where the technology is proving most useful. This resulting lack of understanding of voice systems' capabilities suggests a flawed conceptual model, which again restricts the development of current voice assistants, as many users are still using VUIs to carry out some of the tasks they are already able to get done quickly with GUIs, such as setting the alarm, search online, listen to music, etc. (Furqan et al. 2017). To ensure successful interaction, users of VUI technology must learn the commands that are available to them when using the system (i.e., intents) and what to say to execute these commands (i.e., utterances) (C. M. Myers 2019).

3.3 MENTAL MODELS & COGNITIVE LOAD

Mental models are peoples' internal representations of how something work (Norman 1998). The concept of creating a working mental model for something inconspicuous like voice systems is the main challenge in designing a successful VUI. If discovery tools for VUIs are not intuitively recognizable, the user is forced to remember a palette of utterances to interact with a voice system properly. This can prove particularly cumbersome for users, as this will cause stress to the limited capacity of their short-term memory (Cho 2018). In turn, this can have the undesired effect of leaving the user feeling cognitively overwhelmed and making it hard to recall from memory all of the presented information, let alone undiscovered capabilities of the system (Trippas 2021). For voice interfaces, mental models account for expectations for what to say, who says what, when, and to whom in dialog, or how a voice interaction starts and ends. Further, the term also covers appropriate topics for VUIs and the words used while engaging with a voice assistant. The mental representations that people have of a given voice system affect how users will express themselves to the VUI and how they can interact with it, as well as their expectations of the abilities of the voice assistant (Lee et al. 2020). Good voice design entails actively reflecting on these expectations to take advantage

of them and tap into people's existing knowledge and expectations of the system. Users will have a better experience engaging with a system if their expectations are met (Thymé-Gobbel and Jankowski 2021b).

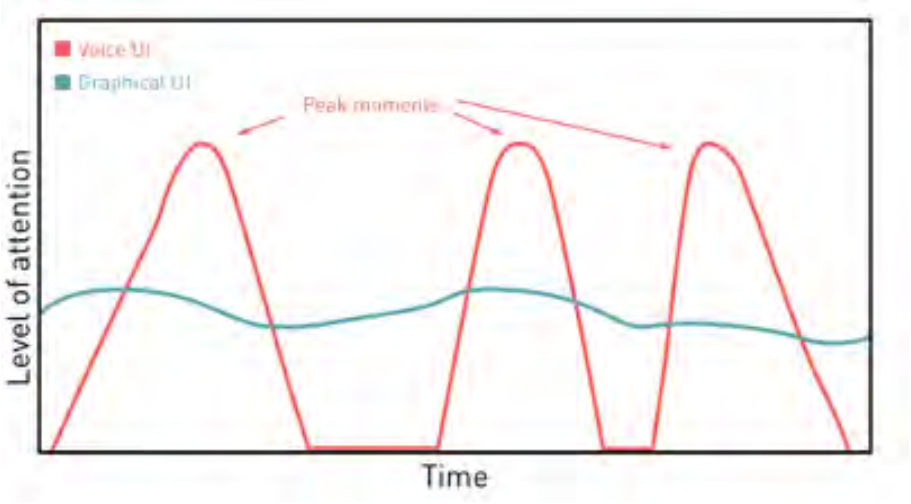


Figure 1: Peaks of Attention: Graphical versus Voice User Interfaces (Westerlund 2021).

How attention and users' mental load affect their cognitive abilities when interacting with VUIs are different from graphical user interfaces. Whereas a GUI is static in its information dissemination, voice is dependent on short bursts or peaks in attention from the user. If the user is inattentive at specific points during the interaction, they will miss crucial information, which might force them to re-play the entire system utterance.

3.4 PROS & CONS OF VOICE USER INTERFACES

Interaction via a purely voice-first system does carry with it its own particular set of issues. As language is so innate and natural to human beings, we have very high expectations of what a system that utilizes voice as its main modality can do. When these high expectations are not met, users get frustrated, as they are confronted with the fact that these "smart speakers" are not as smart as we perceive them to be (Schnelle-Walka and Lyardet 2006).

3.4.1 *Inherent Audio shortcomings*

A primary shortcoming with speech-based interfaces is the sequential nature of audio information presentation, as opposed to visual information that allows for the skimming of a document (Werner et al. 2015). This means that navigation of sound documents will need to adopt a very different approach than what is typically seen in visual navigation. One feature that would alleviate this limitation is the barge-in command (Cohen et al. 2004). Barge-ins enable faster task completion, as lengthy system speech can be abruptly cut off and thus improve the usability of the voice user interface.

3.4.2 *Voice User Interface Challenges*

Additional inherent challenges to designing for voice is (1) *asymmetry*, meaning that speaking is faster than typing in words and sentences on a keyboard. Still, it is slower to listen than it is to read. This incongruity stems from the fact that audio is linear (M. A. Walker et al. 1998). While this voice feature is a drawback in voice-first interactions, it can be a strength in situations where we can add a display to supplement the primary audio interface. Thus, we can use the display to deliver lengthy information while reserving audio for directly manipulating and controlling the interaction (A. Walker et al. 2001). This aspect of speech somewhat ties together with another innate trait of speech; that it is (2) *one-dimensional*. Where the eye is active, the ear is passive, i.e., the ear cannot browse through a large set of presented sorted information in the efficient way the eye can. Audio must wait until being presented, until its turn, and once delivered, it is gone again. This aspect gives rise to (3) the *transiency* of voice. Short-term memory controls voice. Long sentences containing much information will lead many users to forget most of the presented information, especially what was said in the beginning. The logical takeaway from this insight is to design for recognition over recall to ease

the already cumbersome task of keeping large amounts of spoken information in memory (Mittal 2020)]. One advantage of the transient nature of voice is that people can perceive visual and aural information in parallel, allowing the user to get information from both without switching contexts. Also, with this information in mind, voice designers can actively construct conversational pieces to hold the most crucial information towards the end of the sentence, rather than at the very beginning, to ensure the user will not forget the significant bits. Lastly, as we have touched on a few times thus far, speech is (4) *invisible*, meaning that it is difficult to properly indicate to the user what actions are performable and doable, leading back to the core issue of discoverability. An undesired effect of voice being invisible is that users can be left with the impression that they are not in control of the system (Schnelle-Walka 2010). The inherent audio properties of these four items are asymmetry, one-dimensionality, transiency, and invisibility. They are impossible to solve completely. However, it is nevertheless vital to be aware of them to find a workaround – or even try to design voice systems with these features in mind proactively, and maybe we might eventually be able to see them as an advantage rather than general limitations to what can be achieved with the technology (Zhang et al. 2020).

3.4.3 *Design Opportunities for Voice User Interfaces*

However, with this being said, the voice does hold some undisputed advantages over visual interfaces. As talking and conversation might be the most intuitive modalities for us to communicate, every user possesses some level of the prerequisite competencies of interaction with a voice assistant. All they have to do is talk. Speech is the ultimate shortcut – as it enables us to provide speed, simplicity, and ubiquity in that users can get answers to their questions instantaneously. It can save the user much time by jumping directly to where they want to be instead of maneuvering through an entire menu list structure. Voice interaction can be much quicker compared to having to perform multiple taps on a touch screen in a graphical interface. Conversations also allow users to multitask by assisting them in

busy, hands-free, eyes-free situations. An example of a scenario in which voice poses a superior mode of interaction could be as follows: You are late for your bus, running frantically down to the bus stop hoping to make it. The conventional GUI interaction for this scenario would be to enter an appropriate public transportation app, search for your specific bus and then decipher the upcoming timetable to learn if it is even worth your effort to run to catch the bus. Assuming we have a context-aware voice system available, the user would be able to ask the device while running to the bus stop, "When is my bus departing?" and the VUI would provide you with the exact information you are looking for. Obviously, for this scenario to work, the VUI must understand what "my bus" refers to. Nevertheless, instead of slowing down, give your undivided attention to making the search query; well over a minute might end up passing by before you have attained the desired information. In the VUI scenario, you would never have to slow down, or look away from your running path, and you would only be presented with the one specific piece of information you were interested in at that particular time.

3.5 AUDIO SIGNIFIERS FOR VOICE INTERACTION

In conventional graphical interfaces, designers have tried to bridge the gulf of execution by providing visible cues to give users a way of intuitively discovering what commands are possible by only glancing at the interface. In pure voice-first designs, the apparent absence of visual signals leaves two possible strategies for the user to understand what forms of interactions are possible with the system. They can guess or imagine what might be possible within the system or try to remember the possible commands by heart. Any of these strategies, however, will place severe strain on the users' cognitive load, thus making the interface more difficult to work with (Cho 2018; Cohen et al. 2004). However, in a similar way as seen in GUI design, voice designers can include signifiers to inform users of critical system information. These signifiers for VUIs will come in audio icons or earcons. Their primary function is to cue users about possible commands.

3.5.1 *Types of Audio Signifiers*

We generally distinguish between three different types of sound-based signifiers which can prompt user actions or inform users about possible commands. **Nonverbal sounds**, **earcons** or **auditory icons**, are distinctive sounds generated by the system. These are typically associated with specific states or actions. These can be used to elucidate some unknown or unused function. Or they can signal to the user where they are in the menu structure of the VUI (Pearl 2017).

3.5.1.1 *Non-Verbal Audio (NVA)*

Non-verbal audio or NVA is sound other than spoken words. NVAs come in a wide variety. Most notable are earcons, auditory icons, spearcons, and spindices (Ammari et al. 2019; Ilango et al. 2021). Good NVAs, like GUI icons, assist in decreasing the user's cognitive load (Pearl 2019). NVAs serve the same function as a string of confirmation words otherwise would symbolize a specific feature or action undertaken in the VUI environment. A drawback with NVAs is that they are not as universally recognized as explicit wordings (or visual symbols) are. For auditory icons or earcons to be meaningful, users must somehow associate them with their intended meaning. The best way to do that is to pick the corresponding sound to its intended use (Norman 2002). Meaning assignment through association is usually better than explicitly telling people what some introduced sound is supposed to mean. Because short, distinct sounds are quicker to process than words in a sentence. Spearcons deviate from the other techniques of utilizing sound in that spearcons are sped-up versions of explicit verbal signifiers, where the word of whatever is going on in the VUI is vocalized while being sped up (sometimes) almost to the point where what is said is no longer recognizable. Spearcons have consistently been shown to perform superior as opposed to auditory icons (Paikari and Hoek 2018). NVA landmarks are particularly useful for repeat users who learn to associate a sound with a meaning like success vs. failure. Typical uses of earcons include VUI activity status (listening starts, listening stops, processing, no result found, timing out, successful completion, standby), an alarm

Appropriate symbolism	Avoid sounds whose natural symbolism implies something opposite.
Sound quality	Make sure that sounds are of high audio quality and are pleasant to listen to multiple times.
Volume	NVA volume should match the VUI volume, as to avoid sounds be obstructing by being drastically louder or quieter.
Duration	Earcons need to be short and concise, while also having a distinct onset and end.
Theme	Define a family of sounds that provide consistency for the VUI persona and domain.

Table 2: Audio icon guidelines. We consider these in the making of the icons used in this study (Thymé-Gobbel and Jankowski 2021a).

or notification, helpful hint or more information, navigation (backward, forward, home/main menu, subarea of an app), and each item in a list (Hoffmann et al. 2019).

In order to ensure successful user interpretation of NVAs, the following guidelines help overcome common pitfall:

3.5.1.2 *Earcons*

Earcons are different from other NVA's. Earcons use abstract sounds, like chimes or instruments that will change, e.g., pitch or other sonic attributes, to signify some change. Like rising pitch sounds often predict success (successful action), while abrupt, 'harsh' sounds will indicate that something when wrong in the interaction. This means that earcons are further distanced from their actual action than auditory icons and are therefore more reliant on the user's ability to correctly understand the abstraction of the earcon sound to decipher its meaning. The success of earcons depends mainly on the user's ability to understand what meaning is being conveyed with the signal (Abbott 2002).

3.5.1.3 *Auditory Icons*

Auditory icons (audio icons) are similar to visual icons found in GUIs because they both attempt to communicate with users more efficiently by trying to do away with lengthy word explanations. They both strive to resemble the actual action they are meant to perform, i.e., a graphical trashcan icon or the sound of crumpled-up paper. Ideally, they should both connote the action of trashing something.

Auditory icons are similar to visual icons found in GUIs because they both attempt to communicate with users more efficiently by trying to do away with lengthy descriptions of the action they represent.

As seen with visual icons, universal understanding is far from the norm, and their intrinsic meaning becomes even vaguer with sound icons. As with graphical icons, these can be classified by their driving vehicle for meaning-making, either through resemblance, reference, or abstraction, depending on how it relates to the symbol it seeks to represent. As sound design within this realm is still relatively new, very few conventions have been established for why something might sound a specific way. Therefore, to the extent it is possible, making use of resemblance might ensure users' highest degree of recognition. Due to the limitations of earcons, they have shown efficacy primarily in narrow, repetitive contexts (for example, when serving as confirmations for frequent tasks) or as generic attention getters (Large et al. 2019).

3.5.2 *Auditory Icons for Discoverability*

In the words of Donald Norman (Norman 2002), there is no natural user interface. There is only agreed-upon convention. The icons appearing on a computer desktop are not natural, but we have come to regard them as such with time. The same holds for sound cues. The scholarly interest in nonspeech audio cues, including auditory icons (Gaver 1989) and earcons (Blattner et al. 1989) goes back to the mid-80s when they were suggested as tools to improve TTS-only interfaces (B. N. Walker et al. 2013). Since the inception of using audio to ease learnability, different examples of nonverbal audio types has been suggested to now incorporate: *auditory icons*, *spearcons*, *spindex/spindices*, *auditory emoticons*, *auditory scollbars*, *musicons*, etc. (Csapó and Wersényi 2013; Yalla and B. N. Walker 2008). While a plethora of specific types of sound cues exists, we shall limit this section only to auditory icons (or audio icons) as they will be the method of investigation for this study.

Gaver 1989 define an audio icon as "a unique sound from an event, providing a powerful resource for information about a situation." audio icons or warning signals are widely used for various purposes

in digital systems to convey information about, e.g., the status of a system or device itself. These audio icons can be anything from small beeping sounds to full vocal messages. Their function is to convey information about an object, event, or situation. (Cabral and Remijn 2019).

3.5.3 *Voice Navigation*

As VUIs grow, the need for effective navigation commands to help users get around in an audio-only interaction space becomes increasingly important. Consistent voice-only navigation commands enable users to take full advantage of the voice's strengths, i.e., skip past multiple interaction sequences and back again, start over, pause, etc. Through the use of effective audio feedback, the VUI can quickly let the user know if they are on the right path or not (Corbett and Weber 2016). Searching or browsing by voice is generally considered to have greater navigational complexity than most other applications. This is because menus or categories present a wide array of services and information, which can be quite bulky to navigate solely by way of voice (Cohen et al. 2004; Pearl 2019). A common strategy to assist users in knowing where they are in a voice-first application is auditory landmarking, initially proposed by Jenkins in 1985 (Jenkins 1985).

3.6 AUDITORY LANDMARKS

Landmarks are auditory markers that ideally provide associated sounds denoting each particular service, action, or position within the voice application. Landmarks are the audio equivalent of product labels, menu labels, or product door signs (Norman 1998). They are short phrases or audio bits signifying to let the user know that they are currently interacting with a specific feature or are in a particular area of the VUI. Ideally, auditory landmarks should be brief and clear, like a verbal equivalent saying, "Canned food aisle" or "Main menu" (Thymé-Gobbel and Jankowski

2021d). Landmarks exist in the form of non-verbal cues, like short, distinct audio chimes, or direct verbal labels, i.e. spearcons or audio icons (B. N. Walker et al. 2013). Landmarks work as explicit confirmation, and they assist the user in quickly understanding the setup of complex VUIs and in establishing expectations and the right connotations to the upcoming interaction dialog of the VUI's different functionalities. Landmarks provide reassurance that the users are on the correct path. Their effectiveness can be noticed through the way user behavior follows a landmark, i.e., less hesitation and uncertainty in user utterances, faster barge-ins, and quicker requests to start over or go back. Essentially, speeding up the user experience along with fewer mistakes made (Murad et al. 2018). Landmarks are particularly useful for repeat users as they learn to associate the specific meaning of a sound with success or failure.

4

RELATED WORK

In this section, we start by presenting studies on the introduction of current VUI systems, their uses, and the challenges the technology has met. Next, we look at previous research on how users interact with how-to videos as a particular media genre. Then, we examine voice interfaces used for video navigation and prior attempts to rethink remote-like features in GUI settings. Lastly, we present the most recent literature in VUIs supporting content-based navigation for how-to videos.

4.1 SMARTPHONE-BASED ASSISTANTS AND HOME-BASED ASSISTANTS

Although having seen scholarly interest earlier, voice user interfaces only really became a household name with their introduction in mobile applications with the likes of Siri, Google Assistant, and Microsoft Cortana. As their usage grew, so did the scientific interest in learning how people make use of the newly introduced technology (Abdolrahmani et al. 2018; Bentley et al. 2018; Corbett and Weber 2016; Druga et al. 2019; Furqan et al. 2017; Kirschthaler et al. 2020; Purington et al. 2017). The results coming out of these studies suggest that the primary uses of voice technology are mainly focused on carrying out simple tasks like; checking the weather, playing music, setting reminders, setting alarms, etc. (Beneteau et al. 2020; X. Liu et al. 2021). Problematically, the technology is often shown to be too inaccurate. As a result, users often have to adjust their language to be understood by the voice assistant or even outright abandon the technology based on their initial bad experiences.

This can have the undesired effect of confusing many users as to what they can do and what features they can perform with their device (C. M. Myers 2019).

Luger and Sellen 2016 explored the use of smartphone-based voice assistants and comically found them to be "really bad." Their study showed that users expect a broader range of features and functionalities than what current VUIs can provide. Even for questions, the users deemed simple would fall out of scope for what was possible with these new voice assistants — resulting in frustration and rejection of the technology after the first initial encounters. They reported that while 98% of iPhone users had tried out Siri at some point, 70% of these people say they do not make use of Siri regularly. For future design recommendations, they propose a focus on honestly revealing a system's capabilities and limitations to assist users in establishing accurate mental models of the conversational agent's true level of intelligence. Such perceptions will inadvertently affect how the system presents itself and how it conveys information through its interaction with the user (Yang and Aurisicchio 2021).

When voice assistants started to enter the home, a different type of interaction would now be possible. The placement of these systems in, e.g., the kitchen or living room would more easily be able to assist the natural, hands-free environment that these spaces afford. Especially the introduction of Alexa, Google Home, and Apple HomePod honed this new era of personal voice assistants. To researchers, the introduction of smart assistants in the home now opened new avenues for studying interactions. Curiously learning more about VUI use in a home setting would either teach us that it resembles or differentiates itself from the use patterns seen with users of portable voice assistants like Siri on iPhones. (Beneteau et al. 2020) explored how learning and discovering the functionalities of voice interface technology is undertaken in the family home. They found that families decreased their exploration and breadth of use of their Alexa Echo Dots over the six-week period the study went on for. The study participants indicated that they learned of new features during the study through trial-and-error and from people they knew. They tried to learn the functions of the Echo Dot through Echo Dot itself, assuming that the VUI was a natural point of departure to learn about its features.

These endeavors were met with mixed success. The authors suggest that voice designers create home-based voice interfaces using the concept of a "near-peer," in which the VUI is a trustworthy learning partner that aids in increased discoverability of its functionality (Ostrowski et al. 2021).

4.2 INTERACTING WITH HOW-TO VIDEOS

For navigating MOOC videos, Smart Jump Chen et al. 2014 works by suggesting the best position for a jump-back on the timeline to minimize the amount of time wasted the user will have to spend on waiting to come to the desired timestamp due to inaccurate navigation controls. Their analysis of the navigation data showed that most skip-backs are because of 'bad' positions of a previously performed skip-back. In another study, the researchers tested how VUIs might be used for voice interaction in assisting tasks like image editing Laput et al. 2013 or enabling expert users of image editing software to be able to utter a command to minimize the number of shortcuts that the user will need to remember Kim et al. 2019. SceneSkim (Pavel, Goldman, et al. 2015) showed how to parse video transcripts and summaries to allow for expressive video snippet search at several levels of granularity. VideoDigests (Pavel, Reed, et al. 2014)

In our research, we add to this rich line of existing literature of expanded interaction possibilities with how-to videos, specifically focusing on how to augment voice interaction techniques for a wood crafting scenario.

4.3 VOICE INTERFACES, VIDEO NAVIGATION, AND DIRECT MANIPULATION

Currently, voice navigation commands primarily focus on emulating basic 'VHS-like' playback controls like play, pause, rewind, and skip-forward Behrooz et al. 2019. While these play on familiarity

for the user, they also limit the potential for expanding people's mental model of what is possible with voice assistants.

Previous work has looked at potential new interaction techniques suited for navigating videos beyond the scope of basing the interaction primarily on a timeline interface. One study (Dragicevic et al. 2008) concluded that directly manipulating the video content (via mouse dragging over the desired object) is a useful fit for visual search content tasks. Their study showed that letting the user drag a billiard ball along the screen with the cursor, which would correspond to a change in time, is an effective way of conjoining time and content in video navigation. Another example is Swift and Swifter, which helped video navigation by allowing the user to quickly get an overview of future content in the video by showing a small pre-cached thumbnail of the content at different points in the video (Matejka et al. 2012; Matejka et al. 2013). Improvements like these give viewers a straightforward preview of unseen video content in an easily navigable manner, enhancing the browsing experience by letting the user skip straight to the content they are specifically looking for. This method is common in all major video playback applications and websites. (Crockford and Agius 2006) Using basic VCR-remote controls (pause, play, rewind, skip-forward) improves and limits the user's browsing competencies. While these basic manipulation techniques create a starting point for voice video navigation, they were not originally intended to be performed via voice. Moreover, importing navigating controls in this manner can potentially show the undesired effect of just imitating its button counterparts instead of exploring the potential benefits that might exist when using voice controls for video navigation (Srinivasan et al. 2019).

4.3.1 *Time-Based Versus Content-Based Navigation*

Tuncer et al. 2020 presents an in-depth study of how users use online instructional videos to achieve practical tasks, documenting the job of interlacing video and task. To balance the video and their activities, to manipulate artifacts that are part of the task, and to do the task itself, users need to pause

and resume videos repeatedly. They conclude that users work on two separate temporalities, one being the video timeline and the other their timeline for carrying out the task. The visual and physical timeline does not run in perfect unison (which they rarely do). The authors point to the pause button as the key navigation feature for threading together the instructional video and the task being carried out. (Zhi et al. 2018) introduces *VisPod*, an audio player that enables audio content browsing, topic-based, and keyword-based audio-visual navigation. The application works by enabling the user to navigate to specific time points in the audio file by saying words that appear in the audio based on a transcript of the audio content. While VisPod enables augmented audio content navigation capabilities to users, they still incorporate the conventional time-based navigation options. Considering referencing behavior, in Yarmand et al. 2019, the authors examine different video referencing behaviors and techniques in the endeavor to map what and in which situations other content referencing methods are used in YouTube's video commenting section. They found that people are generally more inclined to use temporal location points or time-stamps as their anchor point for reference, more so than referring to the video's actual content. While their study does not directly relate to what we are doing in this study, it is still fascinating to get insight into how people conceptually think about referencing and to what extent they address previously consumed video content. From these studies, the implication seems to be that temporal referencing is likely to receive more use than content-based referencing for different reasons. One is that users are more accustomed to using time as their point of reference. Another reason might be the apparent lack of current interaction techniques supporting proper content-based video interaction. One thing to note is that besides the lack of content-based navigation literature, the research on voice content-based interaction is even sparser. As of now, Therefore, in this research, we continue the work of augmenting voice interactions with a particular focus on instructional how-to videos, specifically focusing on how to integrate content-based voice interaction for navigating a natural voice interaction scenario.

4.3.2 *Voice Interfaces Supporting Content-Based Navigation for How-To Videos*

Recently, Chang, Wang, et al. 2019 studied the behavior of how users currently follow how-to videos when only having remote-control like voice commands available and how users visualize a complete voice navigation system by use of a wizard-of-oz study. The study's most striking and informative result is that users idealize conversational strategies they use in daily conversation with other people, which should be reflected in the design. They identified four key challenges in supporting content-based referencing in VUIs for how-to videos: (1) users prefer short, keyword-based queries as speech commands, as opposed to having to say entire sentences; (2) users are unable to recall the precise vocabulary used in the video; (3) users are not able to consistently remember all available commands and make use of them at appropriate times; (4) more complex conversational strategies allows for more varied commands, as opposed to the reasonably stringent timeline interaction inputs. The added complexity of a more varied accepted vocabulary heightens the change for speech recognition errors and speech parsing delays.

A future iteration of their first study Chang, Huh, et al. 2021 presents a system that supports efficient content-based voice navigation through keyword-based queries. Addressing the limitations of their initial study, they argue that naively translating conventional timeline manipulation into VUIs will result in importing temporal referencing. An ensuing consequence of temporal referencing is a limited mental model of what the system affords, e.g., users are coerced into conceptualizing voice video navigation as if their voice was a remote or the inability to peek into unseen future content through referencing the contents of the video. Allowing for content-based voice navigation, the authors expand on the current interaction capabilities of voice systems by equipping users with a tool that allows them to make peeks into the content better.

In this work, we aim to develop upon the existing work that has been done in video navigation research by exploring novel techniques for assisting users in temporal and content-based referencing

methods for navigation into voice interaction while having instructional how-to video tutorials as our primary focus.

Part III

METHODOLOGY

5

METHOD

To investigate our hypotheses, we conducted a controlled experiment to determine if the proposed augmented navigation options significantly affect the usability when controlling video playback via voice. The participants' main task was to assemble a small wooden robot figurine with the help of a VUI and a video tutorial shown on a provided laptop screen. All materials and tools needed to complete the job were supplied and available at all times. The participants could progress and navigate the how-to video only by voice. The study consists of three conditions. Depending on the conditions the participants were assigned, they were given additional video navigation options.

In this study, we distinguish between two types of video navigation strategies; (1) *temporal referencing* and (2) *content-based referencing*. The difference between the two strategies is what acts as the anchor for navigation in the user's mental model. Temporal referencing is the most used, with people doing VHS-like controls like rewinding or fast-forwarding (Yarmand et al. 2019). For example, viewers use a typical GUI-video player to use temporal referencing by clicking the cursor at some place in the video timeline or clicking the keyboard to go back and forward. For voice user interfaces, a similar action can be performed by issuing a user-specified time command, like, e.g., "rewind 20 seconds" (Zhi et al. 2018).

Contrarily, for content-based referencing, the anchor for navigation is the content of the video in question. Likewise, the GUI example for content-based referencing is when users examine the small pre-cached thumbnail that appears when the cursor hovers over the video timeline. To make

something equal to what is achievable in a GUI interface for voice interfaces, users must be able to directly issue voice commands that describe the content of the particular scene in question, like; ”go to the scene where [...] is happening”. However, current speech technology has yet to design a voice system that effectively supports complex commands based on content for any video.

5.1 PILOT STUDIES

Prior to arriving at our final study design, we conducted pilot studies with four recruited participants before arriving at the final version of the experiment. This approach was chosen to understand user behavior and to test our preconceived notions of how our participants might react to the experiment. This was partly a result of the existing literature claiming that users might not feel entirely comfortable using voice interfaces (Cheng et al. 2018). In the initial trial runs, we allowed the participants to interact with the video material on the laptop via the keyboard. We did this to test whether we would observe significant differences between keyboard and voice input – given that only one or the other is available. Interestingly, the two participants who underwent this pilot study performed 2.7 times more total command inputs (average is 81 commands) than their voice-only counterparts. This intrigued our interest in better understanding why we might see more input when users are allowed to use the keyboard versus their voice. Even though this finding is fascinating, we decided to depart from the assumption that voice-only interaction is an established mode of interaction. Moreover, in the formative study, all participants informed us that they were both familiar with voice technology and had tried it out for themselves at least once. In the subsequent trials, we focused on having the speech technology work satisfactorily to uphold the illusion that they were communicating with a machine. Also, we spent considerable time choosing the sounds so that they would not be too intrusive or steer away focus from the primary task.

5.2 THE DESIGN'S AUGMENTED INTERACTION FEATURES

The apparatus used for the study is comprised of a Wizard-of-Oz setup, emulating a voice-driven controller for a video player. The apparatus supports both temporal and content-based referencing to varying extents depending on which condition is tested. In total, the video player accepts six different types of input with some being deactivated for certain study conditions. These six input commands are (1) *play*, (2) *pause*, (3) *rewind*, (4) *fast-forward*, (5) *user-specified time referencing* (e.g. "go back 13 seconds"), and (6) *content-based sounds* appearing at different important points in the video. The rewind and fast-forward commands are both set to 10 seconds by default. The participants were notified of this. Moreover, the proposed design also supports command queues, i.e. saying multiple commands in a single utterance (e.g. "go back 45 seconds and pause"). For content-based referencing, users can make use of the audio icons appearing at content-specific places in the video for navigation by saying either the name of the specific sound (e.g. "Skip ahead to next duck sound"), or they could also simply refer to the last played sound, if they forgot the sequence of the audio icons (e.g. "go back to last sound").

6

STUDY DESIGN

We used a between-participant study design with the different augmented navigational voice features as an independent measure across three conditions, with one independent variable being tested for in each condition. The independent variables are: (1) *baseline remote-like voice controls*; (2) *audio icons positioned at action points in the video*; and (3) *audio icons placed at 30-second intervals*.

(a) Tool desk

(b) wood-working desk

Figure 2: Images of the two experiment tables used in the study. They are placed immediately adjacent to each other.

Each participant was assigned to partake in one study condition lasting between 17-46 minutes. The participants doing conditions 2 and 3 also participated in a practice trial before commencing the actual experiment. The practice trial is not considered for the total task completion time.

6.1 STUDY PROCEDURE

We recruited the participants by reaching out to people by hanging up posters at various spots on campus to get people to contact us back, as well as an online community call for participation. When people had made contact, we would send them a Calendly link containing general information about the study and a short pre-study screening questionnaire. Lastly, they could book a time to come to the lab for the in-person experiment. The only criteria for invitation were that the participants had previously been exposed to video tutorials, were familiar with using shop tools, and were acquainted with voice user interfaces.

The study setup was made using a between-subjects study design, where participants would only interact with one of the three conditions to complete the task of creating a small wooden figurine. They watched a typical instructional video tutorial for this task and could control the video via different types of referencing styles for voice input.

To provide a preview of the system, we gave the participants of conditions 2 and 3 a brief training session, which comprised of watching and controlling a short video via voice input while simultaneously answering six questions about the video's content. The purpose of the training video was to familiarize users with using the augmented voice commands as controls for a video player. As content-based referencing is relatively new to many, the training video emphasized this aspect to ensure people correctly understood the concept in a VUI context. Upon completing the training session, the leading researcher ran a short debriefing session with the participant to clarify any confusion regarding the system's abilities. If the participant could express the system's workings themselves satisfactorily, the study session moved on to the main experiment. The reason why the participants in the baseline study did not take part in the priming training session is because we assume that users are familiar with voice navigation in general.

Figure 3: Illustration of experiment stages. The experiment consisted of three main stages: (1) Pre-experiment briefing (and an additional priming session for condition 2 & 3), (2) voice interaction task and (3) a post-experiment interview. At stage (1) participants were informed about what will happen during the experiment (and was given a small tutorial task to show them how the system worked if participating in condition 2 & 3). Stage (2) consisted of the main how-to video voice control experiment; upon finishing the session the participant was asked asked to NASA-TLX and SUS questionnaires. Finally, during the last stage (3), participants were invited to take part in a post-experiment interview and were afterwards thanked for their participation.

To properly evaluate the usability and perceived workload of each strategy, we had the participants fill out a System Usability Scale (SUS) and NASA's Task Load Index test (NASA-TLX). At the end of each study, we conducted short semi-structured interviews to collect qualitative feedback for a deeper understanding of the actions of the participants.

6.2 EXPERIMENT TASK

Condition 1 serves as the baseline experiment with no additional features added to the task. This means that participants were given the standard functions you would typically find in a VUI-assisted video playback interface. These include legal TV remote-like navigation options.

The premise of condition 2 is to test how people respond when given the option to control the video via audio icons placed at points in the video where they are required to pay attention to gain the appropriate guidance to complete the task satisfactorily. Condition 3 differs in design by having the audio icons appear at 30-second intervals instead of having them play at points of interest in the

video. Therefore, the audio icons in condition 3 are not directly tied to any specific content in the video. Instead, they work as time-based referenceable anchor points.

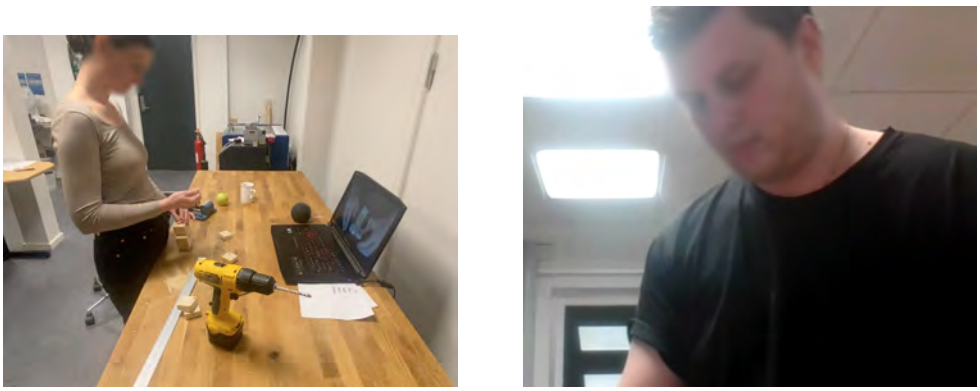
With these two proposed augmented voice navigation variations, we seek to learn whether and how audio icons might improve the usability of voice navigation for how-to videos by investigating how users make use of these icons when they are either directly mapped to the visual content or when they are following a specific time interval.

	Description
Condition 1	The baseline condition; resembles current practices in voice navigation interaction. Available input types: (1) play, (2) pause, (3) rewind, (4) fast-forward, and (5) user-specified time skips.
Condition 2	Content-based audio icon condition: input types (1-5) are available. Audio icons appear at time points in the video where the tutorial requires the user to perform some specific action to progress toward the finished product. For example, an icon can be placed when certain content is shown. The user can then freely go back and forward in the video by referring to the audio icons, like ”go back to the last bark”. The video contains 19 content-based audio icons in total.
Condition 3	30-second interval audio icon condition; input types (1-5) are available. Audio icons play each 30 seconds, looping over the icons. Users can refer to the audio icon they heard last to quickly skip back. Metaphorically, the landmarks function as anchors which are spread out at 30-second intervals. The video contains 18 audio icons in total.

Table 3: Overview of the three different conditions.

The experiment is comprised of a setup stretching over two adjacent rooms. One is the wood shop the participant would stay in, and the other is the adjoining room from which the primary researcher

would observe and control the experiment. The wood shop area had two main tables for the subjects to use while undertaking the experiment. One table, the tool desk, was designated for all the relevant tools and equipment the participants would need to complete the experiment. The other table, the work desk, had a laptop and an Alexa Echo Dot. The laptop was set up with the tutorial video initiated before the participant entered the room. Prior to starting the experiment, the participants were instructed to use the woodworking desk to carry out the actual task and only to use the tool desk to retrieve tools. This was partly done to uniform the experiments and ensure the cameras could capture the participants' actions and facial expressions adequately. Two cameras were used. One camera focused on the entire woodworking table; the other was the activated webcam in the laptop. Aside from recording the participants' facial expressions, the webcam recording was also broadcast live to the primary researcher (wizard) sitting in the room next door.



(a) Camera angle 1 (b) Camera angle 2

Figure 4: Image of what both cameras recording during an experiment session.

6.2.1 *Task Procedure*

After welcoming the participant, the primary researcher first explained the general aim of the study and the procedure. After this step, we asked the participants to read and sign an informed consent form. It was unnecessary to ask for basic demographic information as we already did that in our

screening questionnaire sent to the participants earlier in the process, back when they initially showed interest in the study. The reason for this choice was primarily taken so that we could reject potential participants before they showed up in the lab.

Main Command	**Popular Variants**
Commands available for all conditions	
play	resume, go, start, begin
pause	stop, wait, hold on
fast forward	skip ahead, skip
rewind	go back, back
time-specific	e.g. go back 20 seconds, skip ahead 20 seconds
Commands available for condition 2 and 3	
sound-specific	e.g. go back to last sound, skip forward to next sound
content specific	e.g. go back to the duck, skip ahead two bell sounds

Table 4: List of commands supported by the system.

After finishing up the formalities, we told the participant that the purpose of the study was solely to evaluate the effectiveness of navigating how-to videos via a voice interface and not to test them. The study was set up such that the participants could not easily discern that they were interacting with the wizard, unlike an actual intelligent system. Specifically, this was achieved in part by having the participants and the wizard placed in two different rooms with a live stream of the participant carrying out the experiment transmitted live to the wizard.

During the experiment, the subjects would stand (or sit), assemble the wooden robot by the woodworking desk, fetch tools from the tool desk, and bring them over to the woodworking desk.

While the experiment ran, the participant was the only person in the wood shop room. The participants navigated the video tutorial by uttering their desired commands (like 'stop,' 'resume,' 'go back,' 'skip forward 30 seconds, 'go back to the last doorbell sound,' etc.). The participants in conditions 2 and 3 were instructed that they were able to both say the exact audio icon they wanted to – to either forward or return to - or they could say 'go back to last icon' or similar. We chose to include this function to increase the easability of using the new feature and minimize the chance that participants might be deterred from using the icons due to them not remembering the exact sequence of the audio icons.

Before starting the experiment, the participants were told they could call for the primary researcher to assist them if they had questions about the task or how to navigate the video by voice. Based on what we learned from the pilot studies, we decided that it was better to provide small clues and help the participants if they got stuck during the task rather than having to discard entire user sessions.

6.2.2 *Apparatus*

For all three conditions, the study used two laptop computers and an Amazon Echo Dot 3rd generation connected via Bluetooth to the participant's computer as an external speaker. The computer that the participant would use was a 3.16GHz quad-core Intel i7 laptop computer running Windows 10 64-bit Edition with a 15' display with a resolution of 1920 by 1200 and an NVIDIA GTX 1060 graphics card. The wizard had access to a MacBook Pro 15' 2019 model with a 2,3 GHz 8-core Intel i9 processor. As a token of thanks for participating in the study, all participants were given two microbrew beers or luxury chocolate, approximately 100 DKK. They were given the option to take home their wooden robot.

6.2.3 *Audio Icons*

The audio icons for this study consist of five basic sounds being played continuously in the same sequence. They only appear in conditions 2 and 3. Their sequence is the same across both conditions. The five audio icons are chosen because they are universally recognizable and easy to understand for the participants. The five are: (1) *dog's bark*, (2) *car horn*, (3) *door bell*, (4) *duck quack*, (5) *sheep* in that order. As seen, there is no one logic to the types of sounds chosen. Rather they all act as highly identifiable sound cues. This is intentional for the study not to have any sounds directly associated with the given tasks the sounds might represent, as they seek to be universal rather than specific. From the formative studies, we learned the optimum point where the audio icons were most effective regarding most recognizability while being the least annoying. Therefore, all five sounds were either truncated or spatially compressed to a duration of approximately 150 ms. Further, each audio icon was also subjected to a decrease in its attack gain, ensuring that the icons would not come barging in.

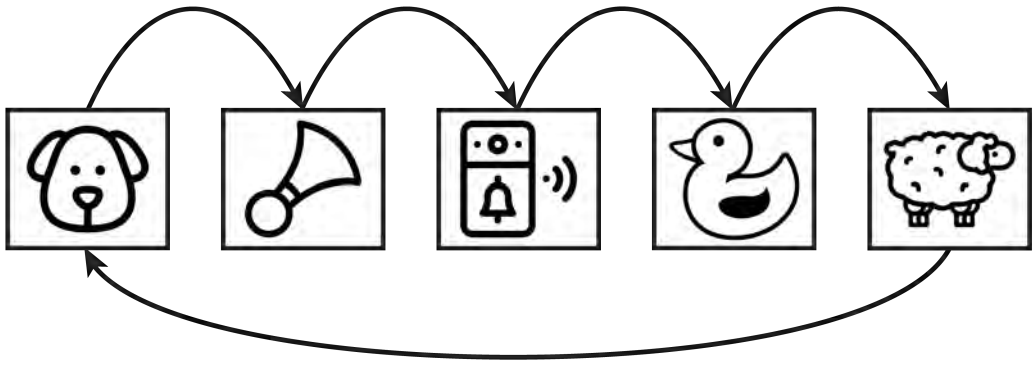


Figure 5: Illustration of audio icons. The illustration shows the audio icons' play sequence and how they loop back to the beginning. The sequence is (1) dog bark, (2) car horn, (3) door bell, (4) duck quack, (5) sheep.

6.2.3.1 *Time Position Overview for Condition 2 & 3*

The first number in the 'Name' column is the icon number in the 5-icon looping sequence, and the last number in parentheses is the number of times the icon has looped.

Condition 2

Icon #	Time	Name
1	00:29	1 - dog (1)
2	01:25	2 - horn (1)
3	02:23	3 - bell (1)
4	02:39	4 - duck (1)
5	03:14	5 - sheep (1)
6	03:50	1 - dog (2)
7	04:11	2 - horn (2)
8	04:21	3 - bell (2)
9	04:43	4 - duck (2)
10	05:08	5 - sheep (2)
11	05:38	1 - dog (3)
12	06:02	2 - horn (3)
13	06:33	3 - bell (3)
14	06:51	4 - duck (3)
15	07:14	5 - sheep (3)
16	07:27	1 - dog (4)
17	07:42	2 - horn (4)
18	07:50	3 - bell (4)
19	08:37	4 - duck (4)

Condition 3

Icon #	Time	Name
1	00:30	1 - dog (1)
2	01:00	2 - horn (1)
3	01:30	3 - bell (1)
4	02:00	4 - duck (1)
5	02:30	5 - sheep (1)
6	03:00	1 - dog (2)
7	03:30	2 - horn (2)
8	04:00	3 - bell (2)
9	04:30	4 - duck (2)
10	05:00	5 - sheep (2)
11	05:30	1 - dog (3)
12	06:00	2 - horn (3)
13	06:30	3 - bell (3)
14	07:00	4 - duck (3)
15	07:30	5 - sheep (3)
16	08:00	1 - dog (4)
17	08:30	2 - horn (4)
18	09:00	3 - bell (4)

6.2.4 *Video Content*

For this study, a wood-working assignment was chosen. The task was to assemble a small wooden robot with the task consisting of drilling holes in a set of pre-cut wooden blocks provided to the participants. This particular task was chosen based on a list criteria: 1 the task needs to be completed in a step-by-step fashion; 2 the user would be hesitant of placing their on devices in this environment due to potential damage to it (e.g. sawdust); 3 the task needs to occupy both hands, and 3 it should be a nuisance to having to actively switch between doing the task while simultaneously having to navigate the video tutorial with your hands. The video we chose to use in this study meets all the criteria above and is representative of videos about making small wooden figurines on Youtube.

Currently, tasks like cooking, knitting, news reading application, origami, and video based navigation have been tested under these criteria (Chang, Wang, et al. 2019; Chang, Huh, et al. 2021; Chung et al. 2020; Sahijwani et al. 2020). With this study, we add to the growing number of possible applications that step-by-step voice navigation tutorials can be applied to.

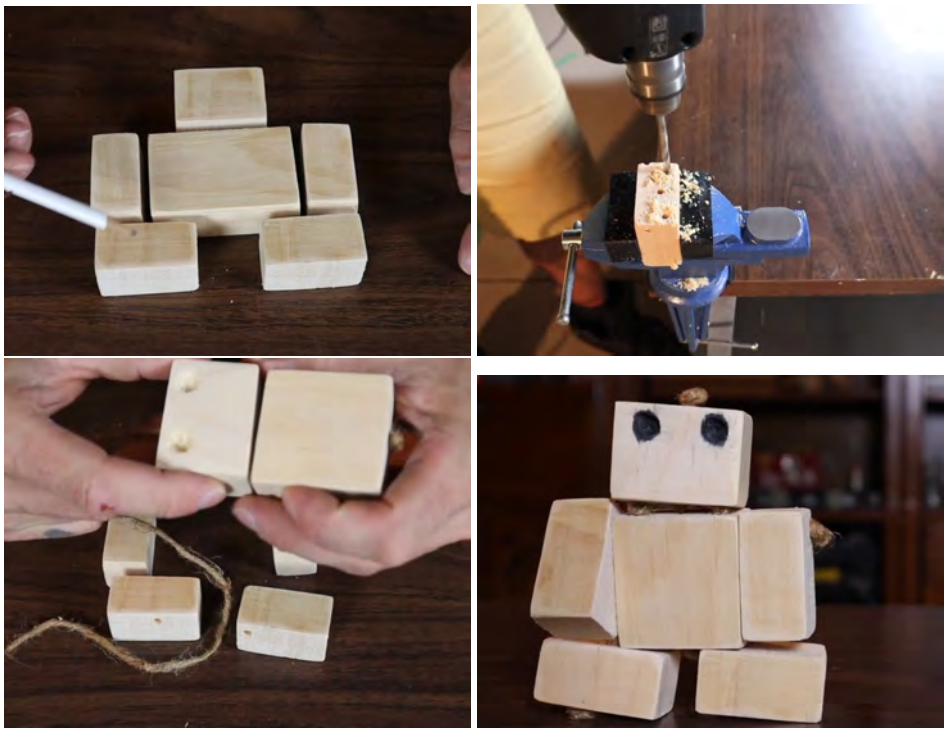


Figure 6: Four images from the video tutorial showing step-wise task progressions.

6.2.4.1 *Target Scene, Length, and Location*

The total running time of the video tutorial is 9 minutes and 12 seconds. Condition 2 has 19 audio icons at points of interest throughout the tutorial. Examples of points of interest are when holes are made in the wooden blocks when the eyes are drilled out; the rope is put through the wooden blocks, etc. In this way, the placements of the audio icons and their intrinsic meaning are not tied to the time between them, but to the content they seek to represent. The introduction of each audio icon signifies a point in the video wherein the user is required to carry out some specific action. Thus, making each a point of interest that demands the participants' visual attention. The how-to video was truncated to roughly the first image in figure 6, thereby leaving out the introduction and video of the instructor cutting the pieces of wood needed for the study. The wood was provided to the participants upon starting the experiment.

The total number of audio icons for condition 3 is 18. The audio icons are placed at a 30-second interval throughout the video and do not directly tie to any specific visual content in the video. In this way, condition 2 and 3 seek to understand how users might react towards having direct audio cues tied to the content in the video versus having evenly spaced-out anchors available to them to help ease navigation.

Video ID	**Title (Duration)**	**Domain**	**Creator**	**URL**
Video 1-1	Wooden Robot DIY [How-To] (397 s.)	Wood Crafts	Crafted Workshop	Link

Table 5: The How-To Video Used in the Study.

6.2.5 *Participants*

24 participants (9 female) between the ages of 20 and 33 (mean=26, std=2.54, see table 6) were recruited using convenience and snowball sampling by advertising the study with posters on the

university campus, accompanied by online recruiting as well. The study was conducted at Copenhagen University's Human-Centred Computing research section in one of the section's labs and an adjacent office. Four participants (all male) partook in initial trial experiments to narrow in the final experiment design, leaving 20 participants in the final sample. The participants were selected based on an online survey they were asked to fill out upon showing interest in the study. The requirements for being considered participants were based on: previous experience with speech technology and their ability to work with shop tools. None were excluded from the study based on these criteria. 5 participants did condition 1, 11 participants did condition 2, and 4 participants did condition 3. The reason for the unbalanced dataset is due to cancellations, Covid-19 and time restraints.

Participant Number	Age	Gender	Prior Experience Watching Watching How-To Videos		
			Wood Crafting	Tool Efficiency	Others
P1	24	M	YES	YES	Exercise Tutorials
P2	33	F	NO	YES	Knitting
P3	20	F	NO	YES	X
P4	26	M	NO	YES	Guitar
P5	25	M	YES	YES	Mathmatics
P6	28	M	YES	YES	Programming
P7	27	F	NO	YES	Makeup
P8	28	M	NO	YES	Gaming
P9	26	F	NO	YES	Programming
P10	28	M	YES	YES	Exercise Tutorials
P11	24	F	NO	YES	Arts and Crafts
P12	28	M	NO	YES	Exercise Tutorials
P13	25	F	NO	YES	Cooking
P14	26	M	YES	YES	Exercise Tutorials
P15	28	M	NO	YES	Programming
P16	27	F	NO	YES	Programming
P17	24	F	NO	YES	Makeup
P18	24	F	NO	YES	Knitting
P19	26	M	NO	YES	Programming, Math
P20	25	M	YES	YES	Programming

Table 6: Background Information of Study Participants.

7

DATA

7.1 THE WIZARD OF OZ (WOZ) METHOD

The Wizard of Oz testing method has seen extensive use as a testing tool for VUI prototypes (Klemmer et al. 2000; Porcheron, Fischer, and Valstar 2020; Large et al. 2019; Barko-Sherif et al. 2020). A key advantage of WoZ is that it enables the researcher to test out imagined scenarios without having to construct every part of the design product (Thymé-Gobbel and Jankowski 2021c). The basic premise of the WoZ method is to have the participants interact with a system they believe to be autonomous or 'intelligent.' However, unbeknownst to the participant, the system is controlled by a 'behind the curtain' human operator. The WoZ method is an effective tool for exploring the user experience of a complex, responsive system during the early stages of development as it enables an easy way of emulating a high-fidelity application (Pearl 2017).

Especially with VUIs, where language understanding can pose a hindrance, the WoZ effectively eliminates this aspect by having the researcher act as the automated machine part of the prototype. Another critical aspect of the WoZ testing scheme is not to let the participant know that the test setup is fake. This ensures the most realistic experience (Porcheron, Fischer, and Reeves 2021).

Based on these advantages in terms of applicability for this study, we conducted a WoZ experiment to verify the effectiveness of the proposed design in terms of its general understandability and cognitive

load requirements. To construct the WoZ model for this experiment, we implemented a set of keyboard hotkey commands corresponding to the positions of the various user input types, both VHS-style and audio icons. These controls enabled the wizard to remotely navigate the video the participants engaged with. We used two laptop computers; one was used by the wizard and the other by the participants. The wizard's computer screencast to the participant's computer to make the illusion that their inputs did control the video. The last element implemented was a live feed between the two laptops through their respective webcams. However, the broadcasting was only visible to the wizard, which enabled him to follow along in the experiment and provide the correct responses to the participant's voice input.

7.2 OBJECTIVE MEASURES

Since our proposed voice feature design is in a Command-and-Control dialog form, the granularity of our recorded measures extends to the commands that fit with what the system accepts. This means that we count and differentiate between six available command types. Currently, this way of interacting is the most common one used for most commercially successful voice systems. Where commands are designed to be no more than one turn-take and bear more resemblance to the user just giving the VUI orders, rather than being a true conversation (Ammari et al. 2019). We consider any additional commands that do not fit into this scheme in the discussion.

Measurement	Description
# of remote-like commands	Count # of basic navigation commands given
# of time-specific commands	Count # of time-specific commands given
# of content-specific commands	Count # of content-specific commands given

Table 7: Description of Measurements.

7.2.1 *Task Completion Time*

During each experiment, we timed the participants from when they made their first command until they finished creating the wooden robot. We did this to evaluate whether differences in time taken to complete the study exists between the conditions. While completion time provides a general measure of the effectiveness of the proposed augmented voice controls, it is essential to note that using the voice command interface in condition 2 and 3 did provide the users with more navigational options, thus incentivizing more command options and exploration. This point is that varying interface options give rise to different task completion strategies, potentially influencing the total time it takes to complete the experiment. Thus, a direct comparison of the competition times across conditions is only partially valid.

7.3 MENTAL WORKLOAD: NASA-TLX

To assess the mental workload of the participants during interaction across the different design iterations of the system, we had the participants fill out a NASA Task Load Index questionnaire upon completing the main task. The NASA TLX is a multidimensional scale designed to obtain workload estimates from users performing a task immediately after completing it (Hart and Staveland 1988). The way the NASA-TLX operationalizes workload is by establishing six dimensions on a Likert scale: *Mental Demand*, *Physical Demand*, *Temporal Demand*, *Performance*, *Effort*, and *Frustration*. When finished, the scores on the questionnaire are summed up to make out the overall workload score (Hart 2006). As holds true for any UX metric, the purpose of NASA-TLX is not to inform the researcher of what to fix. Rather it assists in understanding whether changes to the design have had any effect on the user's mental workload. In (Hart and Staveland 1988), the author makes the case that users' perception of mental workload may show to be a better measure than monitoring objective measures (e.g., heart

rate) as too many potential biases might influence the measure. We used the Raw NASA-TLX scores (Hart 2006), which is the NASA-TLX without the additional weighting process. For assessing the mental workload of VUI tasks specifically, however, the NASA-TLX has been used in several recent studies (Wu et al. 2020; Williams and Ortega 2020; Nouri et al. 2020; Habler et al. 2019; Jeon et al. 2015).

7.3.1 *System's Usability Score*

The System Usability Score (SUS) (Brooke 1995) is one of the most widely used questionnaire types for evaluating the usability of interactive systems (Kocabalil et al. 2018). SUS provides fast and effective methods for quickly assessing a design intervention's usability as perceived by its users. The process evaluates a design's effectiveness, efficiency, and user satisfaction. A strength of the SUS feedback for this project is that it enables repeatable and comparable assessments of the three different experimental conditions undertaken in this study. The questionnaire consists of ten-item Likert scale-type questions, which seek to assess the end user's perceived ease of use and learnability. Each ten-question item consists of a statement expressed from the participant's perspective. As one of the closing steps, after the participants finished the main experiment, they were asked to complete a SUS questionnaire. All participants across all three experiment conditions answered the same questionnaire about their attitudes about their interaction with the VUI system.

As we have yet to see tools and methods designed explicitly for VUI systems, we draw on existing methods, albeit they were initially developed to test the usability of screen-based systems. The SUS' applicability to properly evaluate voice-based user interfaces has previously been under scientific scrutiny. It was found that the SUS is a valuable tool for comparing voice systems (Ghosh et al. 2018; Cordasco et al. 2014). This, in turn, lends credibility and relevance to the test used in this project.

Question number	**Item**
1	I think that I would like to use this interface frequently.
2	I found the interface unnecessarily complex.
3	I thought the interface was easy to use.
4	I think that I would need the support of a technical person to be able to use this interface.
5	I found the various functions in the interface were well integrated.
6	I thought there was too much inconsistency in this interface.
7	I imagine that most people would learn to use this interface very quickly.
8	I found the interface very cumbersome / awkward to use.
9	I felt very confident using the interface.
10	I needed to learn a lot of things before I could get going with this interface.

Table 8: Overview of the SUS Questions.

7.3.2 *Semi-Structured Interviews*

As a last step in the experiment process, we conducted a short interview, asking the participants about their experience using voice control for video navigation. We asked about their experience and what strategy they used to complete the experiment. This was done to uncover how they conceptualize and take on instructional tasks. Then we asked if there were any particular interaction methods, they would have liked to see as part of the interactive features. We also asked them how they think their interaction pattern might have differed had they used a cursor and keyboard to complete the task. We had two questions specifically related to conditions 2 and 3, which relate to the audio icons — asking participants about the logic of the sounds and how they might have been used differently. The last

question was inspired by (Chang, Wang, et al. 2019) and seeks to learn what command gesture is appropriate should the user want to peek into unseen future content.

	Semi-Structured Interviews
Question 1	How was your experience with navigating a video via voice control?
Question 2	What was your overall strategy for completing this experiment?
Question 3	Were there any navigation functions you felt were missing?
Question 4	Do you think you would have provided more inputs if you were able to use the keyboard?
Question 5 (2 & 3)	What do you think logic is behind the placements of the sounds?
Question 6 (2 & 3)	How can the sounds be designed differently to make them more relevant for you?
Question 7	Can you think of a command that will enable you to 'peek' into the future to see the finished product?
Question 8	Do you have any additional comments?

Table 9: Interview Questions. The Questions Are Derived Partly (C. M. Myers et al. 2019; Chang, Wang, et al. 2019) and From the Pilot Studies.

Part IV

RESULTS

8

EXPERIMENT METRICS

We collected a series of experiment metrics for each experiment to quantitatively understand to what degree the different issued commands were used. The main metrics we report in this section are; *Total of commands*, and other specific types of commands are given. For all experiments the *Pause, Play, Rewind, skip forward,* and *time-specific skips* are all valid commands. Further for condition 2 & 3 the total number of *Content-based references* are considered as well. These are further divided into whether the command in question was a direct reference to a specific audio icon or if the participant only uttered, e.g., ”Go back to the last sound.” or ”Skip forward two sounds.” Commands like these are also accepted, as the first iteration experiments showed that users had trouble remembering the specific icon they wanted to go back to. So instead, they could skip back one sound and, in that way, maneuver through the audio icons analogously as chapters. The total time it took each participant to complete the experiment is also noted. However, at the beginning of each experiment, all participants were instructed that the total time it would take them to complete the experiment was not prioritized. This instruction was specifically given to ensure that people would not feel unnecessary stress due to them wanting to complete the experiment fast, as this could potentially give rise to errors in the resulting data.

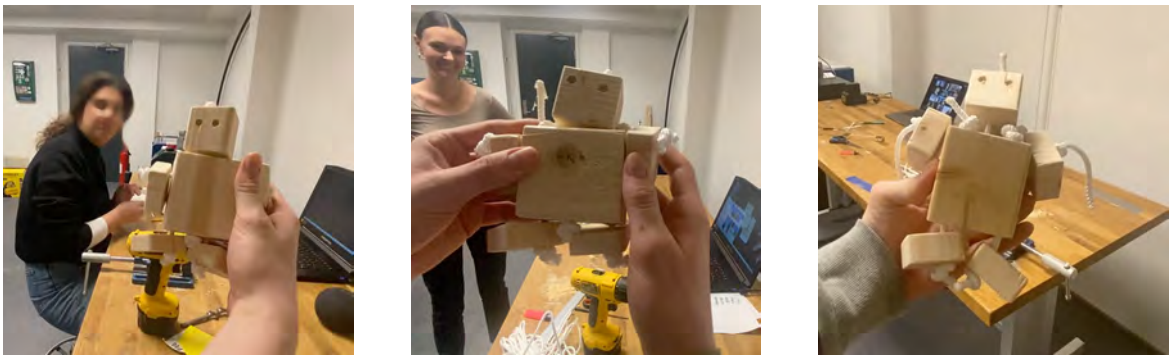


Figure 7: Three Examples of the Finished Wooden Robot.

8.1 NUMBER OF COMMANDS & TYPES OF COMMANDS

In table 10, we see the distribution of the different types of commands available for the participants in the varying types of experiment conditions. The total number of average commands given per condition differs. With condition 1 having the highest number of issued commands, with condition 2 making the second most commands and in third comes condition 3. While the average number of commands differs, the distribution in percentage between the *pause* and *play* commands remains relatively stable in in all the conditons; making up 73.19%-78.32% of all commands given. This finding correlates with results from previous studies (Tuncer et al. 2020).

	Average # of Commands	**Pause**	**Play**	**Rewind**	**Skip Forward**	**Time-Specific**	**Content-Specific**	**Sound-Specific**
Condition 1	40.60	37.93%	40.39%	8.87%	1.48%	11.33%	-	-
Condition 2	30.18	34.94%	38.25%	6.33%	0.90%	7.83%	6.02%	5.72%
Condition 3	28.00	34.82%	38.39%	4.46%	1.79%	17.86%	0.00%	2.68%

Table 10: Average Number of Commands For Each Condition Expressed in Absolute Terms. The Distribution of Command Items are Expressed in Relative Numbers.

8.1.1 *Time-Specific Commands*

For all conditions, time-specific commands were used more often than the pre-set 10-second rewind and fast-forward commands. We see a difference in use between the different conditions for time-specific commands. Condition 1 used the time-specific commands most often, and condition 2 used them the least. As condition 3 had very little engagement with the audio icons, the results of this condition are limited. However, it seems that when participants actively discarded using the audio icons, their use pattern would come to resemble condition 1 more. Condition 2 used the fewest command types while having the highest number of consent-based commands (11.75%).

8.1.2 *Content-Based Commands*

Content-based commands are only measured for condition 2 and 3. The content-based commands are counted as the total number of content references. The total number of content references is divided into two different categories. Either content-specific references or sound-specific references. This distinction is made because the formative studies showed that participants had considerable difficulties remembering the order of the audio icons or which audio icon was played last. Content-specific references are commands where the participants refer to the audio icons *indirectly*. Meaning they referred to the audio icons by saying, e.g., 'go back to the last sound,' 'skip forward a sound,' etc. Sound-specific references are commands where the participants directly refer to the audio icon. These commands can be, for example, 'go back to sheep,' skip forward two dog sounds.'

4 participants in condition 2 and 3 did not use the available audio icon commands once during the experiment. From condition 2 is P5, and from condition 3 are P16, P17, and P19. This means that only one participant in condition 3 used the sounds. Due to the limited number of participants in

	Total Content-Based References	Content-Specific Commands	Sound-Specific Commands
Mean	3.55	1.82	1.73
STD	2.21	1.94	1.27

Table 11: Mean and Standard Deviation For All Content-Based References.

condition 3 who did use the audio icons, we exclude this condition. The underlying reasons why they did not use the sounds will be dealt with later in the discussion section.

In condition 2, the participants uttered 39 content-based commands with an almost equal distribution of 20 content-specific and 19 sound-specific commands. This suggests that participants had major issues remembering which specific audio icon was played last from their current time point. The inability of being unable to remember the exact sequence of the audio icons or which was relevant for some specific time point was voiced throughout the experiment by almost all participants. P2 expressed that she was happy to refer back to the last played sound without remembering which one it was. This enabled her to use the audio icons as video chapters to navigate the video.

Of the total number of commands given in condition 2, content-based references made up 11.75% of all the commands. This number seems to comply with previous studies, confirming the ratio between temporal and content-based referencing is at around 9:1. (Chang, Wang, et al. 2019; Yarmand et al. 2019).

8.2 TASK COMPLETION TIME

During the instruction, we told the participant that their completion time was not a measure of success. Figure shows the average task completion time from the time the participants performed their first command until they finished the last step in the wooden robot task.

Participants were faster when interacting with either of the two variations of the augmented navigational voice commands condition. The baseline condition (condition 1) was the slowest (μ = 2553.2, σ = 421.9). Condition 3 was the fastest (μ = 1695, σ = 647.6), followed by condition 2 (μ = 2286.3, σ = 856.84).

Figure 8: Total task completion time For All Conditions. Lower Scores Indicate Faster Completion Time.

9

NASA-TLX

The difference in cognitive load between the three conditions was not significant. People reported a high task load for the baseline and both audio icon study variations (see table 12). However, we learned that the reason for the high task loads is different between the conditions.

	Mental	**Physical**	**Temporal**	**Performance**	**Effort**	**Frustration**	**X**
Condition 1	50.00	50.00	40.00	68.00	62.00	62.00	55.33
Condition 2	57.27	37.27	40.00	77.27	55.45	52.72	53.33
Condition 3	55.55	42.50	32.50	72.50	62.50	45.00	51.66

Table 12: Cognitive load measured with NASA-TLX (0-100) for 20 participants. There are no significant differences in cognitive load between the three conditions.

For the baseline experiment, participants reported a higher degree of *frustration* attributed to this condition. P1 and P3 said it felt tedious to jump around to the specific temporal location they were looking for by saying, e.g., 'go back' multiple times continuously. Further, the baseline condition produced a higher score for the *physical demand* question. This result coincides with the fact that participants in condition 1, on average, had to make more commands to complete the experiment (see table **??**). As condition 2 and 3 can be seen as more complex (given their augmented features), perhaps this explains why both proposed conversational systems received a higher score on items like *Mental Demand*. To contrast this assessment, condition 1 received a better score on the *Performance*

item (a lower score is considered better for this item). Indicating that audio icons in this experiment did not positively contribute to how successful users were in accomplishing the task's goals.

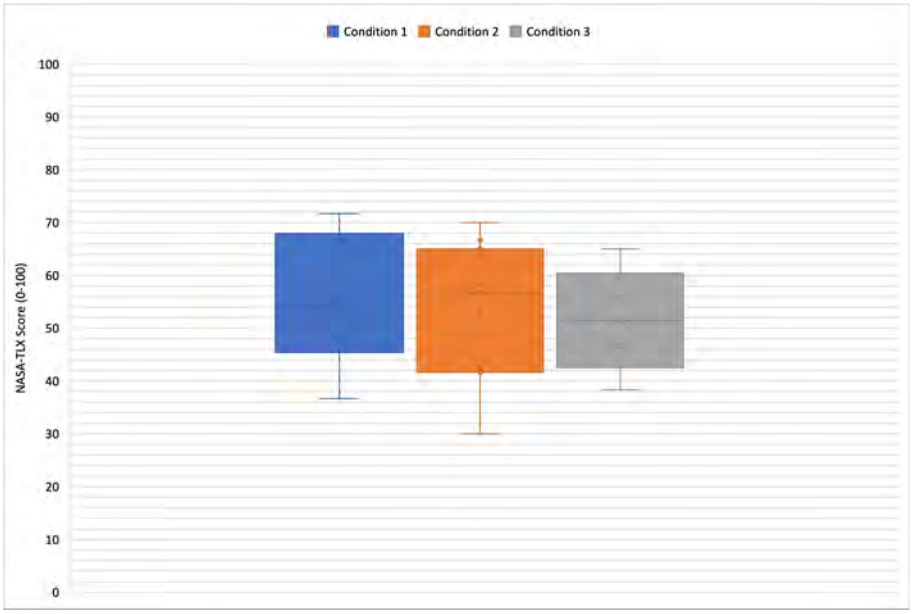


Figure 8: Average NASA-TLX scores for all three condition. Higher scores indicate higher task load.

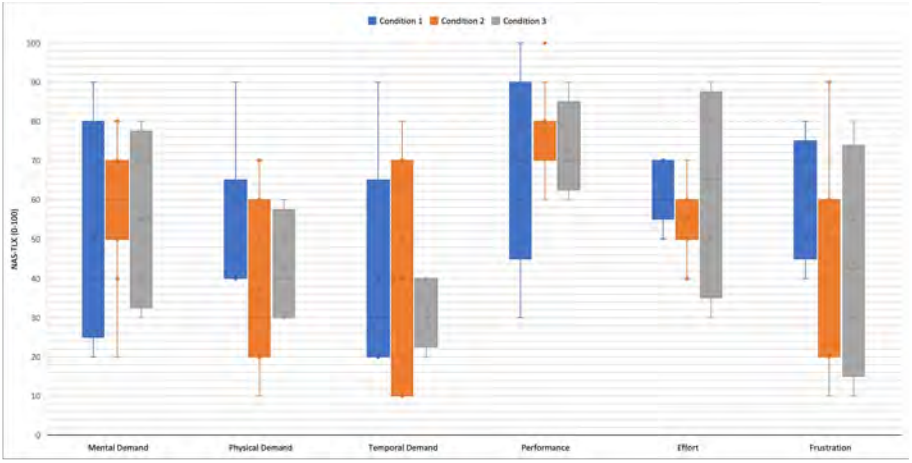


Figure 8: Boxplot of median NASA-TLX scores for the 3 conditions. Higher scores indicate higher task load.

10

THE SYSTEM USABILITY SCALE

After a participant finished the experiment, they were asked to fill out a SUS survey detailing their experience regarding the usability of the particular condition they were part of.

		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	X
Condition 1	M	2.80	2.40	3.25	1.60	2.80	2.40	3.20	2.60	3.20	1.80	2.61
	STD	1.30	1.14	1.50	0.89	1.30	1.14	1.48	0.89	0.45	1.30	
Condition 2	M	4.45	2.45	4.55	1.91	3.55	2.55	3.73	3.09	4.36	1.82	3.25
	STD	0.93	1.29	0.69	1.04	1.04	1.51	1.10	1.64	0.81	1.25	
Condition 3	M	4.00	2.00	4.50	2.25	3.75	1.75	3.50	2.25	4.25	3.50	3.18
	STD	0.00	0.82	0.58	1.50	0.50	0.96	0.58	0.50	0.96	1.00	

Table 13: Average SUS score for each question in each condition. Higher score indicates higher usability rate.

For SUS, this study analyzes the average mean SUS score and the standard deviation for each of the three tested conditions. Participants provided a higher SUS rating for condition 2 (μ = 81 , σ = =11.69) than for condition 3 (μ = 79, σ = =2.4). The baseline experiment condition 1 gave the lowest SUS rating overall (μ = 63, σ = =15.25). By convention, a SUS score of 68 and above is considered an 'above average' design. The data gathered in this study shows that the baseline condition did not meet this criterion. However, placing somewhere in the middle might be expected from a set of remote-like controls as these are the industry standard and therefore do nothing but affirm the participant's already

established ideas of what a voice control in a video setting entails. Both condition 2 & 3 managed to get a relatively high score on the SUS scale, which places both designs in the top 15 % based on an accumulative distribution of SUS scores (Ghosh et al. 2018).

Figure 9: Average SUS scores for all three conditions. Higher scores indicate higher usability.

Additionally, we analyzed each question item to understand the participants' assessment of the perceived usability of the design intervention. See table 8 for overview of SUS questions.

For all questions, except question 7, either condition 2 & 3 scored the highest. However, this result might be somewhat expected as condition 1 is the least complex condition of the three with the least prerequisite knowledge needed to carry out the task. For question 3, condition 2 scored considerably lower than the other conditions. The reason for this might be that it required more training to understand the introduced audio icons and focus on how they relate to the content shown on the screen. In favor of the proposed design intervention, both condition 2 & 3 scored relatively high on question 9, indicating that when participants had somewhat internalized the added commands, they felt comfortable using them.

11

POST-EXPERIMENT INTERVIEWS

11.1 VOICE VERSUS KEYBOARD

During the initial pilot experiments, the test subjects were asked to complete the wooden robot task using the provided laptop's built-in keyboard and essential VHS functions to maneuver the video. In these trial runs, the total number of keyboard inputs was, on average = 109.62 (σ = 20.45). Comparing this relatively high interaction rate with the total number of voice inputs (μ = 40.60) from each documented study condition reveals a considerable difference in the total number of inputs. This discrepancy between keyboard pressings and voice commands is naturally intriguing.

When the voice-command participants were asked whether they believe they would have given more, fewer, or an equal amount of inputs if they had the keyboard at their disposal, 19 out of the 20 participants answered that they were confident they would give more commands via the keyboard than they did with voice. Only P14 responded that he thought he would provide fewer clicks than he did voice commands. He reasoned that he would have to put the task building aside to use the keyboard, which he would not do. Whereas, with voice input, he could work uninterrupted from this nuance. Although others similarly recognized the benefit of not having to put down the tools and wood pieces to maneuver the video, they argued that the 'cost' of a bottom press is significantly lower than issuing a voice command. This ultimately means skipping over your desired video time position is less of an issue, and you can easily maneuver back with a few pushes of a button. Others (n = 7) expressed that

they found it somewhat awkward to talk to a machine. An often-cited reason the participants found the experience uncomfortable was usually due to voice commands not being their favored input type when interacting with digital systems. However, as the experiment progressed, people became more familiar with the system's setup and premise; this issue seemed to diminish in relevance.

For example, P3 said the following: "Because it feels easier (keyboard) and the price of doing an extra click versus doing a full vocal utterance is less costly, in my opinion. I can also carry out these types of commands (keyboard/cursor) much faster because it's not new to me in the way that sound inputs are. So it's not as cognitively demanding for me as saying a full sentence might be."

Similarly, P5 explained how he has voice-enabled devices all around him at home. In his phone, computer, and TV. And even though he has played around with the technology, this has primarily been for the novelty aspect, more so than trying to discover a new preferred way of interacting with his smart devices. He reasons that in most circumstances, he already has the keyboard or remote at his disposal, so the incentive to use something new is unnecessary. However, Both P3 and P5 acknowledged that they could easily see using voice commands as the preferred input methods for doing something concrete like a step-by-step tutorial. Primarily since it provides the freedom to use your hands for other things than controlling the video, as well as not having to consider whether you will get your computer dirty or not.

Additionally, comparing the differences in the effectiveness and execution of voice commands against conventional direct manipulation methods remains an unanswered question (Chang, Huh, et al. 2021). Whether the use of voice commands fundamentally changes the way users interact with a video system is too encompassing for this study to make any meaningful comments on. What we can derive from these results, however, is that there is a difference between the total number of inputs one can expect to see across the two groups. In HCI, we tend to favor the design that requires the fewest total number of interactions. However, this study does not present the necessary evidence to comment on whether this is true for fewer voice commands.

11.2 CRITIQUE OF THE SOUNDS

The participants doing condition 2 & 3 were exposed to the audio icons. In condition 2, participants had the sounds play at a point in the video that required them to perform some action, and in condition 3, participants had the sounds play at 30-second intervals. As part of the prepared post-experiment questions specific to these conditions, the participants were asked what they thought of the audio icons used in the experiment and any prospective critiques or comments on how they could be potentially improved in future design iterations. Some participants recommended using earcons instead of the different audio icons, with a difference in, for instance, pitch, tone, octave, etc. Others pointed to re-making the sounds into something akin to a spearcon, where the icons would directly resemble the visual content it mapped to.

11.2.1 *Not connecting the sounds with the video content*

4 of the participants in condition 2 and 3 voiced how they failed to link whatever action was undertaken in the video with the particular sound that was being played.

P5 stated: ”It’s like you hear a bunch assorted sounds, and then you have to keep track of their intrinsic connection to each other, as well as the sound’s meaning in the video content. And I just don’t find it easier to keep track of all this extra information.”

The fact that there was no apparent link between the audio icons whenever they were running in sequence, as explained by P5, was also stated by other participants. For instance, P13 said the only times she used the sounds was when she had actively paid attention to what video content the audio icons pointed back to. However, keeping constant track of which audio icons were being played while trying to follow the video content and completing the unfamiliar given task was too cognitively demanding.

11.3 VIDEO CONTENT PEEKS

A point raised in (Chang, Huh, et al. 2021) is that no voice command exists, which enables the user to 'peek' into the future of the video to get a glimpse of what the finished product is supposed to look like. In the same manner as Swifter (Matejka et al. 2012; Matejka et al. 2013) have enabled content-based video searching from the small thumbnail 'peeking' into future content a the video. As one of the finishing questions in the post-experiment interview, the main researcher would ask the participants what vocal command they imagined would best cater to this specific content of the video. The response to this question was fairly uniform, as participants would say things like; 'go to summary,' 'show the final product,' 'go to the end,' etc. P1 and P20 both tried to navigate to the end of the video via time-specific commands. However, this process was clunky and required them to do multiple commands to have them arrive at their desired time point. When going back to the beginning of the video, P1 tried to give the command 'go back to start.' Realizing this command was not supported, he had to rewind back in time again by specifying via time jumps.

Part V

DISCUSSION

12

DISCUSSION

12.1 QUANTITATIVE DATA

We conducted a study with 20 participants. Although the sample size did give us sufficient power to reveal differences in attitudes towards using audio icons in VUIs to assist navigating how-to video tutorials, statistical power to differentiate the finer nuances between the individual design proposals are certainly limited.

Considering figure 9, we observe that all three conditions are positively skewed. We see a relatively high variability in condition 2 versus condition 1 & 3 which both exhibit a tight range. Alternatively, the higher degree of dispersion in condition 2 can be interpreted as a higher spread in user preference towards the design intervention. We find the lowest score in the entire dataset given to condition 1. Suggesting that some users might find great value in voice navigation (and expanded navigation options), whereas others have a clear preference towards other input types. Like not using voice controls at all. When comparing the median between condition 2 & 3 we see that their medians are numerically close, suggesting that they both provided an advantage over condition 1. This results affirms that adding these types of additional voice navigation options had a generally positive impact on the usability of the interface.

12.2 USER STRATEGIES

A vital observation during the study is the emergence of discernible user strategies to complete the experiment. Arguably, the participants' experience with prior interfaces. While we used priming to reduce legacy bias, this aspect conceivably still played a significant role in the strategies the participants employed to complete the task.

We observed that while participants were performing either basic control options (e.g., "go back," "fast forward") or time-specific commands ("go back 20 seconds"), they often did the same command multiple times; in this way, they performed shorter incremental time jumps. While it might seem redundant to have to perform a series of the same commands to reach your desired time point, it appeared that the participants purposely did this, as they were not completely sure of the exact time location they wished to return to. This strategy functions as sequential content peeks, which assisted them in deciding whether they had reached their desired time-point. For this reason, while it appears as an apparent interaction to try to reduce, it should be noted that this style of interacting does serve a purpose in its own right when users are not entirely certain where exactly they need to go time-wise but have an idea when they see glimpses of the content that followed it. Therefore, supporting users when they are performing the same command in sequence (e.g., saying "go back" multiple times in a row) by allowing for the VUI to accept abbreviated types of the commands like "back" or "again" might potentially ease usability. Based on the notion that users employ specific navigation strategies, establishing design constructs that actively seek to support entire pre-set navigation options for different user strategies might provide a better-tailored user experience. This recommendation is analogous to how a system initially asks its users if they are novices or experts or if they prefer their pointer and cursor to be inverted.

	Description	*Characteristics*
Start-Stop Dominant	Users follow the video instructions very closely. They will actively stop the video tutorial when their focus is not on the video.	Primarily use the pause and play function. There is less need for navigating the video back and forth.
Time-Specific Dominant	Users take in information to complete a subtask, and will just let the video run until they redirect their focus back to the screen.	Makes extensive use of specific time references instead of pausing the video.
Tool Efficient	These users exhibit exceptional proficiency when working with tools. They are able to complete the task more autonomously.	Will generally make fewer commands overall.
Audio Icons	These users favor audio icons over other types of commands, e.g., time-referencing. Users in this strategy tend to have the most varied use pattern of all the strategies.	Users makes use of audio icons to navigate the how-to video to skip back to the beginning of a subtask, or skip ahead if they have already managed to complete the subtask.

Table 14: Overview of User Strategies Identified in This Study.

12.3 USER EXPECTATIONS ARE NOT MET

When people think of intelligent systems, they presume they can communicate complex, or at least semi-complex utterances to the machine (Knees et al. 2019). While the interaction system deployed in this study is not inherently 'intelligent,' it nudges to a possible future avenue for research within this field. Considering some of the issues raised by the participants in this study, many found it especially difficult to remember the (to them) arbitrary audio icons while simultaneously carrying out an unfamiliar task. At the same time, some pointed to the idea of replacing the audio icons with different sounds, which has more inherent meaning to one another, like a number series (going from 1-5 in a loop or complete sequence), or having a uniform and "pleasant" sound rising in, e.g., pitch, tone, or note. Either of these two suggestions would cater to the need to create some linkage between each audio icon and, in turn, make it cognitively easier for users to understand their navigation choices and the connection between them. However, there is still the issue of connecting the audio icons to the content of the video positions in which they seek to make aware.

As mentioned by some participants, one way of tackling this is to make the audio icons specific for each section they seek to represent (e.g. "go to assembling arms part," "go to eyes part" etc.) and have the voice system vocally confirm the command to eliminate any uncertainty. P2 said she felt it might make the navigation less ambiguous and not require the user to remember many things or keep track of a sound sequence while undertaking a physical task. However, introducing an additional voice means we are now dealing with the voice of the narrator, the user, and the audio icons. As mentioned in (Pyae and Joelsson 2018), multiple voices can potentially be an issue, as it can potentially prove to make the user experience feel clunky, as people experience difficulty distinguishing voices, their hierarchy to each other, and relative importance.

Another solution to this problem could be to use some form of TTS icons or spearcons where the latter have proven most effective in previous studies (Suh et al. 2012; B. N. Walker et al. 2013). These reformed icons could take the shape of *audio chapters*, saying, e.g., "Legs" or "Arms" whenever

appropriate. This design choice will to a more significant degree, succeed in elucidating the link between icon and video content. However, it will come at the cost of universality and would most likely have to be changed depending on cultural conventions, etc. P14 suggested that instead of using the audio icons used in this study, another potentially exciting design choice is to make the icons directly resemble the action which needed to be carried out in the tutorial (e.g., play the sound of a drill machine when the user needs to use this tool or a hammering sound, etc.) While this approach attempts to map and reduce abstraction between audio icon and action more directly, it can prove especially difficult to create appropriate, easily recognizable, universal sounds for all steps in a complex task scenario. For instance, what sound is adequate to communicate that the user now needs to pull a rope through the torso wood block and attach it to the legs).

12.3.1 *Understanding the Higher-Level Intent*

From the Results section, we see how P1 and P5 each try to give the system a command based on their expectations of what an intelligent conversational AI system can do by saying ”Go back to start” and ”Go back to where he puts on the head.” Both utterances were done without knowing whether the system would accept them. Therefore, we can perceive the utterances as natural reactions to what the participants ideally wanted to achieve in those situations. This implies that a voice system for video controlling might benefit from some elasticity regarding what it can understand and accept as a valid request. Essentially, this aspect was voiced by multiple participants to some extent. They reasoned that enabling the VUI to understand the intent behind the utterance would alleviate many of the design shortcomings they experienced in this study.

12.4 ACTION-POSITION VERSUS 30 SECOND TIME INTERVALS

12.4.1 *Participants Not Using Audio Icons at All*

From the results, we see that a total of 4 participants from the content-based conditions (condition 2 (n = 1) and 3 (n = 3)) did not make use of the audio icons. The experiences of these participants are of particular interest, as they might very well provide the strongest arguments against the proposed design intervention.

In condition 2, only P5 did not make use of the audio icons at all. While understanding the icon's logic, he expressed that he had difficulty connecting the audio icons to chapters or subtasks while undergoing the experiment. Upon observing his navigation strategy and considering his explanation, he "[...] is visually oriented. Instead of having people tell me how to do something, I'd much rather just look at how he does it in the video". He paused the video and then did incremental forward or backward skips. In this way, he navigated the video using still images. Very much akin to how visual content-based referencing works. He compared this experiment to how he would typically consume instructional video content on platforms like Youtube. Saying how this specific way of interacting (with keyboard or cursor) is just so internalized for him now. He acknowledged that the audio icons try to complete the same task as, e.g., pre-cached thumbnails on the timeline, but the ease-of-use was not adequately carried over.

Condition 3 suffered from some shortcomings. The data sample is small, consisting of only 4 participants. The relatively unbalanced dataset is due to external factors like participant cancellations, the current Covid-19 situation, and time restraints. To some extent, this limits the condition analysis, with only one out of four participants using the audio icons. Instead, we shall look at the answers given by the participants as to why they did not make use of the audio icons. P19 did not make use of the audio icons. He completed the exercise in only 955 seconds (average is 1577.5 seconds for this condition) and made only 21 commands (6, back, 7 skip ahead, and 11 time-specific skips). When

observing the participant, it was clear this person was above average and familiar with woodworking and crafts. When asking P19 why he did not use the augmented voice controls, he stated that he failed to see how their logic, and as a result, he just ignored them. When asked what he thought the reason behind the audio icons might be, he suggested that from his experience, they were just placed at random time intervals. He further indicated that he thought the actual experiment tested how sonic nuisances influence users' concentration. Arguably, this participant's prior experience working with wood materials made him feel safer doing the task more autonomously. It is feasible to speculate that this particular task might have been too easy for this individual, rendering more or less any interaction redundant. Meaning that he most likely would have been able to complete the wooden robot after watching the how-to video from start to finish in one take. This might explain why he needed to interact with the interface less than most other participants and why he was able to complete the exercise that fast. P16 and P17 exhibited relatively similar interaction patterns. They both understood that the audio icons occurred at some specific time interval and that they could be used as anchor points for navigation. Interestingly P17 pointed to the notion that because the audio icons were placed at a specific set time interval, the resulting interaction was the same as if she had just said, e.g., ”go back 30 seconds”. She attributed this to be the main reason she did not use the audio icons. P16 tried giving the command ”skip,” referring to skipping over a 30-second interval as if they were chapters. In the interview, she said she thought giving this command would have been easier than telling a complete sentence like ”skip ahead 30 seconds” or similar.

12.4.2 *Pause and Play*

Some users favored stopping the video, and others had no problem just letting it run while focusing on the physical task. There have been proposed systems to identify different sections and transitions of instructional video content (Nguyen and F. Liu 2015) or automatic pausing at relevant points to help users while undertaking a task (Pongnumkul et al. 2011). Innovations like these would undoubtedly

cater to many users; however, as our findings suggest, there can similarly be advantages to letting users control their pausing. This seems to support their interaction strategy, emphasizing user control more naturally. For example, they may deliberately fall behind the video as an active method to efficiently progress in the physical task while just slightly following the progress of the video from the corner of their eye or using the video audio track to give them an idea of what is about to happen next.

12.5 DESIGN CHALLENGES

Designing voice commands for how-to videos is not about delineating every command the user can perform but understanding the higher-level user intent behind the command. For instance, P1 asked the system, ”I would like to return to the part when here make the holes for the eyes.” Contextually, the system should understand that this is not a remark but a request. A command like this one holds meaning on more levels. First, the system must discern that the user is actively asking the system to perform an action, and the system must know at what location the eye-hole drilling is placed.

12.5.0.1 *Video Pace*

Some participants also voiced that they would have liked to be able to control the video’s pace. When asking them how they would have used the pace option, P10 said, ”It would have been nice to have the video pace up whenever the narrator was performing the task, and then slow back down whenever he was explaining something.” Supporting a feature like the one explained would inevitably require the user to perform more commands whenever they wish to slow the video down and speed it back up again. In line with similar findings in (Chang, Huh, et al. 2021), a way to address video pacing in a VUI context might be to enable a function that would, for example, speed the video up to 1.25x when the video showed how the task was done, and then pace back down to 0.9x whenever something was explained. So instead of conceptualizing pacing as something the user has to control themselves manually, you employ ready-made pace options for the user, which can be switched on and off at will.

12.5.0.2 *Video-Based Discoverability*

While this study specifically focuses on investigating how audio icons might improve the usability of VUIs for navigating how-to videos, we still have to acknowledge the fact that, by nature, our area of interest does involve both an auditive and visual component. P2 and P9 said they would have liked to have some visual aid to better link the audio with the video. P2 suggested displaying a Rolodex-type thing where you can scroll forwards and backward to past events, "[. . .] maybe also with an image of the animal linked to that command". P5 suggested including a permanent timeline with small thumbnail images of what was happening at each time location. Somewhat akin to the video sequencing tracks in video and image editing software. These suggestions resonate with (Tuncer et al. 2020), who suggested displaying a picture-in-picture frame whenever the user pauses the video. Contextualizing this idea to this study, we can imagine a picture-in-picture frame is shown whenever the video is paused with the video continuing. This gives the user a visual reference point they can return to should they need to. While these suggestions are attractive and can inspire future avenues of research, they can also be seen as a sign of a failed attempt at making a genuinely functioning voice system, as users are drawn back to relying on the screen.

12.5.0.3 *Alignment Pause*

In (Chang, Wang, et al. 2019), they learned a primary action that users tried to perform when interacting with how-to videos is to stop and pause the video at the frames where the video shows the stepwise completed progressions the of the tutorial. P15 was part of condition 2. He only made use of the audio icons one time. When asked why this was the case, he explained that he felt the placement of the audio icons did not match the places he would like to fast-travel to. Instead of going to the beginning of each subtask, he would have preferred the icons to lead him to the summary of each subtask. This would have enabled him to go to the audio icons, align his physical task with the one on the screen and then move on with the task. If he doubted how the video had arrived at a specific

step in the process, he could skip back to his command to see how it was done. Following this design, imaginably, some users might even be able to finish the woodworking task by only referencing the summary (aligning with physical task) images. This can also suggest that it is beneficial to implement both an icon at the start of each subtask and a different icon at each summary of that same subtask to cater to both navigation needs.

Part VI

CONCLUSION AND FUTURE WORK

13

CONCLUSION

In this study we investigated whether and how audio icons can benefit voice navigation in how-to videos by having our participants carry out a practical task of assembling a small wooden robot with the help of a video tutorial and a VUI. The conditions containing audio icons achieved lower task completion times and fewer total voice commands needed. Further, they scored higher on the SUS score, whereas the cognitive load score was inconclusive. Although the inclusion of audio icons to improve navigation of video how-to videos does show promise, the study did identify a number of serious usability challenges associated with this augmented navigation method, as reported by the participants. These usability issues are assistive and insightful for designers of future VUI systems and contribute to establishing more grounded research in area of voice user interfaces.

14

FUTURE WORK

14.1 MORE USE CASES

In this study, we had the participants carry out the practical task of assembling a small wooden robot figurine. We specifically chose this task because it encompasses some level of complexity, its task is comprised of a series of smaller subtasks, and the task is 'messy' in the sense that users will get dirty, which in turn incentives them not to have to touch the keyboard to control the video. Previous studies have used various practical tasks like cooking a dish, knitting, practicing yoga, doing origami, and picking a lock with a paper clip (Tuncer et al. 2020; Cho 2018; Nouri et al. 2020). While all these practical tasks are multi-faceted and varied in form, they all still largely adhere to the task attributes of our study. Future endeavors into studying voice control in how-to video settings can benefit from demonstrating more diversity in terms of the viability of our proposed augmented interaction features by testing them in different task scenarios. Application domains containing complex tasks include home and auto, furniture assembly, and calendar management. Current VUI uses are often reduced to being able to help people perform simple tasks. Therefore, expanding on the scenarios in which VUIs are used to achieve complex multi-step tasks can further assist in broadening users' understanding of the capabilities of VUIs.

14.2 AESTHETICS

The specific five audio icons used for this study were not chosen because of their perceived appeal. Instead, they were selected due to their ease of readability and because they purposely abstracted the audio icon from the video content they aimed to represent. Aesthetics tend not to be prioritized as high as, e.g., usability and learnability in interaction systems (Hornbæk and Bargas-Avila 2011). However, the 'niceness' of a design has the strength to decide whether an artifact will get adopted or abandoned. This implies that beyond what has been demonstrated in this study, future research can potentially look into what effects aesthetics will have on voice navigation.

14.3 ALTERNATIVE AUDIO ANCHORS

As expressed by the participants, the audio cues used for this study were not intuitively understandable. They voiced criticisms like the sounds not carrying any intrinsic meaning in relation to each other. For example, the users felt frustrated because they did not see any resemblance between, e.g., a dog bark followed by a car horn, etc. Further, the participants said they sometimes failed to link the audio icons to the content displayed in each of the chapters beyond being made aware that something was happening on the screen when the audio was initiated.

When asked how these shortcomings could be addressed, the participants mentioned different designs, like each anchor point having its separate number (going from 1-19). With this implementation, the user could refer back to a specific anchor digit. This design choice has the advantage of possessing an inherent correlation with the rest of the number series, and no number relates to more than one time-point. Enabling interaction scenarios like "go back to number 7" when coming to number ten in the video. Further, this design would quite possibly ease the cognitive load on the user, as they are no

longer required to hold the abstract audio series in mind at all times. In the sense that they will not have to remember that duck quacking precedes the doorbell etc.

14.3.1 *Understanding User Strategies*

A critical takeaway from this study is that the users employed different strategies to complete the experiment. Some used a start-stop strategy, where they would mainly try to control the advancement of the video by just stopping it whenever they needed to catch up. Others who were more tool efficient could act more freely and thus did not have to follow the video instruction as closely as other participants. Some participants favored doing time-specific jumps to navigate the video, and others preferred the audio icons more. Where some were able to make do with very few commands, while others used more to complete the experiment. This suggests that the overall use of specific command types like audio icons might be more contingent on the user's particular strategy. Future studies should investigate any potential benefits from further segmenting and analyzing the different strategies used when undertaking how-to video tasks with voice control. This could potentially lead to pre-set navigation options, seeking to work more directly with specific user groups who actively use these augmented command capabilities to learn and understand how to improve them.

14.4 SUPPORT USER EXPECTATIONS

Imaginably, an optimal design choice for navigating timeline-based information is to be able to control by issuing direct commands like, ”go back to where I lost attention and resume from there.” However, current technology is far from reaching a level of sophistication that can support such interaction.

Another aspect that could potentially enhance the user experience is the system's ability to understand and remember context during the interaction. This applies to understanding if the user refers to

a previous command and analyzes the user's viewing pattern, i.e., they may have watched the same section multiple times. Supporting utterances like "Can I watch the parts I had difficulty with one more time?" might improve usability and affirm the user's view of the VUI as intelligent.

14.4.1 *Enable Users to Define Anchor Points*

Instead of the system trying to decide what is of interest to the user, a simple interaction feature is added, enabling user-created bookmarks. With these user-defined anchor points, the user can navigate to points in the video they find interesting. It should be noted, however, the inclusion of bookmarks in addition to, e.g., audio icons positioned at interest points is not mutually exclusionary but can co-exist within the same interface.

14.4.2 *Enable Globals in a VUI Context*

Another aspect of the VUI understanding of the context could manifest in the system supporting global commands that applies for the entire duration of the how-to video. For our study, a suitable example of a global pre-set navigation command is when a user utters, "I want you to pause every time the instructor uses the drill machine." In this way, the system initiates a global command that renders similar subsequent commands redundant. Moreover, with the user actively knowing the video will pause at these specific content-related points, their need to pay attention to the video at all times diminishes.

14.5 DESIGN RECOMMENDATIONS

Table 15 summarizes this chapter's presented suggestions. It should be noted this is a non-exhaustive list of potential design ideas to investigate in future studies on how to improve the usability of voice-driven devices for navigating how-to videos.

1.	Support higher-level user intents
2.	Allow user to make bookmarks
3.	Enables Globals commands (pre-set navigation strategies)
4.	Support context recognition
5.	Support progressive command refinement with multiple uses

Table 15: Overview of Design Recommendations Presented in This Chapter.

BIBLIOGRAPHY

Abbott, Ken (Jan. 2002). "VUI Design Principles and Techniques". In: pp. 87–103. ISBN: 978-1-893115-73-6. DOI: 10.1007/978-1-4302-0850-1_8.

Abdolrahmani, Ali, Ravi Kuber, and Stacy M. Branham (Oct. 2018). ""Siri Talks at You": An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind". In: *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '18. New York, NY, USA: Association for Computing Machinery, pp. 249–258. ISBN: 978-1-4503-5650-3. DOI: 10.1145/3234695.3236344. URL: https://doi.org/10.1145/3234695.3236344 (visited on 07/16/2021).

Ammari, Tawfiq et al. (Apr. 2019). "Music, Search, and IoT: How People (Really) Use Voice Assistants". In: *ACM Transactions on Computer-Human Interaction* 26.3, 17:1–17:28. ISSN: 1073-0516. DOI: 10.1145/3311956. URL: https://doi.org/10.1145/3311956 (visited on 08/27/2021).

Barko-Sherif, Sabrina, David Elsweiler, and Morgan Harvey (Mar. 2020). "Conversational Agents for Recipe Recommendation". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. CHIIR '20. New York, NY, USA: Association for Computing Machinery, pp. 73–82. ISBN: 978-1-4503-6892-6. DOI: 10.1145/3343413.3377967. URL: https://doi.org/10.1145/3343413.3377967 (visited on 09/30/2021).

Behrooz, Morteza et al. (2019). "AUGMENTING MUSIC LISTENING EXPERIENCES ON VOICE ASSISTANTS". en. In: p. 8.

Beneteau, Erin et al. (Mar. 2020). "Assumptions Checked: How Families Learn About and Use the Echo Dot". en. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous*

Technologies 4.1, pp. 1–23. ISSN: 2474-9567. DOI: 10.1145/3380993. URL: https://dl.acm.org/doi/10.1145/3380993 (visited on 03/24/2021).

Bentley, Frank et al. (Sept. 2018). "Understanding the Long-Term Use of Smart Speaker Assistants". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.3, 91:1–91:24. DOI: 10.1145/3264901. URL: https://doi.org/10.1145/3264901 (visited on 05/29/2021).

Blattner, Meera M., Denise A. Sumikawa, and Robert M. Greenberg (Mar. 1989). "Earcons and icons: their structure and common design principles". In: *Human-Computer Interaction* 4.1, pp. 11–44. ISSN: 0737-0024. DOI: 10.1207/s15327051hci0401_1. URL: https://doi.org/10.1207/s15327051hci0401_1 (visited on 09/30/2021).

Brooke, John (Nov. 1995). "SUS: A quick and dirty usability scale". In: *Usability Eval. Ind.* 189.

Cabral, João Paulo and Gerard Bastiaan Remijn (July 2019). "Auditory icons: Design and physical characteristics". en. In: *Applied Ergonomics* 78, pp. 224–239. ISSN: 00036870. DOI: 10.1016/j.apergo.2019.02.008. URL: https://linkinghub.elsevier.com/retrieve/pii/S0003687019300511 (visited on 10/02/2021).

Chang, Minsuk (Oct. 2019). "Data Structures for Designing Interactions with Contextual Task Support". en. In: *The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology*. New Orleans LA USA: ACM, pp. 142–145. ISBN: 978-1-4503-6817-9. DOI: 10.1145/3332167.3356874. URL: https://dl.acm.org/doi/10.1145/3332167.3356874 (visited on 02/14/2022).

Chang, Minsuk, Mina Huh, and Juho Kim (May 2021). "RubySlippers: Supporting Content-based Voice Navigation for How-to Videos". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 97. New York, NY, USA: Association for Computing Machinery, pp. 1–14. ISBN: 978-1-4503-8096-6. URL: https://doi.org/10.1145/3411764.3445131 (visited on 11/03/2021).

Chang, Minsuk, Oliver Wang, et al. (May 2019). "How to Design Voice Based Navigation for How-To Videos". en. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, pp. 1–11. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300931. URL: https://dl.acm.org/doi/10.1145/3290605.3300931 (visited on 06/19/2022).

Chen, Chia-Chen et al. (Feb. 2014). "A smart assistant toward product-awareness shopping". In: *Personal and Ubiquitous Computing* 18.2, pp. 339–349. ISSN: 1617-4909. DOI: 10.1007/s00779-013-0649-z. URL: https://doi.org/10.1007/s00779-013-0649-z (visited on 08/17/2021).

Cheng, Yi et al. (June 2018). "Why doesn't it work? voice-driven interfaces and young children's communication repair strategies". In: *Proceedings of the 17th ACM Conference on Interaction Design and Children*. IDC '18. New York, NY, USA: Association for Computing Machinery, pp. 337–348. ISBN: 978-1-4503-5152-2. DOI: 10.1145/3202185.3202749. URL: https://doi.org/10.1145/3202185.3202749 (visited on 08/27/2021).

Cho, Janghee (Apr. 2018). "Mental Models and Home Virtual Assistants (HVAs)". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI EA '18. New York, NY, USA: Association for Computing Machinery, pp. 1–6. ISBN: 978-1-4503-5621-3. DOI: 10.1145/3170427.3180286. URL: https://doi.org/10.1145/3170427.3180286 (visited on 10/01/2021).

Chung, David et al. (Dec. 2020). "Designing Auditory Experiences for Technology Imagination". In: *32nd Australian Conference on Human-Computer Interaction*. OzCHI '20. New York, NY, USA: Association for Computing Machinery, pp. 682–686. ISBN: 978-1-4503-8975-4. DOI: 10.1145/3441000.3441025. URL: https://doi.org/10.1145/3441000.3441025 (visited on 11/08/2021).

Cohen, Michael H., James P. Giangola, and Jennifer Balogh (2004). *Voice User Interface Design*. USA: Addison Wesley Longman Publishing Co., Inc. ISBN: 978-0-321-18576-1.

Corbett, Eric and Astrid Weber (Sept. 2016). "What can I say? addressing user experience challenges of a mobile voice user interface for accessibility". In: *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '16. New York, NY, USA: Association for Computing Machinery, pp. 72–82. ISBN: 978-1-4503-4408-1. DOI: 10.1145/2935334.2935386. URL: https://doi.org/10.1145/2935334.2935386 (visited on 05/15/2021).

Cordasco, Gennaro et al. (Nov. 2014). "Assessing Voice User Interfaces: The vassist system prototype". In: *2014 5th IEEE Conference on Cognitive Infocommunications (CogInfoCom)*, pp. 91–96. DOI: 10.1109/CogInfoCom.2014.7020425.

Crockford, Chris and Harry Agius (Apr. 2006). "An empirical investigation into user navigation of digital video using the VCR-like control set". en. In: *International Journal of Human-Computer Studies* 64.4, pp. 340–355. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2005.08.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S1071581905001631 (visited on 06/19/2022).

Csapó, Ádám and György Wersényi (Dec. 2013). "Overview of auditory representations in human-machine interfaces". In: *ACM Computing Surveys* 46.2, 19:1–19:23. ISSN: 0360-0300. DOI: 10.1145/2543581.2543586. URL: https://doi.org/10.1145/2543581.2543586 (visited on 09/16/2021).

Dragicevic, Pierre et al. (2008). "Video browsing by direct manipulation". en. In: *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*. Florence, Italy: ACM Press, p. 237. ISBN: 978-1-60558-011-1. DOI: 10.1145/1357054.1357096. URL: http://portal.acm.org/citation.cfm?doid=1357054.1357096 (visited on 06/19/2022).

Druga, Stefania et al. (Mar. 2019). "Inclusive AI literacy for kids around the world". en. In: *Proceedings of FabLearn 2019*. New York NY USA: ACM, pp. 104–111. ISBN: 978-1-4503-6244-3. DOI:

10.1145/3311890.3311904. URL: https://dl.acm.org/doi/10.1145/3311890.3311904 (visited on 02/18/2021).

Furqan, Anushay, Chelsea Myers, and Jichen Zhu (May 2017). "Learnability through Adaptive Discovery Tools in Voice User Interfaces". In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. CHI EA '17. New York, NY, USA: Association for Computing Machinery, pp. 1617–1623. ISBN: 978-1-4503-4656-6. DOI: 10.1145/3027063.3053166. URL: https://doi.org/10.1145/3027063.3053166 (visited on 03/24/2021).

Gaver, William (Mar. 1989). "The SonicFinder: An Interface That Uses Auditory Icons". en. In: *Human-Computer Interaction* 4.1, pp. 67–94. ISSN: 0737-0024. DOI: 10.1207/s15327051hci0401_3. URL: http://www.tandfonline.com/doi/abs/10.1207/s15327051hci0401_3 (visited on 09/16/2021).

Ghosh, Debjyoti et al. (Apr. 2018). "Assessing the Utility of the System Usability Scale for Evaluating Voice-based User Interfaces". In: *Proceedings of the Sixth International Symposium of Chinese CHI*. ChineseCHI '18. New York, NY, USA: Association for Computing Machinery, pp. 11–15. ISBN: 978-1-4503-6508-6. DOI: 10.1145/3202667.3204844. URL: https://doi.org/10.1145/3202667.3204844 (visited on 07/16/2021).

Habler, Florian, Marco Peisker, and Niels Henze (Nov. 2019). "Differences between smart speakers and graphical user interfaces for music search considering gender effects". en. In: *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*. Pisa Italy: ACM, pp. 1–7. ISBN: 978-1-4503-7624-2. DOI: 10.1145/3365610.3365627. URL: https://dl.acm.org/doi/10.1145/3365610.3365627 (visited on 06/15/2022).

Harris, Randy Allen (Jan. 2005). "Chapter 1 - Introduction". en. In: *Voice Interaction Design*. Ed. by Randy Allen Harris. San Francisco: Morgan Kaufmann, pp. 3–31. ISBN: 978-1-55860-768-2. DOI: 10.1016/B978-155860768-2/50001-3. URL: https://www.sciencedirect.com/science/article/pii/B9781558607682500013 (visited on 10/15/2021).

Hart, Sandra G. (Oct. 2006). "Nasa-Task Load Index (NASA-TLX); 20 Years Later". en. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50.9. Publisher: SAGE Publications Inc, pp. 904–908. ISSN: 2169-5067. DOI: 10.1177/154193120605000909. URL: https://doi.org/10.1177/154193120605000909 (visited on 06/15/2022).

Hart, Sandra G. and Lowell E. Staveland (Jan. 1988). "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". en. In: *Advances in Psychology*. Ed. by Peter A. Hancock and Najmedin Meshkati. Vol. 52. Human Mental Workload. North-Holland, pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9. URL: https://www.sciencedirect.com/science/article/pii/S0166411508623869 (visited on 06/15/2022).

Hoffmann, Fabian et al. (Nov. 2019). "User-defined interaction for smart homes: voice, touch, or mid-air gestures?" en. In: *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*. Pisa Italy: ACM, pp. 1–7. ISBN: 978-1-4503-7624-2. DOI: 10.1145/3365610.3365624. URL: https://dl.acm.org/doi/10.1145/3365610.3365624 (visited on 12/15/2021).

Hornbæk, Kasper and Javier A Bargas-Avila (2011). "Old Wine in New Bottles or Novel Challenges? A Critical Analysis of Empirical Studies of User Experience". en. In: p. 10.

Huyghe, Jonathan, Jan Derboven, and Dirk De Grooff (Oct. 2014). "ALADIN: demo of a multimodal adaptive voice interface". In: *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*. NordiCHI '14. New York, NY, USA: Association for Computing Machinery, pp. 1035–1038. ISBN: 978-1-4503-2542-4. DOI: 10.1145/2639189.2670269. URL: https://doi.org/10.1145/2639189.2670269 (visited on 06/01/2021).

Ilango, V., M. Anand Shankar Raja, and V. R. Uma (Feb. 2021). "Multi agent-based framework for interactive learning system with conversational user interface for visually and speech impaired". English. In: *IOP Conference Series. Materials Science and Engineering* 1070.1. Place: Bristol, United Kingdom Publisher: IOP Publishing. ISSN: 17578981. DOI: http://dx.doi.org.

ep.fjernadgang.kb.dk/10.1088/1757-899X/1070/1/012058. URL: http://www.proquest.com/docview/2513047127/abstract/778CFBF0A5464936PQ/1 (visited on 06/16/2021).

Jenkins, la111es J. (1985). "Acoustic Information for Objects, Places. and Events Introduction". In: *Persistence and Change*. Num Pages: 24. Psychology Press. ISBN: 978-0-203-78144-9.

Jeon, Myounghoon et al. (Jan. 2015). "Menu Navigation With In-Vehicle Technologies: Auditory Menu Cues Improve Dual Task Performance, Preference, and Workload". In: *International Journal of Human-Computer Interaction* 31.1. Publisher: Taylor & Francis Ltd, pp. 1–16. ISSN: 10447318. DOI: 10.1080/10447318.2014.925774. URL: https://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=99963694&site=ehost-live (visited on 09/09/2021).

Kim, Yea-Seul et al. (May 2019). "Vocal Shortcuts for Creative Experts". en. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, pp. 1–14. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300562. URL: https://dl.acm.org/doi/10.1145/3290605.3300562 (visited on 02/10/2022).

Kirschthaler, Philipp, Martin Porcheron, and Joel E. Fischer (July 2020). "What Can I Say?: Effects of Discoverability in VUIs on Task Performance and User Experience". en. In: *Proceedings of the 2nd Conference on Conversational User Interfaces*. Bilbao Spain: ACM, pp. 1–9. ISBN: 978-1-4503-7544-3. DOI: 10.1145/3405755.3406119. URL: https://dl.acm.org/doi/10.1145/3405755.3406119 (visited on 05/16/2021).

Klein, Laura (Oct. 2015). "Design for Voice Interfaces". en. In: p. 37.

Klemmer, Scott R. et al. (Nov. 2000). "Suede: a Wizard of Oz prototyping tool for speech user interfaces". In: *Proceedings of the 13th annual ACM symposium on User interface software and technology*. UIST '00. New York, NY, USA: Association for Computing Machinery, pp. 1–10. ISBN: 978-1-58113-212-0. DOI: 10.1145/354401.354406. URL: https://doi.org/10.1145/354401.354406 (visited on 08/20/2021).

Knees, Peter, Markus Schedl, and Rebecca Fiebrink (Mar. 2019). "Intelligent music interfaces for listening and creation". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*. IUI '19. New York, NY, USA: Association for Computing Machinery, pp. 135–136. ISBN: 978-1-4503-6673-1. DOI: 10.1145/3308557.3313110. URL: https://doi.org/10.1145/3308557.3313110 (visited on 02/27/2021).

Kocabalil, A. Baki, Liliana Laranjo, and Enrico Coiera (July 2018). "Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires". en. In: DOI: 10.14236/ewic/HCI2018.21. URL: https://scienceopen.com/document?vid=79db6ce4-7488-4de5-8fae-201702d73fbb (visited on 09/25/2021).

Laput, Gierad P. et al. (Apr. 2013). "PixelTone: a multimodal interface for image editing". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: Association for Computing Machinery, pp. 2185–2194. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2481301. URL: https://doi.org/10.1145/2470654.2481301 (visited on 11/05/2021).

Large, David R., Gary Burnett, and Leigh Clark (Sept. 2019). "Lessons from Oz: design guidelines for automotive conversational user interfaces". In: *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings*. AutomotiveUI '19. New York, NY, USA: Association for Computing Machinery, pp. 335–340. ISBN: 978-1-4503-6920-6. DOI: 10.1145/3349263.3351314. URL: https://doi.org/10.1145/3349263.3351314 (visited on 06/15/2021).

Lee, Sunok, Minji Cho, and Sangsu Lee (Sept. 2020). "What If Conversational Agents Became Invisible? Comparing Users' Mental Models According to Physical Entity of AI Speaker". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4.3, 88:1–88:24. DOI: 10.1145/3411840. URL: https://doi.org/10.1145/3411840 (visited on 08/17/2021).

Liu, Xingyu et al. (May 2021). "What Makes Videos Accessible to Blind and Visually Impaired People?" In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 272. New York, NY, USA: Association for Computing Machinery, pp. 1–14. ISBN: 978-1-4503-8096-6. URL: https://doi.org/10.1145/3411764.3445233 (visited on 11/08/2021).

Luger, Ewa and Abigail Sellen (May 2016). ""Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. New York, NY, USA: Association for Computing Machinery, pp. 5286–5297. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858288. URL: https://doi.org/10.1145/2858036.2858288 (visited on 03/07/2021).

Ma, Xiao and Ariel Liu (July 2020). "Challenges in Supporting Exploratory Search through Voice Assistants". en. In: *Proceedings of the 2nd Conference on Conversational User Interfaces*. Bilbao Spain: ACM, pp. 1–3. ISBN: 978-1-4503-7544-3. DOI: 10.1145/3405755.3406152. URL: https://dl.acm.org/doi/10.1145/3405755.3406152 (visited on 12/15/2021).

Matejka, Justin, Tovi Grossman, and George Fitzmaurice (May 2012). "Swift: reducing the effects of latency in online video scrubbing". en. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Austin Texas USA: ACM, pp. 637–646. ISBN: 978-1-4503-1015-4. DOI: 10.1145/2207676.2207766. URL: https://dl.acm.org/doi/10.1145/2207676.2207766 (visited on 04/13/2022).

Matejka, Justin, Tovi Grossman, and George Fitzmaurice (Apr. 2013). "Swifter: improved online video scrubbing". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '13. New York, NY, USA: Association for Computing Machinery, pp. 1159–1168. ISBN: 978-1-4503-1899-0. DOI: 10.1145/2470654.2466149. URL: https://doi.org/10.1145/2470654.2466149 (visited on 04/13/2022).

Mittal, Deepika (Mar. 2020). *How Nielsen's 10 usability heuristics apply to VUI Design*. en. URL: https://uxdesign.cc/10-usability-heuristics-for-voice-user-interface-design-69ad9ea4f166 (visited on 09/30/2021).

Murad, Christine et al. (Sept. 2018). "Design guidelines for hands-free speech interaction". en. In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. Barcelona Spain: ACM, pp. 269–276. ISBN: 978-1-4503-5941-2. DOI: 10.1145/3236112.3236149. URL: https://dl.acm.org/doi/10.1145/3236112.3236149 (visited on 05/28/2021).

Myers, Chelsea et al. (Apr. 2018). "Patterns for How Users Overcome Obstacles in Voice User Interfaces". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. New York, NY, USA: Association for Computing Machinery, pp. 1–7. ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3173580. URL: https://doi.org/10.1145/3173574.3173580 (visited on 05/15/2021).

Myers, Chelsea M. (Mar. 2019). "Adaptive suggestions to increase learnability for voice user interfaces". en. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*. Marina del Ray California: ACM, pp. 159–160. ISBN: 978-1-4503-6673-1. DOI: 10.1145/3308557.3308727. URL: https://dl.acm.org/doi/10.1145/3308557.3308727 (visited on 05/28/2021).

Myers, Chelsea M., Anushay Furqan, and Jichen Zhu (May 2019). "The Impact of User Characteristics and Preferences on Performance with an Unfamiliar Voice User Interface". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. New York, NY, USA: Association for Computing Machinery, pp. 1–9. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300277. URL: https://doi.org/10.1145/3290605.3300277 (visited on 05/15/2021).

Nguyen, Cuong and Feng Liu (Apr. 2015). "Making Software Tutorial Video Responsive". In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*.

CHI '15. New York, NY, USA: Association for Computing Machinery, pp. 1565–1568. ISBN: 978-1-4503-3145-6. DOI: 10.1145/2702123.2702209. URL: https://doi.org/10.1145/2702123.2702209 (visited on 06/25/2022).

Norman, Donald A. (Aug. 1998). *The Design of Everyday Things*. en. Cambridge, MA, USA: MIT Press. ISBN: 978-0-262-64037-4.

Norman, Donald A. (2002). *The Design of Everyday Things*. USA: Basic Books, Inc. ISBN: 978-0-465-06710-7.

Nouri, Elnaz et al. (Mar. 2020). "Step-wise Recommendation for Complex Task Support". en. In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. Vancouver BC Canada: ACM, pp. 203–212. ISBN: 978-1-4503-6892-6. DOI: 10.1145/3343413.3377964. URL: https://dl.acm.org/doi/10.1145/3343413.3377964 (visited on 02/14/2022).

Ostrowski, Anastasia K. et al. (Mar. 2021). "Small Group Interactions with Voice-User Interfaces: Exploring Social Embodiment, Rapport, and Engagement". In: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '21. New York, NY, USA: Association for Computing Machinery, pp. 322–331. ISBN: 978-1-4503-8289-2. DOI: 10.1145/3434073.3444655. URL: https://doi.org/10.1145/3434073.3444655 (visited on 06/15/2021).

Paikari, Elahe and André van der Hoek (May 2018). "A framework for understanding chatbots and their future". In: *Proceedings of the 11th International Workshop on Cooperative and Human Aspects of Software Engineering*. CHASE '18. New York, NY, USA: Association for Computing Machinery, pp. 13–16. ISBN: 978-1-4503-5725-8. DOI: 10.1145/3195836.3195859. URL: https://doi.org/10.1145/3195836.3195859 (visited on 03/06/2021).

Pavel, Amy, Dan B. Goldman, et al. (Nov. 2015). "SceneSkim: Searching and Browsing Movies Using Synchronized Captions, Scripts and Plot Summaries". In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. UIST '15. New York, NY, USA:

Association for Computing Machinery, pp. 181–190. ISBN: 978-1-4503-3779-3. DOI: 10.1145/2807442.2807502. URL: https://doi.org/10.1145/2807442.2807502 (visited on 06/20/2022).

Pavel, Amy, Colorado Reed, et al. (Oct. 2014). "Video digests: a browsable, skimmable format for informational lecture videos". en. In: *Proceedings of the 27th annual ACM symposium on User interface software and technology*. Honolulu Hawaii USA: ACM, pp. 573–582. ISBN: 978-1-4503-3069-5. DOI: 10.1145/2642918.2647400. URL: https://dl.acm.org/doi/10.1145/2642918.2647400 (visited on 06/20/2022).

Pearl, Cathy (2017). *Designing voice user interfaces: principles of conversational experiences /*. eng. First edition. Beijing: O'Reilly. ISBN: 978-1-4919-5541-3.

Pearl, Cathy (May 2019). "Using Voice Interfaces to Make Products More Inclusive". In: *Harvard Business Review*. Section: Technology and analytics. ISSN: 0017-8012. URL: https://hbr.org/2019/05/using-voice-interfaces-to-make-products-more-inclusive (visited on 09/25/2021).

Pongnumkul, Suporn et al. (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*. UIST '11. New York, NY, USA: Association for Computing Machinery, pp. 135–144. ISBN: 978-1-4503-0716-1. DOI: 10.1145/2047196.2047213. URL: https://doi.org/10.1145/2047196.2047213 (visited on 06/25/2022).

Porcheron, Martin, Joel E. Fischer, and Stuart Reeves (Jan. 2021). "Pulling Back the Curtain on the Wizards of Oz". In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW3, 243:1–243:22. DOI: 10.1145/3432942. URL: https://doi.org/10.1145/3432942 (visited on 09/01/2021).

Porcheron, Martin, Joel E. Fischer, and Michel Valstar (July 2020). "NottReal: A Tool for Voice-based Wizard of Oz studies". In: *Proceedings of the 2nd Conference on Conversational User Interfaces*. CUI '20. New York, NY, USA: Association for Computing Machinery, pp. 1–3. ISBN:

978-1-4503-7544-3. DOI: 10.1145/3405755.3406168. URL: https://doi.org/10.1145/3405755.3406168 (visited on 06/15/2021).

Purington, Amanda et al. (May 2017). ""Alexa is my new BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo". en. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Denver Colorado USA: ACM, pp. 2853–2859. ISBN: 978-1-4503-4656-6. DOI: 10.1145/3027063.3053246. URL: https://dl.acm.org/doi/10.1145/3027063.3053246 (visited on 05/28/2021).

Pyae, Aung and Tapani N. Joelsson (Sept. 2018). "Investigating the usability and user experiences of voice user interface: a case of Google home smart speaker". In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. MobileHCI '18. New York, NY, USA: Association for Computing Machinery, pp. 127–131. ISBN: 978-1-4503-5941-2. DOI: 10.1145/3236112.3236130. URL: https://doi.org/10.1145/3236112.3236130 (visited on 05/15/2021).

Radlinski, Filip and Nick Craswell (Mar. 2017). "A Theoretical Framework for Conversational Search". In: *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. CHIIR '17. New York, NY, USA: Association for Computing Machinery, pp. 117–126. ISBN: 978-1-4503-4677-1. DOI: 10.1145/3020165.3020183. URL: https://doi.org/10.1145/3020165.3020183 (visited on 08/20/2021).

Sahijwani, Harshita, Jason Ingyu Choi, and Eugene Agichtein (Mar. 2020). "Would You Like to Hear the News? Investigating Voice-Based Suggestions for Conversational News Recommendation". In: *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. CHIIR '20. New York, NY, USA: Association for Computing Machinery, pp. 437–441. ISBN: 978-1-4503-6892-6. DOI: 10.1145/3343413.3378013. URL: https://doi.org/10.1145/3343413.3378013 (visited on 11/08/2021).

Schnelle-Walka, Dirk (July 2010). "A pattern language for error management in voice user interfaces". In: *Proceedings of the 15th European Conference on Pattern Languages of Programs*. EuroPLoP

'10. New York, NY, USA: Association for Computing Machinery, pp. 1–23. ISBN: 978-1-4503-0259-3. DOI: 10.1145/2328909.2328920. URL: https://doi.org/10.1145/2328909.2328920 (visited on 06/01/2021).

Schnelle-Walka, Dirk and Fernando Lyardet (Jan. 2006). "Voice User Interface Design Patterns." In: pp. 287–316.

Srinivasan, Arjun et al. (Mar. 2019). "Discovering natural language commands in multimodal interfaces". In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. IUI '19. New York, NY, USA: Association for Computing Machinery, pp. 661–672. ISBN: 978-1-4503-6272-6. DOI: 10.1145/3301275.3302292. URL: https://doi.org/10.1145/3301275.3302292 (visited on 11/05/2021).

Suh, Hyewon, Myounghoon Jeon, and Bruce N. Walker (Sept. 2012). "Spearcons Improve Navigation Performance and Perceived Speediness in Korean Auditory Menus". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56.1. Publisher: SAGE Publications Inc, pp. 1361–1365. ISSN: 2169-5067. DOI: 10.1177/1071181312561390. URL: https://doi.org/10.1177/1071181312561390 (visited on 09/09/2021).

Thymé-Gobbel, Ann and Charles Jankowski (2021a). "Creating Secure Personalized Experiences". en. In: *Mastering Voice Interfaces: Creating Great Voice Apps for Real Users*. Ed. by Ann Thymé-Gobbel and Charles Jankowski. Berkeley, CA: Apress, pp. 545–594. ISBN: 978-1-4842-7005-9. DOI: 10.1007/978-1-4842-7005-9_15. URL: https://doi.org/10.1007/978-1-4842-7005-9_15 (visited on 08/29/2021).

Thymé-Gobbel, Ann and Charles Jankowski (2021b). "Define Through Discovery: Building What, How, and Why for Whom". en. In: *Mastering Voice Interfaces: Creating Great Voice Apps for Real Users*. Ed. by Ann Thymé-Gobbel and Charles Jankowski. Berkeley, CA: Apress, pp. 97–136. ISBN: 978-1-4842-7005-9. DOI: 10.1007/978-1-4842-7005-9_4. URL: https://doi.org/10.1007/978-1-4842-7005-9_4 (visited on 08/29/2021).

Thymé-Gobbel, Ann and Charles Jankowski (2021c). “From Discovery to UX and UI: Tools of Voice Design”. en. In: *Mastering Voice Interfaces: Creating Great Voice Apps for Real Users*. Ed. by Ann Thymé-Gobbel and Charles Jankowski. Berkeley, CA: Apress, pp. 137–171. ISBN: 978-1-4842-7005-9. DOI: 10.1007/978-1-4842-7005-9_5. URL: https://doi.org/10.1007/978-1-4842-7005-9_5 (visited on 08/29/2021).

Thymé-Gobbel, Ann and Charles Jankowski (2021d). “Helping Users Succeed Through Consistency”. en. In: *Mastering Voice Interfaces: Creating Great Voice Apps for Real Users*. Ed. by Ann Thymé-Gobbel and Charles Jankowski. Berkeley, CA: Apress, pp. 315–352. ISBN: 978-1-4842-7005-9. DOI: 10.1007/978-1-4842-7005-9_9. URL: https://doi.org/10.1007/978-1-4842-7005-9_9 (visited on 08/29/2021).

Trippas, Johanne R. (Mar. 2021). “Spoken conversational search: audio-only interactive information retrieval”. In: *ACM SIGIR Forum* 53.2, pp. 106–107. ISSN: 0163-5840. DOI: 10.1145/3458553.3458570. URL: https://doi.org/10.1145/3458553.3458570 (visited on 08/27/2021).

Tuncer, Sylvaine, Barry Brown, and Oskar Lindwall (Apr. 2020). “On Pause: How Online Instructional Videos are Used to Achieve Practical Tasks”. en. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu HI USA: ACM, pp. 1–12. ISBN: 978-1-4503-6708-0. DOI: 10.1145/3313831.3376759. URL: https://dl.acm.org/doi/10.1145/3313831.3376759 (visited on 02/17/2022).

Walker, Ashley et al. (July 2001). “Diary in the Sky: A Spatial Audio Display for a Mobile Calendar”. In: ISSN: 978-1-85233-515-1. DOI: 10.1007/978-1-4471-0353-0_33.

Walker, Bruce N. et al. (Feb. 2013). “Spearcons (Speech-Based Earcons) Improve Navigation Performance in Advanced Auditory Menus”. In: *Human Factors* 55.1. Publisher: SAGE Publications Inc, pp. 157–182. ISSN: 0018-7208. DOI: 10.1177/0018720812450587. URL: https://doi.org/10.1177/0018720812450587 (visited on 09/09/2021).

Walker, Marilyn A. et al. (Jan. 1998). "What can I say? evaluating a spoken language interface to Email". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '98. USA: ACM Press/Addison-Wesley Publishing Co., pp. 582–589. ISBN: 978-0-201-30987-4. DOI: 10.1145/274644.274722. URL: https://doi.org/10.1145/274644.274722 (visited on 05/29/2021).

Werner, Steffen et al. (Sept. 2015). "Can VoiceScapes Assist in Menu Navigation?" en. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 59.1. Publisher: SAGE Publications Inc, pp. 1095–1099. ISSN: 2169-5067. DOI: 10.1177/1541931215591157. URL: https://doi.org/10.1177/1541931215591157 (visited on 10/03/2021).

Westerlund, Daniel (Aug. 2021). *How to Deal with Cognitive Load in Voice Design*. en-US. Section: UX Design. URL: https://careerfoundry.com/en/blog/ux-design/voice-ui-design-and-cognitive-load/ (visited on 10/20/2021).

Williams, Adam S. and Francisco R. Ortega (Sept. 2020). "Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation". In: *arXiv:2009.06591 [cs]*. arXiv: 2009.06591. URL: http://arxiv.org/abs/2009.06591 (visited on 12/15/2021).

Wu, Yunhan et al. (July 2020). "Mental Workload and Language Production in Non-Native Speaker IPA Interaction". en. In: *Proceedings of the 2nd Conference on Conversational User Interfaces*. arXiv:2006.06331 [cs], pp. 1–8. DOI: 10.1145/3405755.3406118. URL: http://arxiv.org/abs/2006.06331 (visited on 06/16/2022).

Yalla, Pavani and Bruce N. Walker (Oct. 2008). "Advanced auditory menus: design and evaluation of auditory scroll bars". In: *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. Assets '08. New York, NY, USA: Association for Computing Machinery, pp. 105–112. ISBN: 978-1-59593-976-0. DOI: 10.1145/1414471.1414492. URL: https://doi.org/10.1145/1414471.1414492 (visited on 09/16/2021).

Yang, Xi and Marco Aurisicchio (May 2021). “Designing Conversational Agents: A Self-Determination Theory Approach”. en. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, pp. 1–16. ISBN: 978-1-4503-8096-6. DOI: 10.1145/3411764.3445445. URL: https://dl.acm.org/doi/10.1145/3411764.3445445 (visited on 09/28/2021).

Yankelovich, Nicole (Dec. 1996). “How do users know what to say?” In: *Interactions* 3.6, pp. 32–43. ISSN: 1072-5520. DOI: 10.1145/242485.242500. URL: https://doi.org/10.1145/242485.242500 (visited on 09/09/2021).

Yankelovich, Nicole, Gina-Anne Levow, and Matt Marx (May 1995). “Designing SpeechActs: issues in speech user interfaces”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’95. USA: ACM Press/Addison-Wesley Publishing Co., pp. 369–376. ISBN: 978-0-201-84705-5. DOI: 10.1145/223904.223952. URL: https://doi.org/10.1145/223904.223952 (visited on 05/29/2021).

Yarmand, Matin et al. (May 2019). “”Can you believe [1:21]?!”: Content and Time-Based Reference Patterns in Video Comments”. en. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Glasgow Scotland Uk: ACM, pp. 1–12. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300719. URL: https://dl.acm.org/doi/10.1145/3290605.3300719 (visited on 02/10/2022).

Zhang, Justine, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil (Oct. 2020). “Quantifying the Causal Effects of Conversational Tendencies”. In: *Proceedings of the ACM on Human-Computer Interaction* 4.CSCW2, 131:1–131:24. DOI: 10.1145/3415202. URL: https://doi.org/10.1145/3415202 (visited on 08/17/2021).

Zhi, Qiyu et al. (Mar. 2018). “VisPod: Content-Based Audio Visual Navigation”. en. In: *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*. Tokyo Japan: ACM, pp. 1–2. ISBN: 978-1-4503-5571-1. DOI: 10.1145/3180308.3180318. URL: https://dl.acm.org/doi/10.1145/3180308.3180318 (visited on 02/08/2022).

Zhong, Yu et al. (2014). "*JustSpeak*: enabling universal voice control on Android". en. In: *Proceedings of the 11th Web for All Conference on - W4A '14*. Seoul, Korea: ACM Press, pp. 1–4. ISBN: 978-1-4503-2651-3. DOI: 10.1145/2596695.2596720. URL: http://dl.acm.org/citation.cfm?doid=2596695.2596720 (visited on 05/28/2021).