"Rewind to the Jiggling Meat Part": Understanding Voice Control of Instructional Videos in Everyday Tasks

Yaxi Zhao
yaxi.zhao@mail.utoronto.ca
University of Toronto
Toronto, Ontario, Canada

Razan Jaber
razan@dsv.su.se
Stockholm University
Stockholm, Sweden

Donald McMillan
donald.mcmillan@dsv.su.se
Stockholm University
Stockholm, Sweden

Cosmin Munteanu
cosmin@taglab.ca
University of Toronto Mississauga
Mississauga, Ontario, Canada

ABSTRACT

Voice interaction has long been envisioned as enabling users to transform physical interaction into hands-free, such as allowing fine-grained control of instructional videos without physically disengaging from the task at hand. While significant engineering advances have brought us closer to this ideal, we do not fully understand the user requirements for voice interactions that should be supported in such contexts. This paper presents an ecologically-valid wizard-of-oz elicitation study exploring realistic user requirements for an ideal instructional video playback control while cooking. Through the analysis of the issued commands and performed actions during this non-linear and complex task, we identify (1) patterns of command formulation, (2) challenges for design, and (3) how task and voice-based commands are interwoven in real-life. We discuss implications for the design and research of voice interactions for navigating instructional videos while performing complex tasks.

CCS CONCEPTS

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

KEYWORDS

Conversational Interaction, Non-Linear Instructional Video, Voice-Based Navigation, Wizard-of-Oz

ACM Reference Format:
Yaxi Zhao, Razan Jaber, Donald McMillan, and Cosmin Munteanu. 2022. "Rewind to the Jiggling Meat Part": Understanding Voice Control of Instructional Videos in Everyday Tasks. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29–May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3491102.3502036

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '22, April 29–May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3502036

1 INTRODUCTION

When people learn physical tasks by watching and following online videos, such as cooking and home repair, voice interaction has been envisioned to facilitate task completion by allowing learners' hands to remain in the task while still navigating through the instructional videos.

Learning how to cook pertains to a complex non-linear environment that requires working memory to put in the correct amount of ingredients in a certain order, tacit knowledge of how to manipulate the ingredients into certain states (e.g., kneading and cutting), domain knowledge to solve problems, and multi-tasking skills to coordinate the preparation of different parts to be completed at the same time. As such, to follow a cooking instructional video, users often engage in non-linear navigation behavior, often jumping back and forth to understand the content [21], check the status, control the pace, navigate to a given point of interest [62], and potentially search for additional information.

Cooking instructional videos represent a challenge and an opportunity for voice user interfaces (VUIs). Commercial voice assistants with screens, such as Amazon Echo and Google Nest, allow users to navigate through professionally-annotated text-and-picture recipes with a variety of voice commands. This annotation enables the user to off-load some cognitive burden, such as repeating a step, answering questions about ingredients, and jumping back and forth between steps with content-based commands (e.g., "forward to adding the egg"). However, when it comes to the millions of user-generated instructional videos on YouTube, current use for voice merely emulates the navigation found in graphical user interfaces for playback control (e.g., rewind and pause), thus insufficiently supporting non-linear instructional video navigation.

Yet such video navigation represents an interesting benchmark case for advances in the design of voice interfaces, as these videos are highly interactive and rich in content, yet natural to use. Additionally, it represents a more realistic goal for improving the engineering aspects of voice interaction than the more common benchmarks of dictation or query-answering dialogue systems.

While much engineering efforts and technology advancement has been made in voice recognition and computational algorithms for VUIs that make the existing features more efficient and accurate, understanding of the human side of spoken interaction with the

technology remains limited [3]. As such, VUIs are only used for simple tasks and often disappoint users [38]. Luger and Sellen [38] highlighted the limited functionality of existing commercially-available voice interfaces and how it causes a gulf between their capabilities and the users' expectations. A recurring theme of research into VUIs is that "user expectations of CA[conversational agent] systems remain far from the practical realities of use" [38]. Users come with a set of expectations about how spoken conversation should work outside the capability of today's speech technology [31]. The design of speech technologies encompasses several dimensions, making it entangled and complex for designers to create comparatively privileged technologies. Investigating speech systems and interfaces design issues, social use, and different possibilities of interactions become crucial in designing such systems.

This study investigates user requirements for designing voice-based playback control for non-linear instructional videos found in cooking. To understand user expectations and requirements in a natural environment yet without the limitations of current commercial speech technology, we conducted an ecologically-valid Wizard-of-Oz study to answer the following research questions (RQs):

RQ1 What are the contexts of user interactions with voice-based navigation in a natural environment (e.g., home-based cooking) following instructional videos?
RQ2 What types of voice-based navigation needs do users have for non-linear instructional videos in an ideal system?
RQ3 What are the challenges and opportunities for designing such voice control?

Through remotely observing how 10 participants (of varied experience in VUIs and cooking) used a wizard-controlled voice playback system to navigate an instructional video to make dumplings from scratch and follow-up interviews, we highlight the technological and social context of using voice control for non-linear tasks (cooking) in the kitchen, patterns of time- and content-based query formulation for such interaction, and challenges and opportunities for designing such voice controls.

Our paper makes the following contributions:

- Empirical insights of contextualized user needs and query formulation patterns from an ecologically-valid elicitation study
- A synthesis of command response challenges and potential solutions for voice control of non-linear instructional videos
- Opportunities for future research and design to better support voice-based non-linear video control for everyday tasks

2 RELATED WORK

2.1 Instructional Video Interaction

Online instructional videos have become a popular means for delivering 'how-to' information to teach new skills. How-to videos available on YouTube and other online resources enable millions of users to learn procedural tasks such as applying makeup, cooking [29, 33], playing music, and fixing bicycles [58]. The structure of instructional videos generally requires the viewer to watch one step of a task, complete that step, and then proceed to the next. To support users to navigate videos easily, Tankaka et al [56] proposed a chart-style video playback interface called PlayWatch that visualized indices as points on a 2D chart. The user could jump directly to a scene indicated by an index point by simply clicking on it, which was found to be highly effective in accessing scenes by precise keywords. Dragivevic et al. [16] presented a method for browsing videos by directly dragging their content, allowing users to control video playback by moving objects of interest along their visual trajectory. The results showed the benefits of direct manipulation through focusing on visual content rather than time. Tuncer et al. [58] documented how pausing was used to coordinate and interweave watching and doing to find task objects, turning actions, keeping up, and fixing problems. The study discussed the role of basic user interface functions in complex tasks involving different forms of engagement with the physical world and with screen-based activity.

Many of the instructional videos that teach physical tasks such as cooking involve interaction with real-world objects. When following the instructions in the videos, users usually need their hands both to execute the tasks and to control the video to manage the delivery of the step-by-step instructions. Researchers have recently proposed voice-based interactions to enhance navigation through instructional videos of physical tasks. Cuendet et al. [15] used speech in combination with a smartphone touchscreen to design an instructional video repository for farmers. The interface was completely text-free, with a "push-to-talk" button to allow users to find the informational video by selecting images or stating the name of their choice on the current level of the navigation tree. Chang et al. [9] conducted a set of experiments through which they examined how different user navigation objectives and intentions affect their word choices in voice command utterances, and reported a lexicon of types of interactions and the motivating factors behind these commands. Their work focuses on the challenges and the opportunities for designing voice navigation interactions for how-to videos. Chang et al. [8] also studied how instructions are followed and developed a voice-controlled video pausing tool that allowed users to manage video while their hands were busy with the task. The system supported efficient content-based voice navigation through keyword-based queries.

These studies on navigating instructional videos with a voice-controlled interface have inspired us to further explore user needs when using voice interactions to follow 'how-to' videos such as cooking, a complex non-linear task. Cooking is a skill that consists of combining, mixing, processing, and handling ingredients using a variety of tools. Technologically understanding the individual parts of food preparation and how they depend on each other from recipes is an ongoing technical challenge [10, 28].

While physically supporting users in general cooking tasks is still a long way off [2], it is suggested that collaborative guidance in following recipes, rather than automating the human out of the interaction, is a more promising goal. One key problem that users of instructional videos face is that they need to perform the tasks shown in the video while navigating through it, particularly for a complex non-linear and time-sensitive task such as cooking. This implies that using a modality other than touch is necessary. It is necessary to understand how users would interact with an ideal

voice-based interface for this task, to design a better voice-based interface to interact with instructional videos,

Our work builds upon previous work of exploring the interaction with a voice-based interface for navigating with how-to videos, but specifically focusing on the requirements for designing idealized voice-based control for non-linear instructional videos.

2.2 VUI Design

Recently, there has been a growth in research that investigates interactions with speech interfaces in different settings [24, 25, 35, 52] and domains [32, 36, 42, 51, 59–61]. Many studies focus on how to be sensitive to the needs of users, and what an appropriate relationship between users and technology might be [14]. For example, the necessity for human-like communication strategies with systems has been the focus of many studies, as that may help people understand and interact with these systems, resulting in more trust and more frequent system use [30]. Fischer et al. [19] examined how the orientation towards progressivity in talk – keeping the conversation moving – might help us better understand and design voice interactions.

Several studies highlight a number of similar problems in speech agent interaction, including troubles situating the interaction in an ongoing conversation, difficulties activating and verifying activation, and challenges learning how to formulate and enunciate queries for devices [5, 13, 17, 34, 38]. When interacting with speech systems, humans must work to adapt their communication patterns to the needs of the conversational agent, rather than the other way around [26], similar to how humans adapt their communication styles and patterns to align with robots [45, 63] and computers [43, 44]. For example, Hara et al. [22] conducted a study with spoken language translation and found that people naturally adapted to the imperfect translation systems by changing how they spoke, and how they comprehended speech. Research continues to demonstrate that people shorten their sentences, use simplified language, and repeat themselves in attempts to be understood by voice interfaces [26, 38, 45]. For example, Cheng et al. [11] investigated the repair strategies that young children faced when they interacted with a game using voice interaction, and the system broke down. The authors showed that children used several strategies like repetition and spoke loudly. Similarly, Beneteau et al. [4] investigated the repair communication breakdowns with digital home assistants between family members. Porcheron et al. [48] suggested that future work on VUIs should, shift "from conversation design to [. . .] request/response design".

Using voice as a means to search holds various potential advantages [57], yet the inability of most conversational systems to manage the interaction and keep track of topics leads to confusion and awkwardness. With sound being an open channel, voice as an input mechanism is inherently insecure as it is prone to replay, sensitive to noise, and easy to impersonate [17, 18].

New users of speech-based interfaces, in general, are more satisfied with these products, complete their tasks in fewer interaction cycles, and are more likely to use the system again [20]. However, as a result of the ongoing interaction issues outlined above, users of voice interface technology often become frustrated and can fail to learn the technology's full capabilities or abandon its use altogether [13, 38].

3 INVESTIGATING NATURAL LANGUAGE VIDEO CONTROL

The surveyed literature highlighted several gaps and opportunities to understand user requirements for voice-based navigation for non-linear instructional videos. In the research presented in this paper, we aim to address such gaps, using cooking as an example application. Cooking is particularly interesting as it requires combining, mixing, and handling ingredients; thus users' hands are busy following the steps of the recipes in the correct order. As such, we conducted a remote Wizard-of-Oz study [27] with each participant to make Asian-style dumplings in which they followed a recipe video and navigated through it with voice commands. Dumpling-making was chosen because the nature of the tasks required the participants to get their hands dirty, providing a natural motivation to perform voice-based commands. In addition, the procedure for making dumplings is non-linear, in which the dough and the filling are prepared independently. To save time, a cook should strategically go back and forth between the dough and the filling preparation, which might require the participants to jump back and forward to different parts of the instructional video, making content-based navigation desirable. This Wizard-of-Oz set up within the context of a familiar, yet complex task, allows us to gain a more thorough understanding of user requirements, without the constraints on query formulation placed by the limitations of commercial speech technologies.

3.1 Participants

We recruited 10 participants (7 females, 3 males; average age: 31.5 years old, min: 14, max: 40). We encouraged the participants to self-categorize their cooking experience into four groups, namely minimal (e.g., never cooked alone), average, above average, and very experienced. Two reported having minimal experience, one average experience, five above average experience, and two very experienced. None of them had made Asian-style dumplings before. Five reported owning and interacting with a voice assistant, while the other five had not used one. Table 1 presents a detailed breakdown regarding the combination of cooking experience and past experience using VUIs.

3.2 Protocol

For the Wizard-of-Oz follow-along cooking session, participants were told that they were testing a newly designed voice navigation assistant with the capability of responding to both time-based (e.g., rewind 3 seconds) and content-based commands (e.g., go back to the filling ingredients) while following a recipe video they had not watched before.

We informed them that this was to be conducted in their own home. We provided a shopping list with suggested ingredients (and instructions that they could substitute some of them based on personal choice or food allergies) and tools to prepare before starting the Wizard-of-Oz session. During the session, they followed the video to make dumplings from scratch (i.e., making the dough,

Table 1: Participants' experience in cooking and VUIs

Participants ID	Experience in cooking	Experience in VUIs
P1	above average	yes
P2	above average	no
P3	above average	no
P4	very experienced	yes
P5	minimal	yes
P6	average	yes
P7	above average	no
P8	above average	no
P9	very experienced	no
P10	minimal	yes

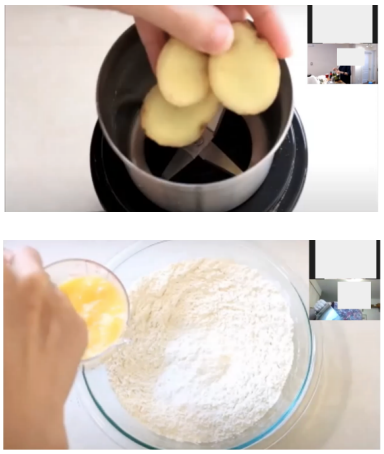

Figure 1: (left) P4 was adding filling ingredients while letting the video play. (right) P9 paused the video to follow what the instructor just did – adding egg-water mixture to the flour.

shaping it into a wrapper, preparing the filling, and then constructing two dumplings) – cooking and eating the results were beyond the scope of the trial. As such, all the requirements this places on the participants for navigating through the video and controlling playback were realistic.

All participants (or their guardians) reviewed and signed a document describing the details of the trial procedure, data collection, and analysis before the study as informed consent. All sessions were conducted using a password-protected Zoom room (Figure 1). At the beginning of each cooking session, the procedure was reiterated. A brief interview was conducted to verify demographic data, probe their experience in cooking and using voice assistants, and how they would typically approach learning to cook a new dish using online instructional videos.

Participants were encouraged to make any navigation-related voice commands that made sense to them. The wizard played the instructional video and shared their screen with each participant; in the meantime, the wizard's video and audio were turned off. Thus each participant would be able to watch the video and navigate it through the wizard (i.e., the perceived navigation system) using time-based and content-based commands of their choice. The wizard had a list of possible areas of commands and navigated the video on behalf of the participant. The participants were not told what exact commands were possible so that they were not limited to technical constraints.

When a participant issued a command towards the system, the wizard would pause the video to listen, similar to Google Nest. The video would play after the navigation task if the participant issued a jump command (rewind/forward). If a received command were out of the planned list, the wizard would play a pre-recorded error message that the system did not understand the command, leaving the users to retry the interaction much like when a commercial voice assistant does not understand a command. Unlike using existing commercial voice assistants, the participants were instructed to say the commands without a wake-word to avoid the effect of wake words on query formulation.

After each follow-along session, we conducted a follow-up interview to understand participants' experience with the video navigation, including their explanation of the commands, what went well, what did not go well, and what they could not achieve.

3.3 Data Analysis

The recordings of the study sessions were transcribed and analyzed using inductive thematic analysis [6]. The analysis focused on 'command actions' – moments in the recordings where the participants directed utterances towards the device in the attempt to control the playing video. Conversations with others in the home, or explicit questions directed at the researcher were not included. The post-session interview data were inductively analyzed to understand participants' overall experience and post-hoc explanations of their commands.

4 FINDINGS

On average, the follow-along session took 59 minutes (min 44 minutes, max 110 minutes) to complete. Interview data suggested that the overall feedback of voice-based playback control was positive, and it was helpful that they did not have to touch their device with food on their hands. The flexibility of using both time-based and content-based commands was reported to be helpful so that they did not have to have a print-out recipe or pre-watch the full video at the beginning.

The findings highlighted four types of high-level interactions or factors: context, navigational command types, dimensions of content-based commands, and challenges for design. Figure 2 presents the themes in summary, followed by a detailed theme-by-theme analysis.

4.1 Context

We elaborate here on the observed high-level interactions that were categorized into Context (Figure 2). The analysis presented here contributes to RQ1 (What are the contexts of user interactions) by showing how the types of interactions are specific to a technical or a social context. When designing specifically for interacting with video through voice, the context of viewing and interacting

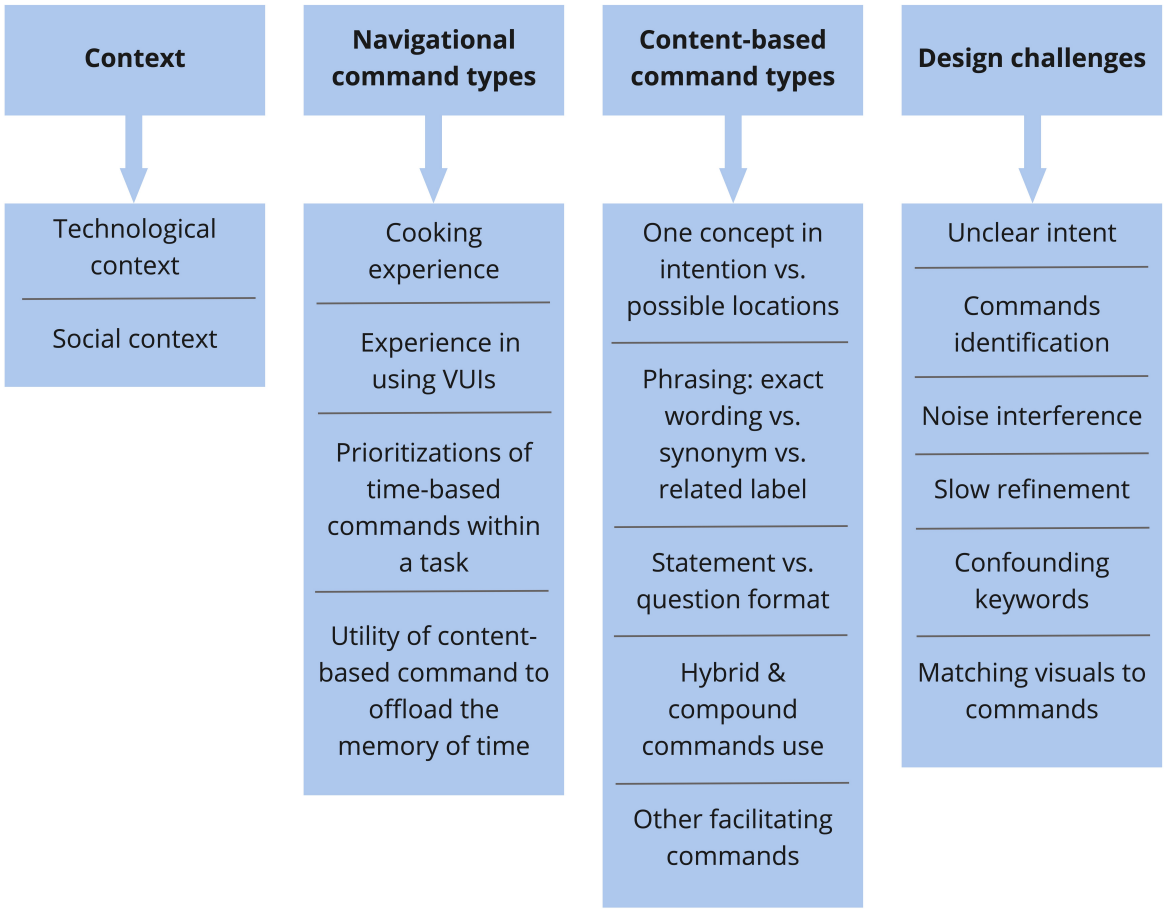

Figure 2: The high-level four types of interactions, and the type of interaction or factors affecting the interaction. These correspond to the high-level themes extracted from the thematic analysis that are explained in details in section 4

provides an important insight into the constraints and opportunities available. With a smaller screen, visual controls either occlude or shrink the video – reducing the detail on the cooking process available to the user when interacting with it and increasing the efficacy of voice in this context. Yet small screens also reduce the space available for possible visual cues for guiding interaction or repair. The shared visual media between agent and user is an obvious opportunity for design. With multiple parties engaging in conversation in and around the device, this shared visual media also provides a shared topic and context for multi-party interaction.

4.1.1 Technological Context. Technical context, including the use of small screen device and multiple VUIs together. Outwith the constraints of the study, our participants reported using small-screen devices (phone: n = 3, tablet: n = 5) to follow along with cooking videos or using printed instructions (n = 2) in their real-life because these devices are portable and do not require much space in a kitchen.

Five of the ten participants reported owning and frequently using at least one VUI at home. Two of them reported using theirs to set a timer. During the follow-along session, as the system/wizard setup was limited to navigation, it did not respond to participants' requests to set up a timer, which is frequently required in cooking scenarios. One participant used their Google home to set a timer when their attempt to do so with this system/wizard failed.

4.1.2 Social Context. Social context, including human and non-human background sounds,as well as out-loud and talking with others. There was much non-command talking or singing during cooking, especially in family settings. During the study, all participants talked with family members, spoke to themselves (e.g., reciting the recipe or instruction), and/or had music or TV playing in the background.

In most cases, when the participants were talking with family members or had background music/TV, the subject of the conversation did not align with the topic of the video, thus would not interfere with the voice command recognition. However, many participants habitually talked to themselves regarding the current or next step, the phrases of which would overlap with the recognizable command list, such as "ginger garlic" (p2).

Participants would also ask questions, and it was sometimes unclear whether they were talking to their family members, themselves, or posing a question to the voice assistant. P5 provided an example as they asked, "is she going to show me how to do that or I have to read the text?" when reading an ingredient list of the dough at the beginning of the video.

4.2 Navigation Commands and Types

In this section, we present the commands used by the participants to control the cooking recipe video. Following this, in section 4.3, we provide further details on the construction and use of the content-based commands issued. These commands are resultant from the instruction to participants to attempt 'any navigation-related voice commands that made sense to them' with no further constraints or priming on the availability of content or time-based commands. As a result of this setup, these commands can be seen as examples of an idealized voice interface for video in this context.

Overall there were 567 commands attempted to control the instructional video, of which 547 fit the criteria to allow the researcher to control the video as directed. On average each participant performed 54.7 actionable (min=28, max=112) and 2 unactionable (min=0, max=5) commands.

Among the actionable commands, the vast majority of them were pauses (37.8%, n = 207, e.g.,, "pause", "stop", "pause video", and "pause, pause, pause"), plays (22.9%, n = 125, e.g., "play", "play video", "start video", and "okay play"), and time-based commands (26.5%, n = 145, e.g., "rewind 2 minutes" and "fast forward"). Sixty-two of them (11.3%) were content-based commands, including jumps (e.g., "rewind to the jiggling meat part") and requests for customized markers (e.g., "add a marker to filling"). The rest of the actionable commands were for adjusting speed (n = 8), volume (n = 7), and subtitles (n = 1).

That the commands are heavily skewed towards pauses is perhaps unsurprising given the importance and ease of using such playback manipulation. As Tuncer et al. [58] discuss, pausing instructional videos can be used to align physically with the video, for example, finding the right tool or ingredient or managing the flow of information coming from the video that must be acted on. Of the 207 pauses, 53% (110) was followed by a play command, accounting for 88% of the play commands recorded. A further 30% (62) were followed by time-based commands, and 11% (22) content-based commands.

The 20 unactionable commands were due to requests of additional information (n=3), ambiguous intent (n=12), noisy environment (n=2), and long pauses between command refinement (n=3).

Looking more closely at these commands, we can see that (Table 2) on average, participants without experience with VUI issued significantly more commands compared to those with VUI experience (Mann-Whitney U = 3.5, $p < .05$). The less experienced participants issued significantly more time-based commands than content-based (Mann-Whitney U = 4, $p < .05$) while there was no significant difference between the two command types for those with experience of VUIs ($p > .1$). Participants with more cooking experience on average issued more commands, but the data was not analyzed with inferential statistics due to only three participants reporting a less than average cooking skill.

4.2.1 Using Time-Based vs. Content-Based Jumps. Participants prioritized giving time-based commands as they were more familiar with them in their daily life interacting with online videos or/and conversational assistants. Time-based commands are relatively shorter than content-based ones, thus more time-efficient for jumping to a close location within a task. In addition, some participants reported that it was easier to recall "rewind/forward" and a time period under a task than the explicit content.

However, when participants were switching between two lines of tasks, content-based jumps were reported to be more helpful as participants would lose track of time. Further, a few participants kept the video playing while working on time-consuming tasks (e.g., kneading the dough) to "have something in the background" (p4) or "to see the next steps" (p7). In this case, participants tended to use content-based jumps as recalling the relevant keywords around the task in hand was more convenient than a precise time.

Participants were also observed to switch between time- and content-based jumps. This could sometimes be seen as a result of an unsuccessful command, but could also be inferred to be the result of a command being executed but not resulting in the action they required. Isolating this change as an example of a repair strategy from naturally subsequent commands was challenging with this data, however we point to this as a possible are of analysis in a future study.

4.3 Breakdown of Content-Based Commands

Unlike time-based commands being straightforward to process and respond to, content-based commands contain more variety and often pose a challenge for voice interaction. Given that all participants were cooking from the same video, and that none of the participants had cooked this particular recipe before or seen this particular video, their formulation of content-based commands to control the video can be seen as an exploration of the expression space. In this section, we present the five dimensions of content-based commands that were identified in the data.

4.3.1 One Concept in Intention vs. Possible Locations. There were three types of match for content-based commands between the concept behind the command and the potential locations of the issued command. One concept in mind resulting in one or more locations in the subtitles and/or visuals.

Firstly, some semantic concepts would only result in one possible location in the verbal instruction. For example, in the entire video subtitles, "oyster sauce" only appeared once; thus any command with this keyword would result in one location. However, a concept like "dough" was a frequently used semantic concept in the process of making the dough, letting the dough rise, checking the dough, and cutting the dough to make dumpling wrappers, etc.. When participants commanded "get back to the dough part" (p7), it resulted

Table 2: Average number of commands per participant categorized by actionability, type, experience with VUI, and experience in cooking

		Experience in CA		Experience in cooking	
		Yes (5)	No (5)	More (7)	Less (3)
Actionability	actionable	41.4 (96.3%)	68 (96.6%)	57.9 (95.8%)	47.3 (98.5%)
	unactionable	1.6 (3.7%)	2.4 (3.4%)	2.6 (4.2%)	0.7 (1.5%)
Type	content	5.4 (12.6%)	8.6 (12.2%)	8.4 (13.9%)	4.0 (8.3%)
	time	9.6 (22.3%)	20.0 (28.4%)	16.6 (27.7%)	10.7 (22.3%)
	play and pause	27.4 (63.7%)	39.0 (55.4%)	33.6 (55.6%)	32.3 (67.3%)
	non-navigational	0.6 (1.4%)	2.8 (4.0%)	1.9 (3.1%)	1 (2.1%)
Total per participant		43	70.4	60.4	48

in multiple possible locations and was unactionable without further refinement.

Secondly, the answer to specific commands can be found in the visual yet not in the video's audio. For instance, p11 asked, "how coarsely should I chop the celery?" In the audio, the instructor only talked about the weight of celery needed, while the visual could demonstrate the coarseness of the chopped celery.

4.3.2 Phrasing: Exact Wording vs. Synonym vs. Related label. Participants demonstrated three strategies to describe a semantic concept. One was to use the exact words used in the instructional video, such as "go-to cutting the dough", "rewind to the part about soy sauce", in which "cut the dough" and "soy sauce" were used by the instructor. Another strategy was to use synonyms of the keywords, such as "opening the dough" ("flatten the dough" in the instruction) and "go back to mixing powders" ("flour" in the instruction). Lastly, participants would also phrase the semantic concept using a related label of the keyword. For instance, in the command "how many grams should the wrappers be?", p1 asked about the weight of each piece of the dough that would be referred to as a wrapper in later parts of the video. While "wrapper" was not a direct synonym for "each piece of the dough", they are different stages of the same ingredient.

4.3.3 Statement vs. Question. As already seen above, participants would command in a statement (e.g., "rewind to making wrapper") or a question format (e.g., "Can you rewind to the kneading part?"). The question format might not contain the navigational action words, such as "rewind", "forward", and "go to"; instead, it implied the action of "go to the part about...". In this case, the semantic concept of the question would be contextualized in the immediate video content they just watched, and the intent was to repair missed information. For instance, when the instructor talked about "cut it [the dough] into thick pieces", p5 asked "how thick?". Right after the instructor said "one tablespoon of vegetable oil", p3 asked "how much oil?".

Beyond the aforementioned content questions (i.e., "WH" questions), participants might also ask alternative questions in which they proposed a restrictive set of alternative answers in the formulation [54], involving confounding keywords that did not belong to the intended semantic concept. When following the process of making the filling, p1 asked, "did you say a teaspoon or a tablespoon of soy sauce?". In this case, only the "tablespoon" is a relevant keyword with "soy sauce", and the keyword "teaspoon" could contribute to irrelevant returns.

4.3.4 Hybrid & Compound Commands. Beyond directly requesting a single location in the video, some participants also issued hybrid commands. The first type of hybrid command asked for a loop of a session that was difficult to follow and contained many consecutive steps, such as wrapping a dumpling. As an example, p3 commanded "play dumpling-making on repeat". Another type of common hybrid command was nested commands, in which the second command needed to be understood in the context of the first one. To follow along with the making of garlic ginger puree (an ingredient of the filling), p8 first requested "rewind to the part about ginger", and then shortly asked "rewind to the part about water [with the ginger]". The short gap between the two commands and their relevance indicated that adding water to the ginger was the only possible location in this context.

In line with previous studies [9], users would request compound commands (i.e., more than one in a row) to achieve two consecutive goals, such as "rewind 20 seconds and stop" (p2). In our study, participants also used compound commands for the same goal, trying to provide as many possible keywords as possible so that at least some of them would be recognized. In this case, the compounds were semantically more complex that more than one concept appeared concurrently to identify one location. For example, p3 asked, "can you rewind to where the ingredients to the dough were added and you start the kneading process?". This might take them a while to complete the command, and some of the information could become confounding.

4.3.5 Other Facilitating Commands.

Conversational partner. Beyond directly navigating through the video, participants, especially those who did not have much cooking experience, required much help with the content. In this particular case, a few participants requested additional information about making the dough, such as "If you're using all-purpose flour, is the egg mixture in place of the water?" and "uh if my dough doesn't look smooth on top should I keep kneading it?". Being able to support this need would truly free their hands from touching their device with unclean hands. Further, participants also noted that it would be great if a VUI could just "answer these questions [e.g., "how much oil?"] without going back [to a specific location]".

Customized marker. Another type of content-based command that participants requested was adding markers with a customized

name that they could refer to later. This was useful when participants were doing a time-consuming task (e.g., kneading the dough) – planning to keep the video playing and then come back to compare their progress with the instruction. It would also be helpful when requesting a loop.

4.4 Challenges for Design

We identified six types of challenges in responding to users' requests that require further design considerations. While these mirror many common issues with the design of CUIs in general, here we highlight the specific challenges that they represent for controlling instructional video in a kitchen context.

(1) Unclear intent. Users would use words or phrases that appeared multiple times in the instructional video, which might require further refinements to land on the intended location. Further, users might follow a command with "more" or "repeat", both of which might indicate different intentions. For instance, with "more" following "rewind 20 seconds", it was unclear if they meant rewinding another 20 seconds or another "rewind", which would be only the default 5 or 10 seconds. Similarly, "repeat" could mean repeating the last command or repeating the content from the last command until now (i.e., requesting a loop). In addition, some participants quickly commanded "rewind, rewind", with which it was unclear whether the command was idempotent similar to "pause, pause" or they meant two "rewinds".
(2) Command identification. Many users repeated the ingredients or actions to themselves they needed to perform, thus eliciting words that overlapped with command keywords that might unnecessarily trigger the voice control. It became more challenging when users issued a command right after self-talk (e.g., "kneading, kneading...uhg... rewind", p6; they did not mean "rewind to kneading"). Further, people might use multiple voice assistants collectively, and it might confuse the intended interlocutor when no wake-word is required.
(3) Noise interference. There were TV sounds, people talking, music playing in a home kitchen, all might make a command undecipherable.
(4) Slow refinement. When users are unfamiliar with the rules of a system or with the instructional content, they might have difficulty deciding what content-based command to use. The commands would take a long time to issue and contain pauses in between, making it challenging to know when a command was completed.
(5) Confounding keywords. In an alternative question and slow refinement of a command, users might use keywords related to unnecessary locations, thus making keyword-based processing inefficient.
(6) Matching visuals to commands. Potentially a command could be matched to visuals when the instructional subtitles did not provide sufficient information, which could maximize the satisfaction of user needs. However, it would still be challenging to decide if a match to visuals is a sufficient solution.

5 DISCUSSION

Through investigating the use of an ideal voice-control system to follow-along cooking videos (as a popular example of non-linear instructional videos) in a home environment, we have highlighted the technical and social contexts, patterns of command formulation, and the challenges in request responding. Improving technology's ability to identify their human communication partners and provide specific clarification responses will ultimately improve humans' and machines' ability to collaborate towards shared understanding in their conversational interactions. To better design VUIs, we need to understand the full range of machine behaviors users expect to leverage when approaching a speech interface. Not simply to implement all possible ways that a user may interact with a system, but rather to predict and mitigate potential breakdowns when these behaviors are not implemented or do not work as users expect. We thus discuss the design and research opportunities towards an ideal VUI as well as the limitations of the current study.

5.1 Respond to External and Internal Context

Purposeful activities are situated in social and physical contexts [55], and an ideal VUI would be aware and responsive to its work environment and progress.

The findings reveal that users tend to use a small-screen portable device to follow-along instructional video in a kitchen environment so that the device does not interfere with the cooking process. Designing for voice navigation assistants should consider the screen size, which means limited space for information overlay or display on the screen and the necessity to reduce unnecessary information through algorithm and user experience design. Further, users might have other VUIs around the kitchen area or other sound sources (e.g., self-talk and music). In this example, unlike most commercial and research-based CUI systems, participants didn't have to use a wake-word to initiate communication with the WoZ. While a non-wake-word setting makes a smooth voice control process, it has to collaborate well with other voices, mainly not to be triggered by commands towards the other VUIs or content-relevant self-talk.

Without wake words, it is crucial to differentiate commands and self-talk that contains keywords. A distinguishable difference between self-talk and content-based commands for voice control was that commands would start with an actionable verb (e.g., "rewind to" or "go back to"). In contrast, self-talk would lack such navigation-related actions (e.g., "knead it again and again"). Yet, this distinction could disappear when the command was in a question format. In this case, another way to differentiate the two could be utilizing non-verbal information, such as gaze [39] and body gestures, as users tend to turn to the speaker when they issue a command. Other modalities can also be useful, such as buttons [7], using prosodic features of the utterance [37], or topic of the intent [1].

Further, the progress of the navigation also provides valuable internal context for a VUI to work efficiently. To tackle the multi-location challenge of possible responses, previous studies suggested using the combination of keywords to display possible locations in a chronological order for users to choose from [8]. However, a simple combination of keywords might generate a long list, which will be overwhelming for a small screen device. Beyond analyzing keywords, leveraging the video location and command history

can further disambiguate user intent when a command is issued. When users issue a hybrid command (e.g., "rewind to the part about water" shortly after "rewind to the part about ginger") or ask to repair missed information (e.g., "how much oil?"), the combination of keywords, pragmatics features, what they have just watched, and/or the time interval between two commands can collectively indicate the most possible location(s) for a command response, thus improving the response and facilitating refinements.

5.2 Build a Shared Vocabulary with Users

A successful conversation requires all interlocutors to "align their linguistic representation during dialogue" [47], thus it is important for an ideal VUI to be able to collaboratively build a shared vocabulary with users by properly guiding users towards an alignment, leveraging the implications in commands, and supporting users with varied experience in the domain area and with VUIs.

In line with the existing practice of commercial VUIs, an efficient strategy to achieve alignment is to guide users to learn the machines' linguistic features and conversational styles to prevent errors. Yet users sometimes do not have the domain familiarity with recalling keywords. Instead of asking for refinement after they issue an inefficient command, suggested keywords for each step could be displayed on each step to reinforce the memory of the content. Further, as Youtube allows content creators to add bookmarks to segment content, this can also be incorporated into the multimodal design (voice and visual). In fact, a participant suggested in the follow-up interview that "it would be nice to see a menu of all the steps".

Beyond that, VUIs should also understand and use the implication of user utterance. In time-based navigation, the playback action words (e.g., "rewind", "back" and "forward") indicate a jump's direction. This also applies to content-based jumps, limiting the possible responses to one direction. Further, based on general language use habits, when people refer to a noun without actions on it (e.g., "go back to the dough"), it likely indicates that they picture the noun as a whole, thus "cutting the dough" or "the ingredients of the dough" would be less likely locations. In addition, when users ask alternative questions ("X or Y?") or issue a long command with refinements, the collection of useful and confounding keywords might result in many unnecessary potential locations. As such, the voice playback control needs to act on the underlying meaning of the complete command, beyond the combination of keywords.

Designing for a more inclusive voice playback control also has to accommodate a wide range of experiences. People unfamiliar with VUIs tend to be less constrained by their expected technical limitations of voice control and be more excited to experiment with commands. The findings align with previous research that voice interfaces continue to be affected by poor usability [12, 41] due to factors such as lack of discoverability, more experienced users are expected to have adapted to the canons of (limited) interactions afforded by current VUIs. Conversely, users unfamiliar with VUIs are more likely to require additional resources to adapt to the limitations of the system (which are increasingly within the scope of design, as the engineering aspects of voice systems have advanced significantly in the past decade [3, 40]. In addition, more experienced home cooks are more comfortable issuing commands with the wordings they are more familiar with and pay attention to more details to get things right while people with less experience in cooking reported "unfamiliar with the ingredient names" thus "not sure what to say".

To create a better-shared vocabulary and alignment between VUIs and users, designers may want to consider approaches that could guide users, understand the meaning of the commands (in opposed to merely keywords), accommodate commonly used language that differs from the one in the instruction, and support those who cannot recall the unfamiliar concepts. This may become difficult to achieve with the user-centered methods more typically employed in interactive voice systems design, such as Wizard of Oz [46]. It may require designers to consider a more extensive use of ethnographic and participatory design methods, which are not widely used outside of some notable examples (e.g. [49, 50]). Designers may also benefit from incorporating insights into interpersonal conversation tactics, such as parent-child and teacher-student communication.

themes

5.3 Take Advantage of the Visuals

In an ideal world, the voice agent controlling the video would have the same level of understanding of the visual content of the media as the user watching it.

As we saw, users chose words that often differed from what the narrator in the video said or wrote in subtitles – especially those with less experience of the pedantry of commercial voice assistants. One reason for this shift in vocabulary could stem from the participants' watching the visuals and using them as a guide to the content. Returning to the case in the title of the paper, the meat did visibly 'jiggle' as the video presenter threw it repeatedly into the bowl in order to tenderize it – but no direct synonym for jiggle was mentioned to make it easily available for computation in the transcript. Part of this can be overcome with extended use as people learn to modulate their grammar, and expectations, to be within the limits of such an agent. Yet it is understandable that the visual component of the video, especially in a system explicitly framed for video control, should be part of the expected shared context of interaction.

One opportunity would be to leverage human-generated closed captions – where the visual scene is described for non-sighted audience members. While YouTube seems to conflate subtitles (where only the speech in a video is written as an overlay) with closed captions, in general, there are very few examples of user-generated content, including true descriptive closed captions. Machine learning techniques are under development [23], but they are far from providing usable closed caption information. While this is an opportunity for more traditional media types – such as television and movies where many countries require closed captions by law – for user-generated instructional videos, this is unlikely to be available.

Another way to develop a shared context involving the visual spectrum would be to leverage currently available image segmentation and object recognition algorithms. Systems such as ImageNet [53] can provide lists of objects visible in any given frame, and metadata such as their color. However, this would not directly provide a shared context as the user would, quite rightly, expect that a system understanding that a tool is in the frame could understand what

that tool is being used for. Providing a small overlay on the screen, simply showing icons of the detected items and highlighting when they appear and disappear from the ongoing video could provide enough of a shared understanding that the system can 'show me the money' while it is unable to 'show me them counting.'

5.4 Limitations

The current study has a few limitations. For one, it was the first time for the participants to watch the instructional video, thus the navigational needs among those who revisit an instructional video might be different. Further, while the current study provides in-depth data generated in a natural environment, a larger sample size will allow better generalization across other contexts. In addition, the current research focused on understanding query formulation and did not intentionally create opportunities to investigate query repairs and further collaboration between humans and the system. Future research can complement the insights from this study with larger-scale studies, adding a non-first time viewer condition, and exploring other forms of interactions.

6 CONCLUSION

We have presented insights from a Wizard-of-Oz elicitation study of 10 participants, who had varied experience in domain expertise and VUIs, using voice-based playback control to follow a non-linear instructional video. The identified contexts, types of command formulation, and the challenges in query response suggest opportunities for improving the design of voice-based video navigation and enhancing human-machine collaboration.

REFERENCES

[1] Saul Albert and Magnus Hamann. 2021. Putting wake words to bed: We speak wake words with systematically varied prosody, but CUIs don't listen. In *CUI 2021-3rd Conference on Conversational User Interfaces*. 1–5. https://doi.org/10.1145/3469595.3469608

[2] Rami Alkhatib, Tarek El Bobo, Afif Swaidan, Jad Al Soussi, Mohamad O. Diab, and Nassim Khaled. 2021. Design of Robotic Manipulator to Hollow Out Zucchini. American Society of Mechanical Engineers Digital Collection. https://doi.org/10.1115/IMECE2020-24389

[3] Matthew P. Aylett, Per Ola Kristensson, Steve Whittaker, and Yolanda Vazquez-Alvarez. 2014. None of a CHInd: Relationship Counselling for HCI and Speech Technology. ACM, New York, NY, USA, 749–760. https://doi.org/10.1145/2559206.2578868

[4] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300473

[5] Dan Bohus and Alexander I Rudnicky. 2003. RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda. (2003), 4.

[6] V. Braun and V. Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. http://dx.doi.org/10.1191/1478088706qp063oa

[7] Fabio Catania, Micol Spitale, Giulia Cosentino, and Franca Garzotto. 2020. What is the Best Action for Children to" Wake Up" and" Put to Sleep" a Conversational Agent? A Multi-Criteria Decision Analysis Approach. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–10. https://doi.org/10.1145/3405755.3406129

[8] Minsuk Chang, Mina Huh, and Juho Kim. 2021. RubySlippers: Supporting Content-based Voice Navigation for How-to Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3411764.3445131

[9] Minsuk Chang, Anh Truong, Oliver Wang, Maneesh Agrawala, and Juho Kim. 2019. How to design voice based navigation for how-to videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3290605.3300931

[10] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. 2018. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *Proceedings of the 26th ACM international conference on Multimedia*. 1020–1028. https://doi.org/10.1145/3240508.3240627

[11] Yi Cheng, Kate Yen, Yeqi Chen, Sijin Chen, and Alexis Hiniker. 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*. ACM Press, Trondheim, Norway, 337–348. https://doi.org/10.1145/3202185.3202749

[12] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, and Cosmin Munteanu. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[13] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?": infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '17*. ACM Press, Vienna, Austria, 1–12. https://doi.org/10.1145/3098279.3098539

[14] Andy Crabtree and Tom Rodden. 2004. Domestic routines and design for the home. *Computer Supported Cooperative Work* 13, 2 (2004), 191–220. https://doi.org/10.1023/B:COSU.0000045712.26840.a4

[15] Sebastien Cuendet, Indrani Medhi, Kalika Bali, and Edward Cutrell. 2013. VideoKheti: making video content accessible to low-literate and novice users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, Paris, France, 2833. https://doi.org/10.1145/2470654.2481392

[16] Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video browsing by direct manipulation. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*. ACM Press, Florence, Italy, 237. https://doi.org/10.1145/1357054.1357096

[17] Justin Edwards, He Liu, Tianyu Zhou, Sandy J. J. Gould, Leigh Clark, Philip Doyle, and Benjamin R. Cowan. 2019. Multitasking with Alexa: how using intelligent personal assistants impacts language-based primary task performance. In *Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19*. ACM Press, Dublin, Ireland, 1–7. https://doi.org/10.1145/3342775.3342785

[18] Huan Feng, Kassem Fawaz, and Kang G. Shin. 2017. Continuous Authentication for Voice Assistants. *arXiv:1701.04507 [cs]* (Jan. 2017). https://doi.org/10.1145/3117811.3117823

[19] Joel E. Fischer, Stuart Reeves, Martin Porcheron, and Rein Ove Sikveland. 2019. Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces - CUI '19*. ACM Press, Dublin, Ireland, 1–8. https://doi.org/10.1145/3342775.3342788

[20] Peter Grasch, Alexander Felfernig, and Florian Reinfrank. 2013. ReComment: towards critiquing-based recommendation with speech interaction. In *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13*. ACM Press, Hong Kong, China, 157–164. https://doi.org/10.1145/2507157.2507161

[21] Philip J. Guo and Katharina Reinecke. 2014. Demographic differences in how students navigate through MOOCs. In *Proceedings of the first ACM conference on Learning @ scale conference (L@S '14)*. Association for Computing Machinery, New York, NY, USA, 21–30. https://doi.org/10.1145/2556325.2566247

[22] Kotaro Hara and Shamsi T. Iqbal. 2015. Effect of Machine Translation in Interlingual Conversation: Lessons from a Formative Study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, Seoul, Republic of Korea, 3473–3482. https://doi.org/10.1145/2702123.2702407

[23] Chiori Hori, Takaaki Hori, Tim K. Marks, and John R. Hershey. 2017. Early and late integration of audio features for automatic video description. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 430–436. https://doi.org/10.1109/ASRU.2017.8268968

[24] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. Convey: Exploring the Use of a Context View for Chatbots. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–6. https://doi.org/10.1145/3173574.3174042

[25] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic Online Evaluation of Intelligent Assistants. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 506–516. https://doi.org/10.1145/2736277.2741669

[26] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors?: lexical and phonetic query reformulation in voice search. (2013), 10. https://doi.org/10.1145/2484028.2484092

[27] John F. Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 193–196. https://doi.org/10.1145/800045.801609

[28] Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en Place: Unsupervised Interpretation of Instructional Recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 982–992. https://doi.org/10.18653/v1/D15-1114

[29] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J. Guo, Robert C. Miller, and Krzysztof Z. Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 4017–4026. https://doi.org/10.1145/2556288.2556986

[30] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In *International workshop on intelligent virtual agents*. Springer, 329–343. https://doi.org/10.1007/11550617_28

[31] Philip Kortum. 2008. *HCI Beyond the GUI: Design for Haptic, Speech, Olfactory, and Other Nontraditional Interfaces*. Elsevier.

[32] Anuj Kumar, Pooja Reddy, Anuj Tewari, Rajat Agrawal, and Matthew Kam. 2012. Improving literacy in developing countries using speech recognition-supported games on mobile devices. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*. ACM Press, Austin, Texas, USA, 1149. https://doi.org/10.1145/2207676.2208564

[33] Sanna Kuoppamäki, Sylvaine Tuncer, Sara Eriksson, and Donald McMillan. 2021. Designing Kitchen Technologies for Ageing in Place: A Video Study of Older Adults' Cooking at Home. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2 (June 2021), 1–19. https://doi.org/10.1145/3463516

[34] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan. 2015. Spoken Content Retrieval—Beyond Cascading Speech Recognition with Text Retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 9 (Sept. 2015), 1389–1420. https://doi.org/10.1109/TASLP.2015.2438543

[35] Q. Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N. Sadat Shami. 2016. What Can You Do?: Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems - DIS '16*. ACM Press, Brisbane, QLD, Australia, 264–275. https://doi.org/10.1145/2901790.2901842

[36] Mike Ligthart, Timo Fernhout, Mark A. Neerincx, Kelly L. A. van Bindsbergen, Martha A. Grootenhuis, and Koen V. Hindriks. 2019. A child and a robot getting acquainted - interaction design for eliciting self-disclosure. In *Proceedings of the 18th international conference on autonomous agents and MultiAgent systems (AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 61–70.

[37] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2016. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, Christchurch, New Zealand, 255–262. https://doi.org/10.1109/HRI.2016.7451760

[38] Ewa Luger and Abigail Sellen. 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)*. ACM, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[39] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: Tama – a Gaze Activated Smart-Speaker. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 176:1–176:26. https://doi.org/10.1145/3359278

[40] Cosmin Munteanu, Gerald Penn, and Christine Murad. 2021. Conversational Voice User Interfaces: Connecting Engineering Fundamentals to Design Considerations. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–3. https://doi.org/10.1145/3411763.3445008

[41] Christine Murad and Cosmin Munteanu. 2019. I don't know what you're talking about, HALexa" the case for voice user interface guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–3. https://doi.org/10.1145/3342775.3342795

[42] Dania Murad, Riwu Wang, Douglas Turnbull, and Ye Wang. 2018. SLIONS: A Karaoke Application to Enhance Foreign Language Learning. In *2018 ACM Multimedia Conference on Multimedia Conference - MM '18*. ACM Press, Seoul, Republic of Korea, 1679–1687. https://doi.org/10.1145/3240508.3240691

[43] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. 1998. Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Lang Speech* 41, 3-4 (July 1998), 419–442. https://doi.org/10.1177/002383099804100409

[44] Jamie Pearson, Jiang Hu, Holly P Branigan, Martin J Pickering, and Clifford I Nass. 2006. Adaptive Language Behavior in HCI: How Expectations and Beliefs about a System Affect Users' Word Choice. (2006), 4. https://doi.org/10.1145/1124772.1124948

[45] Hannah R.M. Pelikan and Mathias Broth. 2016. Why That Nao?: How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4921–4932. https://doi.org/10.1145/2858036.2858478

[46] John Sören Pettersson and Malin Wik. 2014. Perspectives on Ozlab in the cloud: A literature review of tools supporting Wizard-of-Oz experimentation, including an historical overview of 1971-2013 and notes on methodological issues and supporting generic tools. (2014).

[47] Martin J. Pickering and Simon Garrod. 2006. Alignment as the basis for successful communication. *Research on Language and Computation* 4, 2 (2006), 203–228. https://doi.org/0.1007/s11168-006-9004-0

[48] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 640:1–640:12. https://doi.org/10.1145/3173574.3174214

[49] Gisela Reyes-Cruz, Joel Fischer, and Stuart Reeves. 2019. An ethnographic study of visual impairments for voice user interface design. *arXiv preprint arXiv:1904.06123* (2019).

[50] Simon Robinson, Jennifer Pearson, Shashank Ahire, Rini Ahirwar, Bhakti Bhikne, Nimish Maravi, and Matt Jones. 2018. Revisiting "hole in the wall" computing: Private smart speakers and public slum settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3173574.3174072

[51] Lucas Rosenblatt, Patrick Carrington, Kotaro Hara, and Jeffrey P. Bigham. 2018. Vocal Programming for People with Upper-Body Motor Impairments. In *Proceedings of the Internet of Accessible Things on - W4A '18*. ACM Press, Lyon, France, 1–10. https://doi.org/10.1145/3192714.3192821

[52] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the Effects of Embodiment for a Group Facilitation Agent. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–13. https://doi.org/10.1145/3173574.3173965

[53] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. 2020. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning*. PMLR, 8634–8644.

[54] Tanya Stivers, Nicholas J. Enfield, and Stephen C. Levinson. 2010. Question-response sequences in conversation across ten languages: an introduction. *Journal of Pragmatics* 42 (2010), 2615–2619. https://doi.org/10.1016/j.pragma.2010.04.001

[55] Lucy Suchman. 1987. *Plans and Situated Actions: The Problem of Human-Machine Communication*. Cambridge University Press.

[56] K. Tanaka, T. Sasaki, Y. Tonomura, T. Nakanishi, and N. Babaguchi. 2005. PlayWatch: chart-style video playback interface. In *2005 IEEE International Conference on Multimedia and Expo*. 4 pp.–. https://doi.org/10.1109/ICME.2005.1521527

[57] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR '18*. ACM Press, New Brunswick, NJ, USA, 32–41. https://doi.org/10.1145/3176349.3176387

[58] Sylvaine Tuncer, Barry Brown, and Oskar Lindwall. 2020. On Pause: How Online Instructional Videos are Used to Achieve Practical Tasks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376759

[59] Ruolin Wang, Chun Yu, Xing-Dong Yang, Weijie He, and Yuanchun Shi. 2019. EarTouch: Facilitating Smartphone Use for Visually Impaired People in Mobile and Public Scenarios. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–13. https://doi.org/10.1145/3290605.3300254

[60] Steve Whittaker, Vaiva Kalnikaité, and Patrick Ehlen. 2012. Markup as you talk: establishing effective memory cues while still contributing to a meeting. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. ACM Press, Seattle, Washington, USA, 349. https://doi.org/10.1145/2145204.2145260

[61] Cara Wilson, Margot Brereton, Bernd Ploderer, and Laurianne Sitbon. 2018. MyWord: enhancing engagement, interaction and self-expression with minimally-verbal children on the autism spectrum through a personal audio-visual dictionary. In *Proceedings of the 17th ACM Conference on Interaction Design and Children - IDC '18*. ACM Press, Trondheim, Norway, 106–118. https://doi.org/10.1145/3202185.3202755

[62] Kuldeep Yadav, Kundan Shrivastava, S. Mohana Prasad, Harish Arsikere, Sonal Patil, Ranjeet Kumar, and Om Deshmukh. 2015. Content-driven Multi-modal Techniques for Non-linear Video Navigation. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, Atlanta Georgia USA, 333–344. https://doi.org/10.1145/2678025.2701408

[63] Akiko Yamazaki, Keiichi Yamazaki, Matthew Burdelski, Yoshinori Kuno, and Mihoko Fukushima. 2010. Coordination of verbal and non-verbal actions in human–robot interaction at museums and exhibitions. *Journal of Pragmatics* 42, 9 (Sept. 2010), 2398–2414. https://doi.org/10.1016/j.pragma.2009.12.023