Dharan et al., Gerontechnology (2021) 20:2
https://doi.org/10.4017/gt.2021.20.2.436.06

Speech-based prompting system to assist with activities of daily living: A feasibility study

Neeraja Murali Dharan MS[a,*], Muhammad Raisul Alam PhD[a,b,c], Alex Mihailidis PhD PEng[b,c]

*[a]Department of Computer Science, University of Toronto, Toronto, ON, Canada; [b]Department of Occupational Science and Occupational Therapy, University of Toronto, Toronto, ON, Canada; [c]KITE Research, Toronto Rehabilitation Institute, University Health Network, Toronto, ON, Canada; *Corresponding author: n.murali@mail.utoronto.ca*

Abstract

Background: The onset of dementia can negatively affect an individual's ability to perform activities of daily living (ADL). Consequently, these individuals rely on formal or informal caregivers for ADL completion. Cognitive assistive technologies (CATs) can help adults with dementia during ADL performance and potentially alleviate caregiver burden. However, sensors that are commonly used by CATs (i.e., cameras, accelerometers, radio frequency identification tags) have some limitations; they can cause erroneous activity detection if not positioned properly, installing them can be a long and expensive process, and their usage can raise privacy concerns.
Objective: In this study, spoken dialogue was explored as an alternative method of providing input to CATs. The feasibility of using a speech-based CAT to assist with ADL performance was evaluated.
Method: A speech-based intelligent prompting system for ADL assistance was developed. This system was a modification of the Cognitive Orthosis for Assisting with aCtivities in the Home (COACH), a CAT that was designed to assist with the ADL of handwashing. The camera-based hand-tracking sensor in the COACH was replaced with a speech-based conversational agent. A study with 10 healthy adults was conducted to test the feasibility of using speech instead of camera input for the COACH. Results were compared to the outcomes of a previous study with the camera-based COACH.
Results: Results suggested that the speech-based COACH was able to identify completed task steps with 97.5% accuracy and true participant performance with 90% accuracy. In most cases, the performance of the COACH was not negatively affected by the use of speech in place of camera data.
Conclusion: This study is a promising first step in exploring the application of speech in CATs designed for ADL assistance. The results of this paper support the feasibility of using speech as input for CATs and highlight its potential in replacing sensors.

Keywords: Dementia, cognitive assistive technology, human-centered computing, conversational interface, activities of daily living (ADL)

Introduction

Dementia is a neurodegenerative disease characterized by a collection of symptoms, which may include a decline in memory, reasoning, and communication skills (Henderson and Jorm, 2002). It is estimated that as of 2019, over 50 million people are living with dementia globally, and this number is set to increase to 152 million by the year 2050 (Evans-Lacko et al., 2019). The current annual cost of dementia in the United States of America alone is estimated to be 1 trillion USD, a figure that is expected to double by 2030 (Evans-Lacko et al., 2019).

The onset of dementia has many negative effects on an individual's quality of life. For example, he or she may experience a gradual loss of skills needed to perform basic activities of daily living (ADL), such as bathing, feeding, and toileting (Henderson and Jorm, 2002, Evans-Lacko et al., 2019). As a result, dementia patients become highly dependent on formal (paid) or informal (unpaid) caregivers for support. In most cases, a close family member or friend must act as the main caregiver (Chiao et al., 2015). As informal caregivers assume responsibility for a person with progressive dementia, they are prone to experiencing increased stress and burden (Evans-Lacko et al., 2019, Chiao et al., 2015). In a global survey conducted by the Alzheimer's Disease International in 2019, over 50% of dementia carers said that their health suffered as a result of their caring responsibilities, and over 60% experienced a decline in their social lives. On the other hand, dementia patients can feel frustrated due to the loss of independence and autonomy (Mihailidis et al., 2003).

Assistive technologies (ATs) can compensate for a broad range of physical and cognitive impairments (Tao et al., 2020). One subset of ATs termed cognitive assistive technologies (CATs), can support users' cognitive functioning during task performance (Tao et al., 2020). Specifically, some CATs can help adults with dementia perform ADL more independently (Buettner et al., 2012). These devices have the potential to alleviate caregiver burden, delay long-term care admission and improve a patient's quality of life (Buettner et al., 2012). A number of studies have looked into implementing CATs for ADL assistance, employing a variety of sensors for activity monitoring, including accelerometers, radio-frequency identification (RFID) sensors, and cameras. (Boger et al., 2005, Czarnuch et al., 2013, Hattink et al., 2016, Hoey et al., 2010, Pires et al., 2018, Mihailidis et al., 2008, Philipose et al., 2003).

However, there are some limitations to using sensors in CATs. For example, installing sensors in a dementia patient's home can be an expensive and inconvenient process (Hattink et al., 2016, Rudzicz et al., 2015). The positioning of sensors in the environment can negatively affect their ability to identify objects and people, causing erroneous activity detection (Czarnuch et al., 2013, Pires et al., 2018). Lastly, certain sensors (i.e., cameras) can raise privacy issues among adults with dementia and their caregivers, as they may feel uncomfortable with being monitored at all times (Hattink et al., 2016, Boise et al., 2013). Consequently, there exists a need to explore alternative methods of obtaining input data for activity monitoring.

Cognitive Assistive Technologies for Activities of Daily Living: Limitations of Current Activity Tracking Methods

Philipose et al. developed PROACT in 2003, a system that uses Radio Frequency Identification (RFID) technology to recognize ADL. The concept of PROACT is as follows: objects in the environment are tagged with RFID tags and users are asked to employ RFID tag readers when interacting with the tagged objects. Based on this interaction, PROACT can deduce the ADL being performed. A study was conducted with 14 participants to test the feasibility of PROACT. During the study, participants were asked to select and perform 12 ADL. It was found that PROACT correctly inferred the occurrence of an activity 88% of the time (Philipose et al., 2003). However, tags that were placed on refrigerator handles, soap bottles, and faucets had lower detection rates. This was because water and metal absorbed the radio waves that the RFID-tags used, and metal short-circuited the tag antenna. As a result, the system's accuracy suffered when trying to detect the activities of handwashing, making a snack, and preparing a drink. Activities with common prefixes also posed a problem for PROACT. For example, toileting, handwashing, and maintaining oral hygiene all involved the user entering the washroom and interacting with the light switch. The system had difficulty disambiguating these activities when performed together.

Hattink et al. (2016) designed the Rosetta system to support ADL performance for people with dementia. The Rosetta system consists of 3 subsystems, the Elderly Day Navigator (EDN) for providing activity reminders, the Early Detection System (EDS) which tracks behaviour patterns in the context of daily living using sensors, and the Unattended Autonomous Surveillance system & Advanced Awareness and Prevention System (UAS-AAPS) which uses a camera to detect emergencies (e.g., falls). The usefulness and user-friendliness of the Rosetta system were evaluated with 42 persons with dementia and 32 informal caregivers. The system was installed in participants' homes for a period ranging from half a month to 8 months. Overall, dementia patients and caregivers found the Rosetta system very useful and said that it provided feelings of safety and comfort. However, its user-friendliness was not highly rated because it was perceived as being too complex. Patients also felt uneasy being in the presence of sensors at all times. In some cases, the UAS-AAPS system was not activated at all since participants did not want cameras installed in their homes. In addition, the sensors used by the system caused many technical issues, which led to the system being turned off multiple times throughout the study. The authors also mentioned that it was very difficult to plan the installation of the system since it involved 2 full days of technicians visiting the home of the person with dementia.

Lastly, the Cognitive Orthosis for Assisting with aCtivities in the Home (COACH) is an orthotic device that guides adults with moderate to severe dementia through the activity of handwashing (Czarnuch et al., 2013, Mihailidis et al., 2008, Boger et al., 2005, Hoey et al., 2010). The system uses a hand tracker to monitor users non-intrusively. The hand tracker processes images captured using an overhead camera mounted above the sink. Hand positions and interactions between hands and objects are used to determine action observations, which are passed to the planning module. The planning module is responsible for translating these action observations into handwashing task steps and for monitoring the user's progress. Based on the output from the planning module, COACH can either continue to observe the user or provide up to 4 prompts with increasing levels of support: a low-guidance verbal prompt, a high-guidance

verbal prompt, a video demonstration, or a call to the caregiver. A clinical trial with 6 participants showed that they were able to complete 11% more steps independently with COACH, decreasing interaction with caregivers to an average of 60% (Mihailidis et al., 2008).

However, when COACH was deployed in an unsupervised state without individual calibration or configuration by Czarnuch et al. (2013), several limitations were identified with the hand tracker. Tracking failed on bald users, users with rolled-up sleeves, and on darker-toned skin, since the hand tracker relied on skin colour for tracking. The system also failed to correctly track users if they moved out of the soap regions, removed the towel from the towel region, or moved their hands over tracking regions without any interaction since the overhead camera was positioned to track only specific areas of the sink and did not use depth for tracking. These tracking errors lowered the overall accuracy of the COACH, impacting its efficacy.

Research aim

Speech is the easiest and most natural form of communication, making it a highly usable and intuitive choice for human-machine interactions (Michalakis and Caridakis, 2017, Moore, 2005). In recent years, natural language processing has been extensively applied to enhance practical technology, causing a major shift towards voice-based user interfaces. Following this trend, companies like Google and Amazon have designed voice assistants (e.g., Google Assistant, Amazon Alexa) that are capable of processing human speech and responding accordingly via synthesized voices. Users can verbally instruct these assistants to perform a variety of tasks, such as answering questions, setting reminders, and managing their schedules (Vtyurina and Fourney, 2018). Moreover, users can develop their own voice-based applications.

Using speech as the primary method for activity monitoring can allow CATs to overcome existing limitations. A speech-based CAT can be deployed easily via smartphones and smart speakers like Google Home and Amazon Echo, which are commercially available at a fraction of the cost of specialized sensors. The installation process for these devices is minimal and can be easily performed by a caregiver. Using these devices also eliminates the need to install more intrusive methods of monitoring, providing some relief to users who are concerned about their privacy. Another advantage is that even though these devices still have to be close to the user, their exact positioning has less of an impact on accurate activity detection compared to cameras and RFID tags.

Additionally, a speech-based CAT for ADL assistance can be beneficial for adults with dementia. Due to the increasing popularity of smartphones and smart speakers, they are now ubiquitous in our daily lives. It is estimated that in 2020, 81.3% of the Canadian population (Rody, 2018) will use a mobile device and 17.9% (Cakebread, 2019) will use a smart speaker. Therefore, many older adults and their caregivers might already be using these devices or at the very least, be familiar with their existence. This can significantly reduce the intimidation that one may feel when using a new assistive device. In a systematic review conducted by Evans et al. (2015) on dementia-focused technology, the authors reported that most current systems were not effective in real-life situations due to the stigma associated with relying on a device for support. Since devices like smartphones and smart speakers are used by individuals regardless of whether or not they have a cognitive impairment, their use can potentially reduce this stigma. Furthermore, a speech-based CAT can be beneficial because it models verbal support provided by a caregiver (O'Neill and Gillespie, 2008). In doing so, it minimizes the learning curve that is typically associated with using a CAT and promotes social interaction.

Despite the potential benefits of speech-based CATs, little empirical work has been done to develop them. An example of work performed in this area is the General User Interface for Disorders of Execution (GUIDE) by O'Neill et al (2010). GUIDE is a context-aware prompting system that uses verbal prompts and questions to guide users through a task. It accepts verbal feedback from users in the form of 5 one-word commands. Trials were conducted to test the efficacy of GUIDE in scaffolding 8 transtibial amputees with a cognitive impairment through the process of donning prosthetic limbs. Results showed that using GUIDE reduced omissions and errors in 6 participants (O'Neill et al., 2010). Further studies with GUIDE involving 1 individual with vascular dementia showed that the participant adapted to the system during the first session itself (O'Neill and Gillespie, 2008). This supported GUIDE's goal of minimizing the learning curve associated with using an assistive device. However, its limited speech recognition capability was a major drawback, as users were limited to using only 5 words to communicate with it, which can be unnatural and restrictive.

Study objectives

The overall objective of this research is to study the feasibility of using a speech-based CAT, in the form of an intelligent prompting system, to assist adults who have dementia with ADL performance. Specifically, this paper should answer the following questions related to the system:

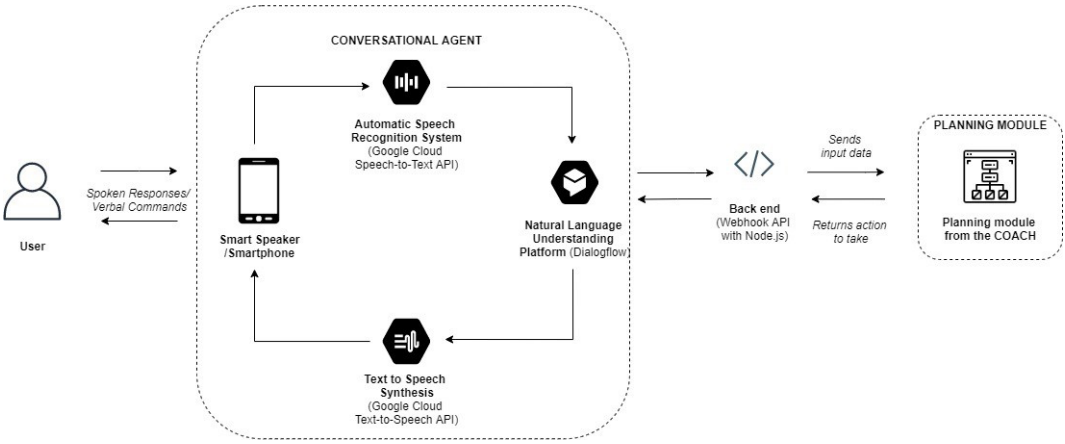

Figure 1. Architecture of the speech-based COACH.

(1) Can the prompting system use spoken language input in place of sensor data to identify whether the correct steps are being completed?
(2) Can the prompting system use spoken language input in place of sensor data to determine the user's task performance?
(3) How will the prompting system's overall system performance be affected when spoken language input replaces sensor data?

To answer these questions, a prototype for a speech-based prompting system was developed. Handwashing was chosen as the focus of this research due to the availability of previous studies on this particular ADL (Czarnuch et al., 2013, Hoey et al., 2010, Mihailidis et al., 2008). The starting point for this paper was the prototype of the COACH by Mihailidis et al. (2008). The COACH was modified by replacing the hand tracking sensor with a speech-based conversational agent. The new system, which will be referred to as the speech-based COACH, still uses the same planning module as the original system for task step and user state estimations, but instead of using sensors, it obtains action observations by conversing with the user. The system architecture for the speech-based COACH is described in the Methods section. A study was conducted with healthy adults to evaluate the feasibility of the system. Results obtained from this study were compared to the system outcomes from a previous study with the COACH conducted by Czarnuch et al. (2013).

Methods

Participants

Study participants consisted of adults over the age of 18. The inclusion/exclusion criteria for participation included the ability to speak and understand English fluently, stand freely at a sink, and verbally communicate with the speech-based COACH while completing the task of handwashing.

System

A conceptual representation of the speech-based COACH is presented in *Figure 1*. The system has 2 main components: the conversational agent, and the planning module which was adapted from Mihailidis et al. (2008). The 2 components are integrated via the use of a webhook service. The conversational agent verbally communicates with the user to infer what he or she is doing. This is done by asking questions related to task performance. The user's responses are mapped to action observations, which are passed to the planning module. The planning module then estimates the current step that the user is performing and provides an action to take. If the action suggested by the planning module is observation (i.e., doing nothing), the agent continues to pose questions to the user about their task status. If the action suggested is providing a prompt, it is conveyed to the user before the agent continues the conversation. Each component is described in more detail in this section.

Conversational agent

The conversational agent was built with Dialogflow, an end-to-end development suite by Google. Dialogflow agents use unique identifiers called intents to categorize a user's intentions and actions, based on Dialogflow Contexts and training phrases. Specific intents can be triggered based on the user's responses. Dialogflow Contexts provide the agent control over which intents to activate at specific points in the conversation while training phrases dictate which activated intent to trigger based on the user's response (Google Cloud, 2020a). Once an intent is triggered, the actions mapped to it are executed. The agent can be configured to provide dynamic responses through the use of fulfillment (Google Cloud, 2020b). When fulfillment is enabled, the agent responds to a triggered intent by calling an external webhook service. Developers can create their webhook services to perform customized actions for matched intents.

Figure 2. The intents created in Dialogflow for the 'Soap On' step in handwashing.

Dialogflow was chosen due to its comprehensiveness; in a study comparing 6 different natural language understanding platforms (Canonico and De Russis, 2018), it was identified as being the most complete for building conversational agents. Moreover, it uses Google's machine learning algorithms to understand end-user expressions and extract structured data (Google Cloud, 2020c), allowing the agent to learn from training phrases and match user expressions to intents more accurately. Applications built with Dialogflow can be deployed on a variety of Google Assistant-enabled platforms, such as Google Home, Amazon Alexa, and smartphones.

Designing conversational agent

The successful completion of handwashing requires 5 steps to be performed in the following order: turning the tap on, putting soap on, rinsing hands under the water, turning the tap off, and drying hands with a towel. Each step is represented as an intent in Dialogflow. Each of these 5 intents has nested intents that can be triggered based on the user's responses. *Figure 2* represents intents created for the 'Soap On' step. The 'soap_on_question' intent asks the user whether they have put soap on their hands. If the user indicates that he or she has completed the task, the 'soap_on_confirm' intent is triggered, and the user is asked to confirm their action. If the user confirms, the intent for the next handwashing step (i.e., 'wash_hands_question') is triggered. If the user says no, either the 'soap_on_no' or 'soap_on_confirm_no' intents are triggered. The structure presented in this example was adapted to design intents for all 5 handwashing steps.

When an intent is triggered, Dialogflow provides a verbal response to the user. The dialogue that the agent uses to converse with the user was designed based on Wilson et al.'s work (Wilson et al., 2012). In this work, the authors explored 12 formal caregivers' use of task-focused verbal strategies when assisting individuals who have Alzheimer's disease with the task of handwashing. Some strategies that led to successful task completion included the use of 1 preposition at a time, close-ended questions, and paraphrased repetition. Caregivers also frequently used the patient's name and provided verbal praise. In contrast to recommendations from clinical literature, verbatim repetition was not frequently used by caregivers.

To closely mimic the verbal assistance provided by caregivers, the above strategies were employed when designing the conversational agent's dialogue. The dialogue for the task is organized in a sequential manner (*Figure 3*), which means that a previous step must be completed before the user can move on to the next. This allows the system to focus on the completion of 1 step at a time. The conversational agent determines whether the user has completed each step by posing close-ended questions and by asking for confirmation. For instance, it asks the user questions such as, *"is the tap turned on?"* and

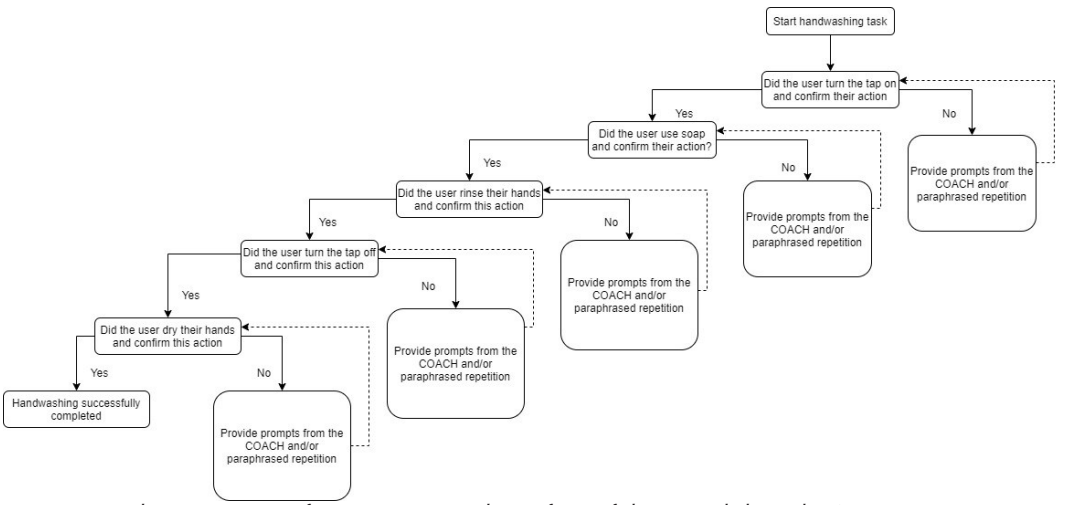

Figure 3. Dialogue structure for conversational interface of the speech-based COACH.

"have you dried your hands?". The user can respond to these questions using natural language. If the user says that he or she has completed a task, the system asks for confirmation. For instance, it will ask the user questions like, *"are you sure the tap is turned on?"* and *"can you confirm that you have soap on your hands"*. If it seems that the user is having difficulty answering, the question is paraphrased. Instead of asking *"is the tap turned on?"*, the system will ask *"is the water running now?"*. For each of the 5 steps, 5 to 6 paraphrased repetitions can be randomly selected by the system, to avoid verbatim repetition as much as possible. A verbal prompt from the COACH's planning module may also be provided. The agent provides verbal praise once a user has completed a step successfully. Lastly, the user's name is frequently used by the agent while conversing with him or her.

Training conversational agent

Each intent was trained with task-specific training phrases. This was done to ensure that when the agent interacts with a user, it can trigger the correct intents based on how closely the user's expressions match the training phrases provided for each intent. A paradigm for measuring communication breakdown, called trouble indicating behaviour (TIB), was used as a guideline for developing task-specific training phrases (Rudzicz et al., 2015). These TIBs were used to formulate phrases that the user might say in response to each intent, specifically if they are confused or have difficulty performing a step. TIBs were used because past research showed that individuals with Alzheimer's disease were more likely to exhibit these behaviours compared to those without the disease. By training the agent with phrases that correspond to each of the 12 TIBs, the system has a better chance of detecting behaviour that indicates confusion and can assist the person with dementia when necessary. All 12 TIBs, as defined in (Rudzicz et al., 2015), and examples of what the user might say in the context of handwashing for each TIB, are presented below.

(1) Neutral or non-specific requests for repetitions (local): Minimal queries indicating non-understanding, which did not identify the problem specifically. For example:
User: What? Huh?

(2) Request for confirmation repetition with reduction: Partial repair of trouble source, often in the form of a question. For example:
Speech-based COACH: Did you rinse your hands?
User: Rinse my hands?

(3) Request for confirmation complete repetition: Recapitulatory echo questions, often with pronounce alteration. These follow a similar pattern and therefore must be distinguished from expressions of incredulity or disapproval. For example:
Speech-based COACH: John, did you turn the tap on?
User: Did I turn the tap on?

(4) Request for confirmation repetition with elaboration: Same as TIB 3, but with the inclusion of additional semantic content. For example:
Speech-based COACH: John, did you turn the tap on?
User: Did I turn the tap on by pulling on the lever?

(5) Request for specific information: Contains a specific semantic concept, content, word, or referent to the previous or recent turn. For example:
Speech-based COACH: Did you dry your hands with a towel?
User: What do I dry my hands with?

(6) Request for more information: A non-specific request (i.e. without direct mention of semantic concepts in a recent utterance). For example:
User: I don't understand. Tell me more. What do you mean?

(7) Corrections: The result of a violation in the quality of message or message inaccuracies. Here, semantic confusion often originates from the individual not indicating the TIB. For example:
Speech-based COACH: Did you turn the tap on for me?
User: No, I am putting soap on.

(8) Lack of update/lack of continuation: Verbal behaviours including (a) minimal feedback when back-channel responses indicate non-understanding or lack of contribution on topic extension; (b) overriding, where the participant does not allow the floor; and (c) topic switch, where 1 participant abruptly changes the topic. For example:
Speech-based COACH: Did you put your hand under the water?
User: Oh, it's a bit too hard for me to do that.

(9) Hypothesis formation: Guessing behaviours involving words or speaking for or on behalf of the other participant. This does not include hypothesis in the form of rhetorical questions (which are instead categorized as TIB 5). For example:
Speech-based COACH: Can you dry your hands?
User: You want me to dry my hands with a towel.

(10) Metalinguistic comment: This includes talk about talk that explicitly refers to nonunderstanding of message content, the interpersonal manner in which the message was conveyed, or the production of the message. For example:
User: I don't understand. I can't remember.

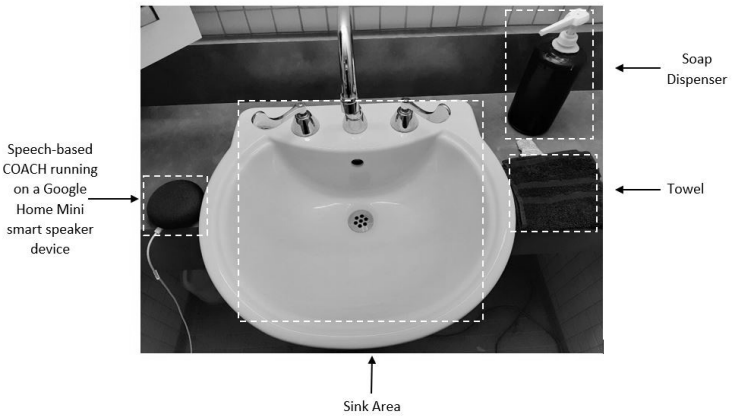

Figure 4. Study setup for the speech-based COACH.

(11) Reprise/minimal dysfluency: Reprises in which partial or whole repetition or revision of the message occurs. Minimal dysfluencies indicate difficulties producing a message that involves sound, syllable and word repetition, pauses, and fillers. These are deemed more excessive than the typical dysfluencies that occur in typical speech.
User: Errr, I want to, I rinsed my hands.

(12) Request for repetition (global): Minimal queries indicate a non-understanding of the previous section of the talk. For example:
User: Wait go back to the part about turning the tap on, you just lost me.

COACH planning module
The planning module, adapted from the version of COACH presented in (Mihailidis et al., 2008), consists of the belief monitoring system and the policy. When an intent is triggered by the conversational agent, an action observation is sent to the belief monitoring system based on the user's response. The belief monitoring system computes a belief state from the provided action observation. A belief state is a probabilistic estimation of the current state of the user and environment. Each computed belief state is passed on to the COACH's policy, which is a lookup table denoting the best course of action for the system to take based on the generated belief state. Possible actions include doing nothing, providing a low-guidance verbal prompt, providing a high-guidance verbal prompt, providing a video demonstration, or calling the caregiver. The selected action is sent back to the conversational agent for execution. A partially observable Markov Decision Process (POMDP) is used by the planning module to model the handwashing task. A POMDP was chosen because of its ability to consider different sources of uncertainty during the decision-making process.

Deployment of speech-based COACH
The conversational agent was automatically integrated with Google Cloud APIs for text-to-speech and speech-to-text capabilities. The agent can be deployed on a variety of platforms, including smartphones and smart speakers such as Google Home and Amazon Alexa. The agent is connected to the webhook service on the back end through the use of proxy software. The webhook service runs COACH's planning module and connects the two components.

Procedure
Ethics approval for the study was granted by the University of Toronto's Research Ethics Board (REB #38741). The study was conducted in a washroom environment with a working sink at the University of Toronto's Rehabilitation Sciences Building (*Figure 4*). The installation included a Google Home Mini smart speaker enabled with Google Assistant, a processing component (computer), an audio recording device, and cabling. The conversational agent ran on the Google Home device while the back end and planning module ran on a computer. Each participant performed a total of 3 trials, where each trial comprised of 1 handwashing event. Each correct handwashing event was made up of 5 steps performed in the following order: turning the tap on, putting soap on, rinsing hands, turning the tap off, and drying hands with a towel. In each trial, participants were asked to verbally communicate with the speech-based COACH while washing their hands. Particularly, they were asked to answer questions asked by the system and listen to any prompts provided to them. In the first trial, participants washed their hands normally and correctly, performing all 5 steps. In the second and third trials, participants were asked to act as if they had forgotten one of the steps of handwashing. The step to be forgotten in each trial was selected in a randomized, controlled manner and provided to participants by the study coordinator at the beginning of the trial. Participants were asked to eventually complete the 'forgotten' step and move forward with the rest of the task, whenever they chose to do so.

Data analysis
Audio recordings were collected while each participant interacted with the speech-based COACH during trials. A coding scheme was developed based on previous studies with COACH (Mihailidis et al., 2008, Czarnuch et al., 2013). The coding scheme consists of discrete participant behaviours

Table 1. Step-based effectiveness of speech-based COACH at identifying completed or missed steps.

		Number of occurrences					
Participant action	**System outcome**	**All steps**	**Turn on water**	**Use soap**	**Rinse hands**	**Turn off water**	**Dry hands**
Step completed by user	Identified complete (True Positive)	121 (80.7%)	28	20	26	22	25
	Identified incomplete (False Negative)	7 (4.7%)	0	2	1	3	1
Step missed by user	Identified incomplete (True Negative)	14 (9.3%)	2	5	0	5	2
	Identified complete (False Positive)	8 (5.3%)	0	3	3	0	2
Total steps in trials		150	30	30	30	30	30

corresponding to handwashing steps. The speech-based COACH's ability to identify individual task steps performed by users and measure overall task performance was assessed based on ground truth data obtained from these recordings.

Task step identification

Audio recordings for each participant were manually annotated to determine the true user state. The belief states generated by the system were compared to this data. COACH's step identification performance was categorized using common categories from signal detection theory (Wickens and Hollands, 2000). For each step of each trial, the speech-based COACH's performance was scored based on the user's performance and the associated system outcome. If the user performed a step and the system correctly changed state, then a True Positive was scored. Anything other than a correct state change was considered a False Negative. If the user did not perform a step, a True Negative was scored if the system did not change states, and a False Positive was scored otherwise. Based on this information, 3 categories of system performance measures were calculated: sensitivity (Equation 1), specificity (Equation 2), and accuracy (Equation 3). Sensitivity is the likelihood that the speech-based COACH will identify a completed step as complete while specificity is the likelihood that it will identify an incomplete step as incomplete. Accuracy represents the likelihood that the system will measure true user performance.

Sensitivity=(TP)/(TP+FN) (1)
Specificity=(TN)/(TN+FP) (2)
Accuracy=(TP+TN)/(Total Outcomes) (3)
TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

Overall user performance

The speech-based COACH was scored on how well it identified participants' actual performance in the task of handwashing. Participant data were organized according to the number of handwashing steps they completed during trials. Based on this, the number of trials where the speech-based COACH identified all steps completed by the users was recorded.

Results

Participants & trials

10 adults ranging from ages 24 to 50 (Mean = 29 years) consented to participate in the study. Each participant performed 3 trials, generating 30 trials in total. All trials were included in the analysis.

System performance

System performance was measured by analyzing each of the 5 steps performed during trials. 150 steps were analyzed in total. *Table 1* shows data representing the system's effectiveness at identifying whether participants completed or missed each of the 5 steps. It is important to note that even though participants were eventually asked to perform the missed step and continue with the rest of the task, the system outcome was determined by analyzing the duration for which the step remained incomplete. Of the 150 steps, participants completed 128 steps (85.4%). Of those steps, 121 (80.7%) were successfully identified as completed by the speech-based COACH (True Positive) and 7 (4.7%) were falsely identified as incomplete (False Negative). Participants were asked to miss a step in the second and third trials (20 missed steps in total). However, 2 of the participants missed additional steps during the trials, resulting in 22 missed steps (14.6%). The system identified 14 (9.3%) of these steps as being incomplete (True Negative) but falsely reported 8 (5.3%) steps as being complete (False Positive). Based on the system performance statistics, it can be concluded that the speech-based COACH can correctly identify a completed step 94.5% of the time (Sensitivity) and correctly identify a step that is not completed by the user as incomplete 63.6% of the time (Specificity). Furthermore, the system is likely to measure true participant performance 90% of the time (Accuracy).

Overall task performance

Of the 30 trials, participants completed all relevant steps and successfully washed their hands 29 times (96.7%) and the system correctly identified the task as being completed all 29 times. 1 participant failed to complete the task and triggered the 'call caregiver' feature while pretending to not turn the tap off. In this scenario, the system was able to track all 3 of the steps that the user completed (Turn on the water, put soap on,

Table 2. Comparison of system performance measures between speech-based COACH and the camera-based COACH from (Czarnuch et al., 2013).

System measure	Speech-based COACH	Camera-based COACH from (Czarnuch et al., 2013)
Sensitivity	94.5%	46.6%
Specificity	63.6%	97.5%
Accuracy	90%	55%

put hands under the water). Overall, participants completed an average of 4.93 out of 5 steps, which was also the average completed steps identified by the speech-based COACH.

Comparison with COACH

To answer the research questions proposed in this paper, the results of this study were compared to the outcomes of a previous study with the COACH by Czarnuch et al. (2013). The COACH from Czarnuch et al. (2013) will be referred to as the camera-based COACH for this discussion. The camera-based COACH used a hand tracker and the same planning module as the speech-based COACH for task step estimation. The system was configured to run in an unsupervised state in a washroom at the Toronto Memory Program. Twenty participants contributed 41 hand-washing trials to the study. System performance was reported based on this data. Even though the authors conducted their study with dementia patients as opposed to healthy adults, it was still possible to compare the results of the two studies since the focus of the comparison was on system performance, and not on user performance. The purpose of this comparison was to demonstrate the feasibility of the speech-based COACH, using the camera-based COACH as a reference.

Comparing task step completion

Table 2 shows the 3 system measures (sensitivity, specificity, and accuracy) that were obtained for the speech-based COACH and camera-based COACH from (Czarnuch et al., 2013). The likelihood of measuring completed steps was 94.5% for the speech-based COACH and 46.6% for the camera-based COACH, while the likelihood of measuring incomplete steps was 63.6% for the speech-based system and 97.5% for the camera-based system. The speech-based COACH was likely to measure true participant performance 90% of the time, while the camera-based COACH was likely to measure true participant performance 55% of the time.

Comparing overall task performance

Participants successfully washed their hands 29 times and the speech-based COACH was able to correctly identify the task as being completed all 29 times (100%). On average, participants performed 4.93 out of 5 steps, which was also the average completed steps identified by the speech-based COACH. In contrast, the camera-based COACH correctly identified the task as complete 8 (30.8%) out of the 26 times that participants washed their hands (Czarnuch et al., 2013). An additional step, take towel, was included when evaluating the performance of the camera-based COACH, which resulted in 6 total steps instead of 5. Overall, participants performed an average of 5.05 out of 6 steps but the camera-based COACH correctly identified a significantly lower average of 2.37 steps as being performed (Czarnuch et al., 2013).

Discussion

Task step performance

Overall, the speech-based COACH identified more completed steps compared to the camera-based system, with a sensitivity measure of 94.5%. This was mainly because the speech-based system allowed participants to verbally convey their progress and confirm the completion of each step, which reduced the number of False Negatives. On the other hand, the camera-based COACH only relied on observations from the hand tracker and did not receive any information directly from the user, which may have led to more False Negatives (Czarnuch et al., 2013).

Another reason for the higher sensitivity was because the exact positioning of objects around the sink did not impact the speech-based system's task step detection, but this was not the case with the camera-based system. Czarnuch et al. (2013) reported that the camera-based COACH failed to detect the completion of the 'Dry Hands' step in two trials because the towel was removed from the towel region. It was also found that the hand tracker sometimes reported hands as toggling in and out of different regions. This led to unstable action observations that were not sent to the planning module for state estimation. These technical errors were reduced when speech was used in place of the camera. As a result, user responses generated more stable observations for the planning module, allowing for more accurate tracking of task steps.

The speech-based COACH was able to identify incorrect steps 63.6% of the time, while the camera-based COACH achieved 97.5%. The speech-based system's lower performance could be attributed to the fact that users were no longer confirming their actions when a step was not completed. Instead, they only expressed that they did not complete the step. Since the system could not observe users' exact hand positions while they were stuck on a step, action observations had to be largely estimated. This might have affected the planning module's ability to determine accurate user progress. To overcome this issue, it would be beneficial to

ask users follow-up questions to determine their exact actions when a step is not completed. The camera-based COACH displayed better performance with regard to specificity because it directly observed users' actions when they forgot a step. Moreover, users tended to slow down when they were stuck on a step. Based on results from (Czarnuch et al., 2013), when the user slowed down, the system was more effective at identifying task steps. This was because slowing down created a temporal delay between the different steps, allowing the system enough time to properly detect the user's actions.

Lastly, it was found that the speech-based COACH was likely to measure true participant performance 90% of the time, while the camera-based COACH was likely to measure true participant performance 55% of the time. As mentioned previously, users' ability to confirm the steps that they performed increased the overall accuracy of the speech-based COACH because this led to action observations that were similar to ground truth data. On the other hand, many of the limitations that were associated with the hand tracker lowered the overall accuracy of the camera-based COACH. These results are very promising, as it shows that using speech in place of a sensor can potentially lead to less erroneous detection and improve COACH's ability to estimate true user performance more accurately.

Overall task performance

The speech-based COACH correctly identified handwashing as complete all 29 times (100%) that it was completed. On average, the participants performed 4.93 out of 5 steps, which was also the average completed steps identified by the speech-based COACH. In contrast, the camera-based COACH correctly identified the task as complete 30.8% of the time (Czarnuch et al., 2013). The participants performed an average of 5.05 out of 6 steps with the camera-based COACH, but it only correctly identified an average of 2.37 steps as being performed (Czarnuch et al., 2013). This was mainly because the speech-based COACH adapted to the sequential nature of handwashing, which was not the case with the camera-based COACH. Czarnuch et al. (2013) concluded that when at least 1 step was missed, the camera-based COACH was not able to adjust its belief state to accommodate. This led to the false reporting of completed steps as being incomplete. However, with the speech-based COACH, this problem was eliminated, since the dialogue was structured sequentially. Participants could not move on to the next step of handwashing until they had completed the previous one. This prevented them from skipping steps and allowed the system to adjust its belief state more easily.

Limitations

There are several limitations with this study that must be acknowledged. Firstly, the sample size for this study was relatively small. Data was collected from only 10 participants before restrictions were put in place due to the ongoing Covid-19 pandemic, which prevented the recruitment of more participants for the study. This makes it difficult to draw any significant conclusions about the wide-scale performance of the speech-based COACH. The 10 participants who were recruited for this study consisted of healthy adults. Even though it was sufficient to use data from healthy participants to evaluate system performance, there are still many aspects that have not been evaluated. For instance, healthy participants generally complied to answering questions correctly and did not have any hearing or speech issues. This might not be the case if the study is conducted with adults who have dementia. Moreover, one of the symptoms displayed by individuals with dementia is communication difficulty and/or breakdown. The speech-based COACH might not be suitable for individuals with this pathology. Therefore, more testing that includes adults with dementia must be conducted before any significant conclusions can be drawn about this system. This study was also conducted in a controlled laboratory environment. Long-term studies in real home settings are necessary to verify the long-term acceptability of the speech-based COACH. Lastly, even though the speech-based COACH can be extended to support other ADL, it might not be suitable for modeling activities that require information from the environment or activities where the nature of the task might vary regularly. Further studies will be required to evaluate the efficacy of modeling various ADL using this approach.

Future work

The results of this paper provide sufficient reason to believe that there is potential in using speech as input for CATs. The next step of this research would be to evaluate the performance and efficacy of the speech-based COACH by conducting clinical trials with older adults with various degrees of cognitive impairment. The efficacy of the system would be evaluated based on (1) whether the speech-based COACH provides the correct prompts to older adults and (2) whether older adults can complete the handwashing task with less dependence on caregivers. It is equally important to investigate the perceptions of older adults with dementia towards a speech-based system for ADL assistance. Thus, further studies should be conducted to evaluate the acceptability and usability of the device through the experiences of target users. It can also be beneficial to adopt a user-centered approach when evaluating the feasibility of the speech-based COACH with

adults with dementia, as this can allow us to consider the needs of the target population and actively involve its members in the design process.

Currently, the speech-based COACH is designed to support the task of handwashing. In the future, the system should be extended to support other ADL. Some ADL that might be suitable for a speech-based system include tasks involving multiples steps that must be completed in a certain order as well as tasks that require privacy, such as dressing and toileting. Another direction for research would be to explore a hybrid solution for ADL assistance that uses a combination of speech and sensors to assist users. For instance, the speech-based COACH can be combined with the camera-based COACH to develop a hybrid solution that uses the camera for non-intrusive monitoring and speech to communicate with the user when needed.

Conclusions

This research paper presents the first step in exploring whether speech-based interactions can overcome the limitations of sensors and improve the usability of CATs for adults with dementia. This study in particular aimed to understand whether speech could be used in place of sensors to provide necessary input data to a prompting system for task step and overall user state estimation. A prototype based on the COACH by Mihailidis et al. (2008) was developed, which used speech instead of the hand tracker to provide action observations to the planning module. The system was tested with 10 healthy adults and results were compared to a previous study with the COACH by Czarnuch et al. (2013). Analysis of the comparison showed that the speech-based COACH identified more completed steps and estimated true participant performance more accurately compared to results reported for the camera-based COACH. However, its performance suffered when the user did not explicitly say what they were doing. In these cases, the hand tracker outperformed speech because it was able to observe users' actions directly. The sequential structure that was adopted by the speech-based COACH was beneficial for its performance with regard to identifying overall task performance. This study was able to answer the following research questions that were posed.

(1) Can the prompting system use spoke language input in place of sensor data to identify whether the correct steps are being completed?
It is possible to use spoken language instead of sensor data to identify whether the correct steps are being completed. The speech based-COACH was able to outperform the camera-based COACH with regard to measuring the number of completed steps and true participant performance. However, more work is needed the improve the system's performance in identifying incomplete steps.

(2) Can the prompting system use spoke language input in place of sensor data to determine the user's task performance?
The prompting system can use spoken language input in place of sensor data to determine the user's overall task performance. Even though more data from the target population is needed to further evaluate this criterion, the results so far are promising. The speech-based COACH was better at inferring the user's overall task performance than the camera-based COACH due to the sequential nature that it adopted.

(3) How will the prompting system's overall system performance be affected when spoken language input replaces sensor data?
The overall system performance of the COACH was not negatively affected when spoken language input replaced sensor data. The system functioned as expected and the results obtained via the use of speech were comparable to those obtained with the camera.

In general, the results of this paper support the feasibility of a speech-based CAT for ADL assistance and highlight its potential as an alternative to sensors. However, this is only the first step of the evaluation and further studies must be conducted to investigate the usability and efficacy of a speech-based CAT for adults with dementia.

References

Boger, J., Poupart, P., Hoey, J., Boutilier, C., Fernie, G., and Mihailidis, A. (2005). A decision-theoretic approach to task assistance for persons with dementia. In IJCAI, pages 1293 1299. Citeseer.

Boise, L., Wild, K., Mattek, N., Ruhl, M., Dodge, H. H., and Kaye, J. (2013). Willingness of older adults to share data and privacy concerns after exposure to unobtrusive in home monitoring. Gerontechnology: international journal on the fundamental aspects of technology to serve the ageing society, 11(3):428.

Buettner, L. L., Yu, F., Burgener, S. C., et al. (2012). Evidence supporting technology-based interventions for people with early-stage Alzheimer's disease. Journal of Gerontological Nursing, 36(10):15-19.

Cakebread, C. (2019). Who Will Win the Smart Speaker War in Canada. Accessed April 2020: https://www.emarketer.com/content/who-will-win-the-smart-speaker-war-in-canada.

Canonico, M. and De Russis, L. (2018). A comparison and critique of natural language understanding tools. Cloud Computing, 2018:120.

Chiao, C.-Y., Wu, H.-S., and Hsiao, C.-Y. (2015). Caregiver burden for informal caregivers of patients

with dementia: A systematic review. International nursing review, 62(3):340-350.

Czarnuch, S., Cohen, S., Parameswaran, V., and Mihailidis, A. (2013). A real-world deployment of the coach prompting system. Journal of Ambient Intelligence and Smart Environments, 5(5):463-478.

Evans-Lacko, S., Bhatt, J., Comas-Herrera, A., D'Amico, F., Farina, N., Gaber, S., et al. (2019). World Alzheimer report 2019: Attitudes to dementia [internet]. London.

Google Cloud (2020a). Contexts: Dialogflow. Accessed April 2020: https://cloud.google.com/dialogflow/docs/contexts-overview.

Google Cloud (2020b). Fulfillment: Dialogflow. Accessed April 2020: https://cloud.google.com/dialogflow/docs/fulfillment-overview.

Google Cloud (2020c). Training Phrases: Dialogflow. Accessed April 2020: https://cloud.google.com/dialogflow/docs/intents-training-phrases.

Hattink, B., Meiland, F., Overmars-Marx, T., de Boer, M., Ebben, P., van Blanken, M., Verhaeghe, S., Stalpers-Croeze, I., Jedlitschka, A., Flick, S., et al. (2016). The electronic, personalizable Rosetta system for dementia care: exploring the user-friendliness, usefulness and impact. Disability and Rehabilitation: Assistive Technology, 11(1):61-71.

Henderson, A. S. and Jorm, A. F. (2002). Definition and epidemiology dementia: a review. Dementia, 3:1-68.

Hoey, J., Poupart, P., von Bertoldi, A., Craig, T., Boutilier, C., and Mihailidis, A. (2010). Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process. Computer Vision and Image Understanding, 114(5):503-519.

Michalakis, K. and Caridakis, G. (2017). IoT interface for healthcare applications. In Proceedings of the 10th International Conference on Pervasive Technologies Related to Assistive Environments, pages 232-233.

Mihailidis, A., Boger, J. N., Craig, T., and Hoey, J. (2008). The coach prompting system to assist older adults with dementia through handwashing: An efficacy study. BMC geriatrics, 8(1):28.

Mihailidis, A., Carmichael, B., Boger, J., and Fernie, G. (2003). An intelligent environment to support aging-in-place, safety, and independence of older adults with dementia. In UbiHealth 2003: The 2nd International Workshop on Ubiquitous Computing for Pervasive Healthcare Applications.

Moore, R. K. (2005). Research challenges in the automation of spoken language interaction. In COST278 Final Workshop and ITRW on Applied Spoken Language Interaction in Distributed Environments.

O'Neill, B. and Gillespie, A. (2008). Simulating naturalistic instruction: the case for a voice mediated interface for assistive technology for cognition. Journal of Assistive Technologies, 2(2):22-31.

O'Neill, B., Moran, K., and Gillespie, A. (2010). Scaffolding rehabilitation behaviour using a voice-mediated assistive technology for cognition. Neuropsychological rehabilitation, 20(4):509-527.

Philipose, M., Fishkin, K. P., Perkowitz, M., Patterson, D., and Hähnel, D. (2003). The probabilistic activity toolkit: Towards enabling activity-aware computer interfaces. submitted to CHI, 3.

Pires, I. M., Garcia, N. M., Pombo, N., and Flórez-Revuelta, F. (2018). Limitations of the use of mobile devices and smart environments for the monitoring of ageing people. In ICT4AWE, pages 269-275.

Rody, B. (2018). Canada's prolific smartphone market skews to iOS: study. Accessed April 2020: https://mediaincanada.com/2018/01/23/canadas-prolific-smartphone-market-skews-to-ios-study.

Rudzicz, F., Wang, R., Begum, M., and Mihailidis, A. (2015). Speech interaction with personal assistive robots supporting aging-at-home for individuals with Alzheimer's disease. In ACM Transactions on Accessible Computing (TACCESS), pages 1-22.

Tao, G., Charm, G., Kabaciński, K., Miller, W. C., and Robillard, J. M. (2020). Evaluation tools for assistive technologies: a scoping review. Archives of Physical Medicine and Rehabilitation.

Vtyurina, A. and Fourney, A. (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, pages 1-7.

Wickens, C. D. and Hollands, J. G. (2000). Signal detection, information theory, and absolute judgment. Engineering psychology and human performance, 2:24-73.

Wilson, R., Rochon, E., Mihailidis, A., and Leonard, C. (2012). Examining success of communication strategies used by formal caregivers assisting individuals with Alzheimer's disease during an activity of daily living. Journal of Speech, Language, and Hearing Research.