Cross-Modal Repair: Gaze and Speech Interaction for List Advancement

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

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

ABSTRACT

Interacting with long lists of instructions or ingredients continues to be a challenge for conversational interaction. In this paper, we conducted a user study to experiment with the use of 'cued-gaze' – waiting for the user's visual attention – to manage the delivery of instructions with a voice agent. In a Wizard-of-Oz setting, 12 participants were instructed to build a simple Lego tower by a conversational agent and were able to advance in the list using either speech interaction, or gaze interaction. The increasing use of speech agents in real-world cause users to encounter failures in interactions, so in this task the agent was designed to fail when providing the list of instruction to explore how the participants proceeded to recover from common failures. This showed that, for this use case, cross-modality repair was more effective than reformulation of speech.

CCS CONCEPTS

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

KEYWORDS

conversational user interface, gaze interaction, speech interaction, user study

ACM Reference Format:
Razan Jaber and Donald McMillan. 2022. Cross-Modal Repair: Gaze and Speech Interaction for List Advancement. In *4th Conference on Conversational User Interfaces (CUI 2022), July 26–28, 2022, Glasgow, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543829.3543833

1 INTRODUCTION

Despite the challenges of interacting with computers with voice commands, devices incorporating conversational user interfaces (CUIs) are becoming pervasive in everyday life. They present hands-free and screen-less interaction and encourage users to engage using a natural and familiar modality that supposedly does not require them to learn new technical concepts or interaction methods. However, the CUIs available today are set up to fail when these expectations meet the natural spoken interactions with and between humans [7].

CUI 2022, July 26–28, 2022, Glasgow, United Kingdom

ACM ISBN 978-1-4503-9739-1/22/07.
https://doi.org/10.1145/3543829.3543833

Users come with a set of expectations about how spoken conversation should work that is outside the capability of today's speech technology [33]. A recurring theme of research into conversational agents (CAs) is that "user expectations of CA systems remain far from the practical realities of use" [40]. For example, in real use, there is often overlapping speech, less clearly spoken commands, and back-channel noise in home environments. Consider participants in Luger and Sellen's study of CA users, who "described making use of a particular economy of language. Dropping words other than 'keywords', removing colloquial or complex words, reducing the number of words used, using more specific terms, altering enunciation, speaking more slowly/clearly, and changing accent was the most commonly described tactics" [40]. When speech technologies fail, users become increasingly frustrated with the device and tend to blame themselves [40]. Porcheron et al. [50] suggest that future work on voice user interfaces (VUIs) should, for the time being, shift "from conversation design to [. . .] request/response design".

To design conversational agents that could meet this longer-term goal, we need to understand the full range of behavior users expect to leverage when they initially approach a conversational interface. Natural conversation is a complex speech-exchange system, which Harvy Sacks called a 'machinery'. He stated that "Human conversation consists of a generic speech-exchange system that is continually adapted by speakers to different activities and situations" [53]. Conversations with the CUIs could be improved by applying the formal knowledge of human conversation with turn-taking systems, sequence organization, and repair strategies. One such mechanism of the machinery of human-human interaction is the use of gaze as a complex channel, alongside the speech itself, to manage attention, expectation, turn-taking, and the progressivity of interaction. One aspect of this complex machinery that we take advantage of here is the use of 'cued-gaze' [53] in human-human communications used for one party in a conversation to indicate their readiness to hear. Combining this with the longstanding problems of lists of options or instructions being provided by speech agents [40] being either slow or difficult to manage, we designed an interaction where the list progressivity was tied to the 'cued-gaze' of the user towards the speech agent.

Since conversation is the primary modality for interaction with speech interfaces, visual feedback is not always available. Little is known about how using gaze could affect users' intentions to interact with the system when failures occur. In this paper, we conducted a user study to compare the interactions of 'cued-gaze' and spoken commands in advancing a list and when things go wrong. The following hypothesis guided this work:

- H0: Users will be able to advance through a list of instructions by giving visual attention to the agent as fluently as using verbal instructions.

- H1: Switching modality to affect a recovery will take longer and cause more frustration compared to staying with the same modality.

We conducted a laboratory study of single user interaction with Tama, with the hypothesis that gaze interaction would result in the participants having a better impression of the interaction when the system fails [34, 69] and be drawn away from their ongoing task less [13].

In a Wizard-of-Oz setting, participants (N=12) were instructed to build a simple Lego tower by a smart agent, advancing the list of instructions on what blocks to add using either speech or gaze. The agent was designed to fail at specific points when providing the instructions to the user, and the users were faced with the problem of recovering from repetitions – where the system continued to utter the same step – and errors – where the system made only an error sound when they attempted to interact with it.

In the quantitative results, we show that 'cued-gaze' is as effective method of progressing lists as speech and also that it is (in certain circumstances) *more* effective when errors of interaction can be addressed by switching modalities to gaze. We also present the results of video analysis of the interactions and repair strategies carried out by our participants, showing the complexities of possible spoken repair strategies as opposed to the relative simplicity of switching modality, and the challenges of effectively communicating the meaning behind errors experienced when interacting with CUIs. In our discussion, we point to other opportunities for 'cued-gaze' and the family of overt-gaze interactions of which it is a part.

2 BACKGROUND

2.1 Speech Technology Development and Interaction

Over the last few years, we have seen immense growth in the research and development of speech systems. One of the most popular applications of speech technology is smart speakers like Amazon's Alexa and others. Today, we can talk to these devices in our home and make them perform actions on our behalf, like search for movies, present instructions from a cooking recipe, or turn devices on or off. Recently, there has been a growth of research investigating interactions with speech agents. Looking at Smart Speaker use in particular, Sciuto et al. [57] explore the experience of households who are using a conversational agent in their homes. Their findings showed how families integrate conversational agents into their daily life. By analyzing the logs of 75 Alexa users, they concluded their results into four themes: how people initially use Alexa, the physical placement of the device, the daily patterns of conversational usage, and how children interact with the device. Porcheron et al. [50] examined how users of smart speakers practically and interactionally situate the device's talk within their ongoing conversational setting at home and how users' formulate and direct queries to the device. These studies highlight a number of similar problems in speech agent interaction, including; troubles situating the interaction in an ongoing conversation, troubles activating and verifying activation, and troubles learning how to formulate and enunciate queries for these devices.

Interaction with speech agents tends to take a highly-task oriented question and answer format, rather than being social or conversational, which makes some users feel frustrated when interacting with such systems [16]. Speech systems need to maintain an understanding of context over multiple turns of interaction [10], and ask appropriate clarifying questions to guide the user [37]. For example, progressivity is a central feature of everyday conversation, as examined by Conversation Analysis. Hence, Fischer et al. [19] examined how the orientation towards progressivity in the talk -keeping things moving- might help us better understand and design for voice interactions. While agent's ability to engage in dialogue has been studied quite extensively [9, 63], the conversational style of these agents has received less attention. However, it has been shown that people's perceptions of conversational agents are influenced by the interaction style of the agent [47]. Speech agents often fail to fit into multi-party conversations, for instance, causing disruption when family members interact with the smart speaker during a meal while others are conversing concurrently (e.g., [50]). Precisely because of this gap, researchers take issue with calling conversational systems conversational [40, 50]. A conversational system must be designed with the ability to manage the floor of interaction and, therefore, have mechanisms for handling turn-taking, grounding, interruptions, and repair.

While the technology underlying speech interfaces have improved in recent years, our understanding of the human side of speech interactions remains limited [3]. Luger and Sellen [40] highlighted the limited functionality of existing commercially-available voice interfaces and how it causes a gulf between their capabilities and the users' expectations. Users come with a set of expectations about how spoken conversation should work outside the capability of today's speech technology [33]. Significant gaps still exist in using theoretical frameworks to understand user behaviors and choices and how they might be applied to specific speech interface interactions.

2.2 Gaze in Human-Human Conversation

In social science, researchers have been exploring gaze interaction since the mid-sixties. Much of the early work on gaze focused on the role of gaze in conversation [2, 32]. Goffman [22] observed that the direction of eye gaze plays a crucial role in the initiation and maintenance of social encounters. Kendon [32] conducted a detailed exploration of the function of gaze in face-to-face conversation. He classified looking, or avoiding to look, at the conversational partner as an indication of monitoring, regulating, concentrating, or expressing emotion. Kendon summarised attentive gaze in conversation by saying that people tend to look at the other participant more when listening than when speaking and that the speaker's glances at the other person tend to be shorter than those observed during listening [32]. In the study of the organisation of summons-answer Schegloff [55] proposes that the occurrence of a first item in a sequence, such as a summons establishes the relevance of the next item. Thus, the absence of an answer to a summons might be noted by the repetition of the summons, until an answer is obtained, which then allows the summons to move on to further talk. So, if a recipient fails to gaze at a speaker after an initial restart, that can cause the production of a new restart, which will affect the

repeating of the summons. In particular, restarts provides a speaker with the ability to begin a new sentence at the point where the recipient gaze is obtained, or alternatively to request a gaze from the hearer.

Goodwin [23] proposed two 'rules' between the speaker and the hearer in face-to-face conversation. The first is that a speaker should obtain the gaze of his recipient during the course of a turn at talk. Goodwin noted that "*when the speaker has the gaze of the recipient, a coherent sentence is produced. To have the gaze of a recipient thus appears to be preferred over not having such gaze, and this preference appears to be consequential for the talk the speaker produces within the turn*". In this way, gaze is an important cue that indicates that the hearer is listening to the speaker. This is consistent with the possibility that gaze is one means available to recipients for displaying to a speaker whether or not they are acting as hearers. The second rule states that a hearer should be gazing at the speaker when the speaker is gazing at the hearer. The speaker can look away from the recipient, but the recipient should not look away from a speaker looking at them. Obtaining the recipient's gaze within the turn is relevant to the speaker.

However, interacting with a CUI is not the same as interacting with a human – and this difference must be understood and taken into account during the design of such technology. Understanding human interaction with CUI and conversational interaction between people is a crucial step in improving the designing of such systems. There has been a significant body of work showing how people interact with robots and speech agents by drawing on their expectations of human-human interactions [1, 4, 5, 8, 26]. Usually, users' behavior differs considerably in terms of the usage of interaction modalities [27]. Therefore, to design multimodal input interaction for existing and new applications, it is necessary to understand how users would use those modalities in different situations [31]. Different modalities can be used to complement each other, to enable a natural and intuitive interaction [62]. Using gaze interaction is interesting as it could overcome some of the practical limitations of using speech as the only interaction modality when interacting with a speech system. By following other persons' gaze we gain access to their attentional focus, which is essential for understanding their intended interaction with technologies for example. A large body of the research has relied on gaze-cueing paradigms, in which the influence of static gaze cues on attentional processing is examined (e.g. [20, 36, 49]).

Porcheron et al. [50] suggested that a conversational system must be designed with the ability to manage the floor of interaction and, therefore, have mechanisms for handling turn-taking, and grounding, interruptions, and repair. Building on human-human interaction, this paper examines the use of gaze to initiate interaction when advancing in a list of instructions, which is a common communication task with voice-controlled interfaces in domestic environments such as cooking [14, 15, 64].

2.3 Communication Breakdowns and Repair

The increasing popularity of digital home assistants, like the Amazon Echo, and conversational assistants, like Siri, increase users' expectations of voice as an effective communication method with machines [40]. However, humans must work to adapt their communication patterns to the needs of the machines, rather than machines adapting to humans [28]. People shorten their sentences, use simplified language, and repeat themselves in attempts to be understood by voice interfaces [28, 40, 48]. As a result, users of voice interface technology become frustrated and can fail to learn the full capabilities of the technology or abandon use altogether [7, 16, 40]. True conversational capabilities have not yet been fully realized [50], and we need to understand better how human-technology "conversations" can be improved.

The field of HCI has a well-established body of literature on how humans verbally communicate with computers, robots, and other devices. In 1987, Suchman framed acts of human-machine interaction as a dialog between communication partners [60]. From this perspective, the designer's work is to enable human and machine collaboration towards a shared understanding through continuous acts of collaborative communication repair when breakdowns occur. Communication repair as presented in [7] "refers to the work of restoring shared understanding after conversational partners misunderstand each other. Essentially, the person who is talking needs to rephrase or say something different because the person they were talking to did not understand what they were saying [43]". Adjustments of speaking style to accommodate the listener can take many forms and depend on people's communication and language abilities [65]. It was observed that people simplified their speech to align with the robot, which highlights the compromise that users make between effective interaction and natural speech [66]. Research continues to demonstrate that humans adapt their communication styles and patterns to match the machine, both with robots [48, 61, 68] and computers [45, 46], rather than the other way around. Humans shorten their sentences [48], use repetition [6], increase volume [12] and hyperarticulate [45] as repair strategies. These modification strategies are motivated by a desire to achieve successful communication with computers [12, 46].

Despite the "conversational" interface with conversational agents, people are not yet able to talk to technology in the same way that they talk to other people [40, 51]. Users of conversational assistants often need to shorten their queries to keywords since increased utterance length can increase the likelihood of speech recognition errors, both with conversational agents and with other humans [28, 38]. With current systems, the burden of ensuring a successful communication interaction with a conversational agent continues to fall to the human in the conversation, with little support from the conversational agent itself [17, 51].

While error handling could be implemented by adding explicit error states when developing speech agents, speech repair detection aims to detect and resolve such occurrences in a more general case. Humans and machines perceive speech differently. In the design of conversational agents, we need to understand the range of behaviour that users expect and the range of behaviour that they may attempt to employ when interacting with them. This will allow us to understand potential breakdowns, and provide support strategies within the interaction that match users recovery procedures in challenging human-human interactions.

Figure 1: The Tama Gaze-Aware Smart Speaker Platform.

Figure 2: Speech agent instructing a user how to build a tower of Lego.

3 EXPERIMENT

3.1 The Speech Agent

This study was conducted using the Tama smart speaker platform [41] (Figure 1). Built on the open-source Mycroft.ai conversational agent [56], the Tama platform is capable of both speech interaction and gaze input and output. It uses a retractable spherical head containing two full-colour LED eyes with 180 degrees of movement laterally and 60 degrees vertically to provide gaze feedback. The gaze input is detected by two OMRON HVC-P2 gaze detection cameras built into the body, which also houses a 7- microphone array (ReSpeaker v2.0), a speaker, and a Raspberry Pi v.3B. This platform was extended with a browser-based Wizard-of-OZ control panel allowing a researcher to control eye movements and color, and define and control utterances to be spoken by the text-to-speech system.

For this study, the wizard interface was used to control the device's actions. The wake-word was disabled, and the eye gaze was set to follow the user even when they were not looking at the system. Subjects were instructed on building a Lego tower in a series of spoken, numbered steps. For this experiment, Tama had two conditions. In the *Speech Interaction* condition, the instruction list was advanced with a variation of the spoken command 'next step.' In the *Gaze Interaction* condition, Tama advanced the list using cued-gaze; when the system had finished presenting an instruction, it would wait until it received visual attention from the user before starting the next one. In both conditions, the user could use simple natural language commands to navigate through the list and ask the system to move backward, forward, or jump to a specific step. However, the wizard ignored any commands not directly about navigating the list of instructions. In both conditions, the trial starts, and the eyes go green, which means that it is listening to users, and the eyes go pink when processing or replying. We engineered failures in the interaction during the experiment to study the recovery strategies in each condition.

3.2 Types of Failure

In this study, the interactions were designed to include failures in both conditions. The failures were informed by taxonomies of failures in previous studies of interaction with speech agents [19], and in HRI [25], and represented typical reported speech agent malfunctions. The failures when interacting with the system could be task-oriented (e.g., providing wrong answers or repeating the same instruction) or interaction-oriented (e.g., giving no response or not activating). All failures caused delays and increased the time users needed to complete the task.

- **No Response.** The agent shows attention but plays the Google Assistant error sound simulating being unable to act upon or process user input.
- **Repetition.** The agent appears to respond to the interactions but repeats the previous instruction, similar to [15, 39].

In both conditions, the interactions continued with the next item in the list of instructions – even if the user took no action to remedy the failing interaction – after five errors were presented to the user. If they attempted a recovery, the next instruction would be given, and the interactions would continue. In the speech condition, a recovery attempt was deemed successful if it included a reformulation of the *verb* in the intent - for example, changing the command from 'advance instructions' to 'go to the next step.' In the gaze condition, a recovery attempt was deemed successful if they changed the gaze action they presented to the device from mutual gaze to another gaze-based gesture or changed modality to give a spoken command to the system. As such, interactions are consistent across different conditions. The timing and the number of failures were predetermined per condition in each sequence. They were counter-balanced per condition (the two variations in each trial contained the same failures but were reordered so that the failures would not become predictable and expected). We have introduced three failures, one error, and two repeats for each condition. As noted below, the individual interactions in this experiment were designed to be atomic. To make the repeats more obvious (as they would be in a more complex task with familiar sequentiality, such as cooking), each instruction in the list started with its place in the instruction list, i.e., 'Step 2'.

3.3 Lab Experiment

To ensure that the experiment wasn't confounded by speech recognition or gaze detection errors, we used a human wizard to control the output of the system [52]. The wizard sat in an adjacent room, watching through a high definition video call with the view of the participant and Tama, and controlled the system's speech output

Table 1: Successful Interactions

	Group	N	Mean	SD	SE
Successful Interaction Length (seconds)	*Gaze*	170	9.626	5.119	0.393
	Speech	170	10.201	4.926	0.378

Table 2: Induced Error Interactions

	Condition	N	Mean	SD	SE
Problematic Interaction Length (seconds)	*Gaze*	30	25.652	10.814	1.974
	Speech	30	33.633	15.941	2.910

and gaze feedback. The wizarding system included an experiment control interface (Figure 2) that managed the configuration and experimental conditions while recording all actions and interactions with the system and related timings.

The study was designed as a within-participants experiment with counter-balanced conditions. In each condition, the participant sat at a table with all the Lego pieces required to complete the task. Tama was placed on the table in front and to the right of the participant, 0.5 m away from the seating position of the participant, at a 90 degree angle relative to them. The experiment was captured by a one GoPro Hero 7 camera facing the user to capture their interactions, with a second on a tripod above and to the left of the user to provide an overview. As noted above, the participants were also in the frame of a third camera providing the video link to the wizard.

3.4 Tasks

The system (via the wizard) responded with the same task-based dialogue policy and interaction protocol in both conditions. Each condition started with a spoken introduction to the condition and task, followed by a short demonstration task with five steps to ensure that the user understood the instructions and the style interaction for the condition.

The task was to follow a twenty-step list of instructions to construct a simple Lego tower comprising 36 different colored and sized bricks. This was chosen as an instructional interaction similar to following a multi-step recipe, for example, which is a common referential communication task with voice-controlled interfaces in domestic environments [14, 15]. However, in contrast to a cooking task, there were no cascading consequences where earlier actions would influence the performance or ability to complete later steps in the Lego construction task. Keeping each step in the list of instructions atomic enabled the focus of the experiment to be placed squarely on exploring the different interaction modalities used to advance through the list of instructions and explore the recovery strategies that the users employed when a failure condition was presented to them. This can also be seen as a way to normalize the experiences, expectations, and skills of the participant – much in the same way as random strings are used in typing experiments [18, 29, 54].

Participants did not need to worry about where to put the blocks, and they were instructed to build straight up. Yet, the instructions were non-trivial, so users had to interact with the agent to know what color and size of Lego block needed to be found and how many of them should be added for that particular step.

The experimental setup included a variety of blocks in five different colors (red, white, blue, yellow, and green), each presented in five sizes named for the number of connecting 'dots' on the top of the block (two, four, six, eight, and ten). Not all sizes and color combinations were used in the building of the tower.

Participants were told that they were being timed on building the tower, with a time penalty added for each missing or wrong block, and that the winner would receive a small prize on top of the 100 SEK[1] gift card for participation in the study.

After each condition had been completed the participants were asked to fill out a Subjective Assessment of Speech System Interfaces (SASSI) questionnaire [24] related to their interaction experience. The SASSI questionnaire has been used to measure the usability of diverse applications such as speech-based access to health information, social robots, and conversational agents [11, 44, 58]. The questions were presented as 7-point Likert scale items.

3.5 Participants and Procedure

12 participants were recruited from the local university by word-of-mouth (6 females and 6 males, average age were 28) and were provided gift cards as compensation for their time. Participants signed a consent form before participation and were instructed that the study was about experimenting with different ways of delivering lists of instructions with a voice agent. First, participants filled out an entry questionnaire that collected demographics and their familiarity with speech technologies and speech interfaces. Most of them (11 participants) had interacted at least once with speech technologies, and 4 had a smart speaker at home.

Then the task and the system were introduced to the participants. The experimenter explained and ensured that the participants understood that the system was based on a commercial speech agent but that the wake-word had been turned off and it had been restricted to only respond to commands that were directly in relation to navigating through the list of instructions. While the initial explanation was scripted, the experimenter checked that this was understood and provided expanded explanations where necessary. While the participants were deliberately not given example commands to use with the speech interaction condition, they were informed that most commands related to the task that the researchers had tested had worked as expected. Following this, the experimenter returned to control the system, and the first of the counter-balanced conditions commenced. Each started with an introduction to the condition and the task articulated by the system, followed by a short demonstration task with five steps to ensure that the user understood the instructions and the style interaction for the condition.

After the five-step pre-task was complete, the participants were given the option to repeat it as necessary until they were comfortable with the interaction. The main building task would then be triggered, which started with the instruction, to begin with a fresh

[1] Around 10 euro or U.S. dollars

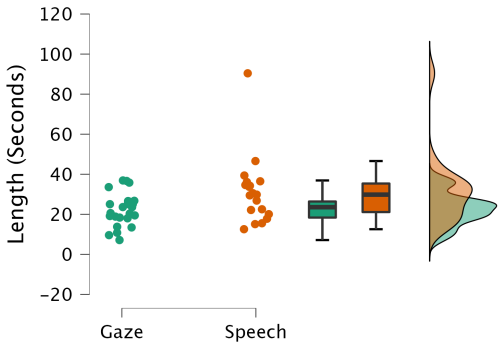


Figure 3: Modality Switches and Reformulations

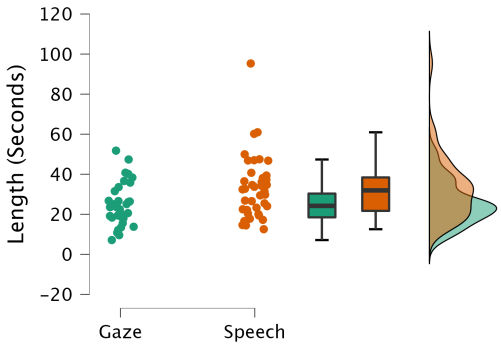


Figure 4: Induced Error Interactions

base sheet of Lego. The agent would then step through the list of twenty instructions, advancing according to the interaction paradigm of the condition. It would also provide additional task-based information if necessary, and would reply to clarification questions related to navigation through the list or repeating the previous instruction. If a user deviated from the interaction protocol (e.g., "What is the best way to build the tower?"), Tama would respond "I am sorry, I do not understand."

When the final block had been placed, the system announced that the tower had been completed. The participants then filled in a SASSI questionnaire for that interaction paradigm before repeating the task in the second condition. After the final condition and questionnaire, the participants were debriefed and informed that the system was not autonomous and that the errors were engineered to understand better how to design more robust interactions for spoken and gaze interaction.

3.6 Data

Beyond the demographic questionnaire, there were three types of data collected for each participant: the video recording of their interactions with the system, the system logs of actions and interactions, and the SASSI questionnaire answers for each condition. Each interaction was timestamped, transcribed, and coded depending on its condition and the outcome of the interaction. Essentially, for every participant we acquired a log of what was said and what was recognised by the system for both gaze and speech conditions.

Due to technical errors, of the 12 participants who completed the study, we collected 10 full sets of interaction logs and video data for 12 Speech interaction conditions and 11 Gaze interaction conditions. As a result, the following quantitative results are based on 10 participants, and the video analysis is based on 11.

Table 3: Modality Switches and Reformulations

	Condition	N	Mean	SD	SE
Time	*Gaze*	25	22.414	8.086	1.617
(seconds)	*Speech*	19	31.248	17.039	3.909

4 RESULTS

In reporting the results, focus initially on the hypothesis outlined in the introduction, namely that A)*Users are able to advance through a list using gaze* and that B)*Switching modalities* from gaze to voice in the gaze interaction condition to deal with an error would take longer and be more frustrating to users than reformulating a spoken command. The assumption is that the code-switch involved in changing interaction modality would take additional time and be more taxing for the participant [21, 30, 67].

In examining these interactions, we look at the length of time taken to complete the building steps. Following from previous work [41], we use this length as a proxy for success – where successfully interacting with the system takes less time due, in part, to the lack of hesitations and repetitions. Outside of the engineered fail states, there were no interactions where (after the initial training stage) the participants were unable to advance the list on command.

4.1 Advancing a List with Gaze

In taking each interaction as an independent point, when the interaction went smoothly, we are able to show that there is no significant difference between the Gaze and Speech conditions (Welch's t-test, t−1.054, p0.293). In total, there were 340 successful interactions included from the 10 recorded tests.

This result supports the hypothesis that taking advantage of cued-gaze to advance through a list of instructions with a speech agent is a viable interaction strategy.

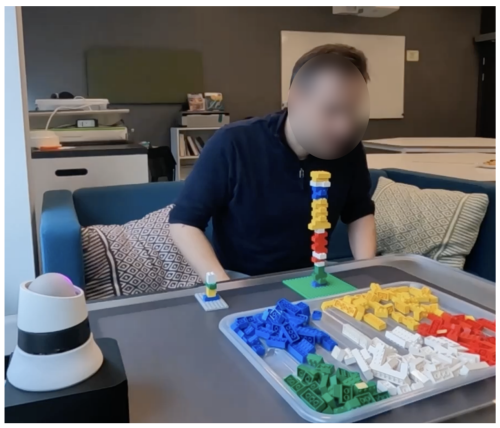
4.2 Modality Switches and Reformulations

During each condition, there were three error interactions induced, up to a total of five times, in which the agent either repeated the previous utterance or produced an error noise until the participant

Table 4: Reformulations and Repair

	Test	t	df	p	Cohen's d
Length	*Gaze < Speech*	−2.088	24.158	0.024	−0.662
(Seconds)	two-sided	−2.088	24.158	0.047	−0.662
	Speech < Gaze	−2.088	24.158	0.976	−0.662

Note. Welch's t-test.



P: Next step↑
S: *beep*
P: (0.1)Next(.)step
S: *beep*
P: (1.8) Ne:xt st:ep
S: *beep*
P: (1.9) What is the next step_
S: *beep*
P: (2.7) Next step, please↑
S: *beep*
P: (2.5) Do you understand?
S: I am sorry I do not understand
P: (0.4) The (.) next (.) step please_
S: *beep*
P: (4.2) Go to Step 15↑

Figure 5: Reformulating Speech

changed the way in which they attempted to advance the list – either changing modalities from gaze to speech, or reformulating the spoken command.

In total, there were 30 errors induced in the 10 trials included in the quantitative analysis (Table 2). While there was a longer mean length of time taken to recover from errors in the speech condition than in the gaze condition (as can be seen in Table 2 and Figure 4), there was no significant difference between the time taken to complete the interactions in the gaze and speech conditions overall (Welch's t-test, t−2.540, p0.993).

However, of these, not all resulted in the participants reformulating in the allotted 5 attempts. In the speech condition, 5 repeated interactions and 6 errors were not reformulated, and in the gaze condition, 2 repeated interactions and 4 errors ran for the full 5 attempts without the user recovering. This resulted in 25 gaze interactions, and 19 speech interactions that included a successful recovery, the details of these interactions can be seen in Table 3 and Figure 3. Looking at the timings of just these interactions, there was a statistically significant difference (Table 4, p 0.024), indicating that these interactions were shorter in the gaze condition than in the speech condition (all be it with a small effect size).

This disproved our initial hypothesis that the context switch involved in changing modality from gaze to speech would take longer than staying with the same modality and reformulating the spoken command. In order to understand why this happened, we examined the videos of *how* the participants recovered from the errors in the interactions.

4.3 Reformulation in List Control

In the transcript from P5 shown in Figure 5 we can see one issue with the comparison conducted in the quantitative analysis above. In designing the wizard, we opted for reformulation of the command to take the form of changing the verb in the intent uttered. However, as we can see from P5's attempt to recover from the 3rd error they encountered, there are multiple strategies available to attempt to repair interactions with a failing speech agent. After the error tone, P5 over enunciates the same command in the first two attempts, first slightly, then with more emphasis on 'N' in 'Next Step'. When this fails, the participant opts for lengthening the utterance and

P: *looks to system*
S: Step 7 place a blue six
P: *adds block, looks to system*
S: Step 7 place a blue six (first rep)
P: *Picks up block and pauses (4.3)
with it over the tower*(shown)*before
replacing the block in the pile and
looking again*
S: Step 7 place a blue six
P: (3.1) Next↑
S: Step 8 add a blue four

Figure 6: Switching Modalities

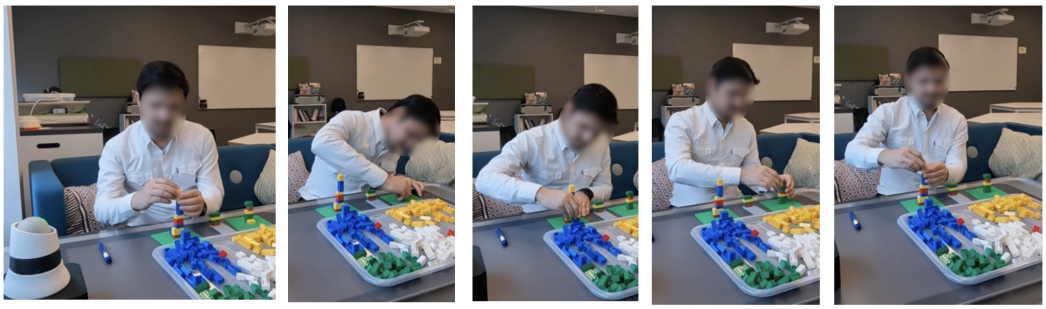


Figure 7: Misunderstanding the Error

attempts 'What is the next step' and adds a 'please'. Asking 'Do you understand?' and getting a standard reply, the participant then reformulates to match the wizard's criteria with the command 'Go to step 15.'

4.4 Switching Modalities

As we can see from the example from P3 shown in Figure 6, when the participants switched modalities to advance through the list, the flow of interaction worked well. In this case, P3 notices that there is a repetition happening the first time the command is repeated triggers another repetition with gaze, which she doesn't act upon, then verbalized the command 'Next.' This was the first induced error in the gaze condition for this participant, although in this case Gaze was the second of the conditions they experienced meaning that they were primed to be aware of repeated steps. In contrast to the multiple attempts to ensure understanding with small changes in the spoken commands in the previous section, here the switch of modalities is a clear change in interacting with the system.

Of the 11 participant videos analyzed, 7 interacted with the system in this way when their gaze failed to advance the list as expected. In both gaze and speech conditions, there were a number of different ways that the system 'going wrong' went wrong.

4.5 Understanding Errors

For a surprising number of our participants, the instructions that they should build the 'correct' tower based on the twenty instructions they were to be given by the system combined with the step number being stated at the start of each one wasn't enough to encourage enough attention to notice the repetitions. Of the 7 unrecovered repetition interactions in 3 cases, the participants added the same block 5 times.

Even when they noticed that there was something wrong with the interaction, in the case of the error sounds rather than repeated spoken instructions or noticing that the repetition is taking place, a number of participants would simply *brute force* the error. Looking or speaking in exactly the same way until they listened to the error 5 times and the system progressed.

Another challenge for users is understanding the errors, this can be seen in Figure 7. In all of the error interactions for P4 the participant didn't attempt to repair the interaction with the system. Instead, they took the repetition and error noises to mean that *they* had made a mistake in building their Lego tower. As we can see in this example, each time the agent produced an error noise, the participant moved the previously placed block around his construction until the maximum five errors had been presented to them. To a lesser extent, this behaviour was also seen in other participants, where for P8 their initial reaction to the repetition was to attempt to more firmly connect the previous Lego block to the tower and for P1 and P9 to demonstrate their 'correctness' by stating out-loud the contents of the previous command 'It was blue four, I'm sure' *(P9)* for the wizard, camera, or robot to hear.

Table 5: Mean-SD table for SASSI main effects, and t-test results.

	Gaze		Speech				
Factor name	Mean	SD	Mean	SD	t	df	p
System response accuracy	4.33	1.81	4.08	1.89	0.94	22	0.36
Likeability	5.37	1.56	5.27	1.58	0.18	22	0.86
Cognitive demand	3.45	2.02	3.28	1.74	0.97	18	0.17
Annoyance	4.07	2.06	3.42	2.09	1.01	20	0.32
Habitability	3.85	1.76	3.35	1.90	1.35	22	0.19
Speed	5.63	1.35	5.79	1.25	-0.73	21	0.47

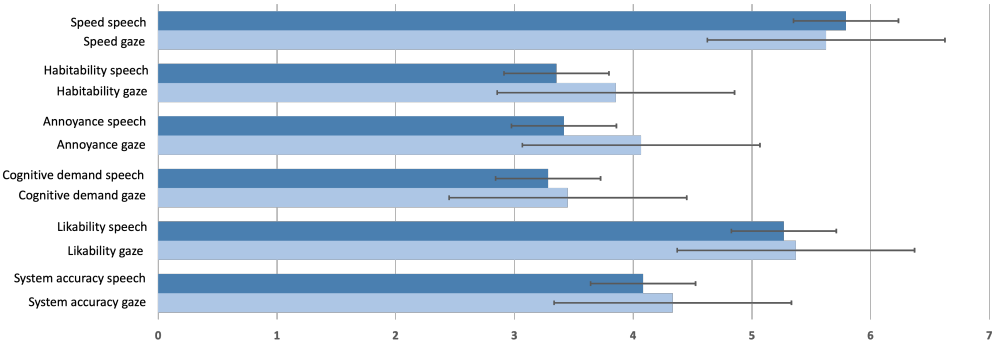


Figure 8: SASSI mean ratings bar chart. Error bars indicate standard error of the mean.

4.6 User Perception of the Interactions

It is worth noting that the SASSI questionnaire contains some items with a negative weight. Therefore, when coding the rankings, such items have to be reversed to ensure consistency. The final ratings for each of the 6 factors presented in Table 5 are computed by averaging the corresponding items within that factor. Consequently, after reversal of ratings and averaging, we can state that the higher the rating for a factor, the more positive it was perceived to be. The SASSI scores analysis was based on the data from all 12 participants, where each participant interacted with the Tama robot in both Gaze and Speech conditions to complete 20 instructions. 7 participants interacted using Gaze first.

To determine if the interaction condition was affecting the subjective SASSI ratings of the participants, a series of t-tests were conducted. The measurements were the six factors from the SASSI questionnaire: system response accuracy, likeability, cognitive demand, annoyance, habitability, and speed. Table 5 and Figure 8 shows the mean and standard deviation for each measurement. Gaze was evaluated as better in all factors except for speed; however, the t-tests confirmed that these differences were not significant in any of the factors. This indicates that participants have perceived interaction with speech and gaze similarly in all types of interactions (i.e., successful interactions and those in which we induced failure).

5 DISCUSSION

There has been a lot of discussion over the years on conversational system design drawing from conversation analysis and social science (e.g., [41, 42]). Here we have shown how relatively small parts of the complex embodied and structured methods humans use to manage interactions with each other can be used to augment certain types of CUI interaction. When taking this further, we plan to add the ability to progress through lists with a cued gaze to real-world tasks in the domain of cooking. The ability to signal to the system not to read the whole of a recipe while the user is engaged in one small part of it, and without adding additional interaction steps the user must learn and perform beyond paying overt visual attention to the system holds promise, as we have shown here.

While the task performed by the participants is much less complex than the interactions uncovered in single or multi-party meal preparation (e.g., [35, 70]), the core interaction presented here provides the opportunity to be interwoven with other methods of interaction to better fit the diverse contexts and needs of users engaged in such a common and complex task.

In embedding this interaction sequence in an autonomous, or semi-autonomous, system trial in a more complex setting, we hope to uncover more methods by which gaze, and speech can be used to augment and support the progressivity of list interaction. Beyond this, when the interaction space is opened to include other intents, it will be of great interest to observe and probe where users attempt to interact in the same way – expecting certain replies to be tied to ongoing attention but not others, or observably drawing attention away from the system in an attempt to pause without being forced to talk over the speech agent.

We don't feel that the goal here should be to mimic human-human interaction. As Shneiderman [59] argues, the complexities of face-to-face communication are effective for human-human interaction, but systems often the lack human-level understanding of these complex social signals and this causes problems when they are naively applied to human-computer interaction. "By appreciating the differences between human-human interaction and human-computer interaction, designers may then be able to choose appropriate applications for human use of speech with computers" [59].

6 CONCLUSION

In conclusion, we have shown that taking advantage of cued gaze to indicate that the agent should advance through a list is a valid option for the design of gaze-enabled speech agent interaction, comparable in our controlled lab experiment to using voice commands.

In comparison with reformulating a spoken command, the counter-intuitive finding that the modality-switching cost was negligible in terms of user experience as measured by the SASSI questionnaire and significantly less in measured interaction time was surprising.

Through this we have shown that in providing multiple modalities of interaction for the same intent, CUI designers can provide users with tools to better aid them in recovering from errors. We see this as one step forward in the ongoing challenge to design learnable, robust, and universal conversational user interfaces.

REFERENCES

[1] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational Gaze Aversion for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction (HRI '14)*. ACM, New York, NY, USA, 25–32. https://doi.org/10.1145/2559636.2559666

[2] Michael Argyle, Luc Lefebvre, and Mark Cook. 1974. The meaning of five patterns of gaze. *European journal of social psychology* 4, 2 (1974), 125–136. https://doi.org/10.1002/ejsp.2420040202

[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] Nikolaus Bee, Elisabeth André, and Susanne Tober. 2009. Breaking the Ice in Human-Agent Communication: Eye-Gaze Based Initiation of Contact with an Embodied Conversational Agent. In *Intelligent Virtual Agents*, Zsófia Ruttkay, Michael Kipp, Anton Nijholt, and Hannes Högni Vilhjálmsson (Eds.). Springer, Berlin, Heidelberg, 229–242. https://doi.org/10.1007/978-3-642-04380-2_26

[5] Nikolaus Bee, Johannes Wagner, Elisabeth André, Thurid Vogt, Fred Charles, David Pizzi, and Marc Cavazza. 2010. Discovering eye gaze behavior during human-agent conversation in an interactive storytelling application. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/1891903.1891915

[6] Linda Bell and Joakim Gustafson. 1999. Interaction with an animated agent in a spoken dialogue system. In *Sixth European Conference on Speech Communication and Technology*. Citeseer.

[7] 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*. 1–13.

[8] Timothy Bickmore. 2002. Towards the design of multimodal interfaces for handheld conversational characters. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. Association for Computing Machinery, New York, NY, USA, 788–789. https://doi.org/10.1145/506443.506598

[9] Timothy Bickmore and Justine Cassell. 2000. How about this Weather? Social Dialogue with Embodied Conversational Agents. (2000), 5.

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

[11] Dan Bohus and Alexander I. Rudnicky. 2008. Sorry, I Didn't Catch That! In *Recent Trends in Discourse and Dialogue*, Laila Dybkjær and Wolfgang Minker (Eds.). Springer Netherlands, Dordrecht, 123–154. https://doi.org/10.1007/978-1-4020-6821-8_6

[12] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, and Janet F. McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics* 42, 9 (Sept. 2010), 2355–2368. https://doi.org/10.1016/j.pragma.2009.12.012

[13] Allison Bruce, Illah Nourbakhsh, and Reid Simmons. 2002. The role of expressiveness and attention in human-robot interaction. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, Vol. 4. 4138–4142 vol.4. https://doi.org/10.1109/ROBOT.2002.1014396

[14] 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

[15] 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

[16] 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

[17] Stefania Druga, Randi Williams, Cynthia Breazeal, and Mitchel Resnick. 2017. "Hey Google is It OK if I Eat You?": Initial Explorations in Child-Agent Interaction. In *Proceedings of the 2017 Conference on Interaction Design and Children (IDC '17)*. ACM, New York, NY, USA, 595–600. https://doi.org/10.1145/3078072.3084330

[18] Anna Maria Feit, Daryl Weir, and Antti Oulasvirta. 2016. How We Type: Movement Strategies and Performance in Everyday Typing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4262–4273. https://doi.org/10.1145/2858036.2858233

[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] Alexandra Frischen, Andrew P. Bayliss, and Steven P. Tipper. 2007. Gaze Cueing of Attention. *Psychol Bull* 133, 4 (July 2007), 694–724. https://doi.org/10.1037/0033-2909.133.4.694

[21] Chiara Gambi and Robert J. Hartsuiker. 2016. If you stay, it might be easier: Switch costs from comprehension to production in a joint switching task. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 42, 4 (2016), 608–626. https://doi.org/10.1037/xlm0000190

[22] Erving Goffman. 1964. The Neglected Situation. *American Anthropologist* 66, 6_PART2 (Dec. 1964), 133–136. https://doi.org/10.1525/aa.1964.66.suppl_3.02a00090

[23] Charles Goodwin. 1980. Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. *Sociological Inquiry* 50, 3-4 (July 1980), 272–302. https://doi.org/10.1111/j.1475-682X.1980.tb00023.x

[24] Kate S. Hone and Robert Graham. 2000. Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering* 6, 3-4 (2000), 287–303. Publisher: Cambridge University Press.

[25] Shanee Honig and Tal Oron-Gilad. 2018. Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology* 9 (2018), 861.

[26] Ryo Ishii and Yukiko I. Nakano. 2008. Estimating User's Conversational Engagement Based on Gaze Behaviors. In *Intelligent Virtual Agents*, Helmut Prendinger, James Lester, and Mitsuru Ishizuka (Eds.). Springer, Berlin, Heidelberg, 200–207. https://doi.org/10.1007/978-3-540-85483-8_20

[27] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding* 108, 1 (Oct. 2007), 116–134. https://doi.org/10.1016/j.cviu.2006.10.019

[28] 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

[29] Bonnie E. John and David E. Kieras. 1994. *The GOMS Family of Analysis Techniques: Tools for Design and Evaluation,*. Technical Report. Carngie-Mellon University Pittsburgh PA Dept Of Computer Science. https://apps.dtic.mil/sti/citations/ADA309174

[30] Florian Kattner, Larissa Samaan, and Torsten Schubert. 2019. Cross-modal transfer after auditory task-switching training. *Mem Cogn* 47, 5 (July 2019), 1044–1061. https://doi.org/10.3758/s13421-019-00911-x

[31] Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz. 2004. Identifying the Addressee in Human-human-robot Interactions Based on Head Pose and Speech. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04)*. ACM, New York, NY, USA, 144–151. https://doi.org/10.1145/1027933.1027959

[32] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26 (Jan. 1967), 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

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

[34] Spyros Kousidis and David Schlangen. 2015. The Power of a Glance: Evaluating Embodiment and Turn-Tracking Strategies of an Active Robotic Overhearer. In *2015 AAAI Spring Symposium Series*. https://www.aaai.org/ocs/index.php/SSS/SSS15/paper/view/10261

[35] 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

[36] Stephen R. H. Langton, Roger J. Watt, and Vicki Bruce. 2000. Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences* 4, 2 (Feb. 2000), 50–59. https://doi.org/10.1016/S1364-6613(99)01436-9

[37] 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

[38] Manja Lohse, Katharina J. Rohlfing, Britta Wrede, and Gerhard Sagerer. 2008. "Try something else!" — When users change their discursive behavior in human-robot interaction. In *2008 IEEE International Conference on Robotics and Automation*. 3481–3486. https://doi.org/10.1109/ROBOT.2008.4543743

[39] Gale M. Lucas, Jill Boberg, David Traum, Ron Artstein, Jonathan Gratch, Alesia Gainer, Emmanuel Johnson, Anton Leuski, and Mikio Nakano. 2018. Getting to Know Each Other: The Role of Social Dialogue in Recovery from Errors in Social Robots. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI '18)*. Association for Computing Machinery, New York, NY, USA, 344–351. https://doi.org/10.1145/3171221.3171258

[40] 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

[41] 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

[42] Robert J. Moore and Raphael Arar. 2018. Conversational UX Design: An Introduction. In *Studies in Conversational UX Design*, Robert J. Moore, Margaret H. Szymanski, Raphael Arar, and Guang-Jie Ren (Eds.). Springer International Publishing, Cham, 1–16. https://doi.org/10.1007/978-3-319-95579-7_1

[43] Tova Most. 2002. The use of repair strategies by children with and without hearing impairment. (2002).

[44] Sebastian Möller, Roman Englert, Klaus-Peter Engelbrecht, Verena Hafner, Anthony Jameson, Antti Oulasvirta, Alexander Raake, and Norbert Reithinger. 2006. *Memo: towards automatic usability evaluation of spoken dialogue services by user error simulations.*

[45] 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

[46] 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

[47] Florian Pecune, Jingya Chen, Yoichi Matsuyama, and Justine Cassell. 2018. Field Trial Analysis of Socially Aware Robot Assistant. (2018), 9.

[48] 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

[49] Ulrich J. Pfeiffer, Kai Vogeley, and Leonhard Schilbach. 2013. From gaze cueing to dual eye-tracking: Novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews* 37, 10, Part 2 (Dec. 2013), 2516–2528. https://doi.org/10.1016/j.neubiorev.2013.07.017

[50] 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

[51] Martin Porcheron, Joel E. Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?": Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. ACM, Portland, Oregon, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[52] Laurel D. Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction* 1, 1 (2012), 119–136.

[53] Harvey Sacks, manuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn Taking for Conversation. *Language* 50 (1974), 696–735. https://doi.org/10.2307/412243

[54] Timothy A. Salthouse. 1986. Perceptual, cognitive, and motoric aspects of transcription typing. *Psychological Bulletin* 99, 3 (1986), 303–319. https://doi.org/10.1037/0033-2909.99.3.303

[55] Emanuel A. Schegloff. 1968. Sequencing in Conversational Openings1. *American Anthropologist* 70, 6 (Dec. 1968), 1075–1095. https://doi.org/10.1525/aa.1968.70.6.02a00030

[56] Derick Schweppe. 2022. Mycroft – The Open Source Privacy-Focused Voice Assistant. https://mycroft.ai/

[57] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[58] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld. 2007. HealthLine: Speech-based access to health information by low-literate users. In *2007 International Conference on Information and Communication Technologies and Development*. 1–9. https://doi.org/10.1109/ICTD.2007.4937399

[59] Ben Shneiderman. 2000. The limits of speech recognition. *Commun. ACM* 43, 9 (Sept. 2000), 63–65. https://doi.org/10.1145/348941.348990

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

[61] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of Style in Information Seeking Conversation with an Agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, 1171–1180. https://doi.org/10.1145/3397271.3401127

[62] Kristinn R. Thorisson, David B. Koons, and Richard A. Bolt. 1992. Multi-modal natural dialogue. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '92*. ACM Press, Monterey, California, United States, 653–654. https://doi.org/10.1145/142750.150714

[63] David Traum and Jeff Rickel. 2002. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*. 766–773.

[64] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173782

[65] Geraldine P. Wallach and Katharine G. Butler. 1994. *Language learning disabilities in school-age children and adolescents: Some principles and applications.* Allyn & Bacon.

[66] Sebastian Wallkötter, Michael Joannou, Samuel Westlake, and Tony Belpaeme. 2017. Continuous Multi-Modal Interaction Causes Human-Robot Alignment. In *Proceedings of the 5th International Conference on Human Agent Interaction - HAI '17*. ACM Press, Bielefeld, Germany, 375–379. https://doi.org/10.1145/3125739.3132599

[67] Christopher D. Wickens and Yili Liu. 1988. Codes and Modalities in Multiple Resources: A Success and a Qualification. *Hum Factors* 30, 5 (Oct. 1988), 599–616. https://doi.org/10.1177/001872088803000505

[68] 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

[69] Tomoko Yonezawa, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. 2007. Gaze-communicative Behavior of Stuffed-toy Robot with Joint Attention and Eye Contact Based on Ambient Gaze-tracking. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI '07)*. ACM, New York, NY, USA, 140–145. https://doi.org/10.1145/1322192.1322218

[70] 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. *(CHI '22)*. New York, NY, USA. https://doi.org/10.1145/3491102.3502036