ARticulate: One-Shot Interactions with Intelligent Assistants in Unfamiliar Smart Spaces Using Augmented Reality

MEGHAN CLARK, University of California, Berkeley, USA
MARK W. NEWMAN, University of Michigan, USA
PRABAL DUTTA, University of California, Berkeley, USA

Smart space technologies have entered the mainstream home market. Most users currently interact with smart homes that they (or an acquaintance) have set up and know well. However, as these technologies spread to commercial or public environments, users will need to frequently interact with unfamiliar smart spaces where they are unaware of the available capabilities and the system maintainer will not be present to help. Users will need to quickly and independently 1) discover what is and is not possible, and 2) make use of available functionality. Widespread adoption of smart space systems will not be possible until this discoverability issue is solved. We design and evaluate ARticulate, an interface that allows users to have successful smart space interactions with an intelligent assistant while learning transferable information about the overall set of devices in an unfamiliar space. Our method of using Snapchat-like contextual photo messages enhanced by two technologies—augmented reality and autocomplete—allows users to determine available functionality and achieve their goals in one attempt with a smart space they have never seen before, something no existing interface supports. The ability to easily operate unfamiliar smart spaces improves the usability of existing systems and removes a significant obstacle to the vision of ubiquitous computing.

CCS Concepts: • **Human-centered computing → User studies**.

Additional Key Words and Phrases: augmented reality, intelligent assistants, smart homes, smart spaces

ACM Reference Format:
Meghan Clark, Mark W. Newman, and Prabal Dutta. 2022. ARticulate: One-Shot Interactions with Intelligent Assistants in Unfamiliar Smart Spaces Using Augmented Reality. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 7 (March 2022), 24 pages. https://doi.org/10.1145/3517235

1 INTRODUCTION

Intelligent assistants are an increasingly popular way for users to interact with Internet of Things technologies in the home, with over 80 million smart speakers like the Echo and Google Home sold thus far in 2021 [19]. However, despite their growing ubiquity in smart spaces, a problem is emerging on the horizon—discoverability.

To support new users, smart space interfaces must help users easily discover 1) what functionality is available in the space, and 2) how to invoke the desired functionality. Currently, smart space users discover what devices are available and how to interact with them by asking or observing the person that set up the system. Many studies have reported that smart homes are often maintained by a primary technical user, or "pilot user," but must also be used by "passenger users" such as roommates, spouses, or guests [14, 25, 26, 30, 35, 51]. Though these passenger users are often not involved in the installation and configuration of the smart space system, they

Authors' addresses: Meghan Clark, mclarkk@berkeley.edu, University of California, Berkeley, Electrical Engineering and Computer Science, Berkeley, USA; Mark W. Newman, mwnewman@umich.edu, University of Michigan, School of Information, Ann Arbor, USA; Prabal Dutta, prabal@berkeley.edu, University of California, Berkeley, Electrical Engineering and Computer Science, Berkeley, USA.

2474-9567/2022/3-ART7
https://doi.org/10.1145/3517235

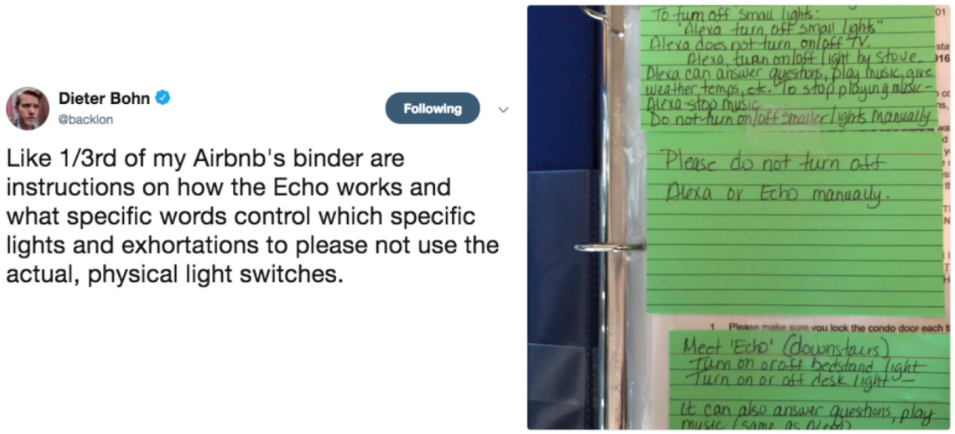


Fig. 1. Ad-hoc discovery method for residential guests. The host provides guests with example voice commands that establish the proper name of each controllable device. The host must also specify what the interface does *not* control (the TV), presumably because it is difficult for guests to determine from failed voice interactions alone that it is not controllable. Note that if users are unable to figure out the interface, they will manually turn off the devices, disrupting any automation routines. Reproduced with permission.

must nevertheless figure out what the system can do and how they can make the system do it. These passenger users currently rely on the pilot user to tell or show them available functionality [25].

As smart space technologies become more widely adopted, there will be an increasing number of passenger users in contexts where the system maintainer is inaccessible, especially in commercial and public spaces. The current reliance on out-of-band social communication for discovering what devices are available and how to interact with them will become a significant obstacle to usability and adoption of smart spaces. In a space like a smart office or smart conference room, the system maintainer is unlikely to be present to teach the large number of passenger users how to use the system. Further, if the capabilities of the smart space change dynamically as mobile smart devices become increasingly common, not even the system maintainer may know what functionality is available at any given moment. These problems have already begun to emerge in residential situations where the smart home maintainer may not be present, such as the AirBnB home rental business (see Figure 1). Helping users quickly discover and invoke functionality in these unfamiliar smart spaces is critical if we wish to enable the widespread adoption of smart spaces, as every user is likely to become a passenger user at some point.

Our goal is to design an intelligent assistant interface that supports smart space discovery—that is, it allows users to walk into an unfamiliar smart space and accomplish a smart space task in a single attempt, or, if the task is not possible, not attempt to do it at all. We refer to this as a *one-shot interaction*. Additionally, such a discovery interface should also help users learn transferable information that could unlock the use of other modalities in the smart home ecosystem, such as voice interfaces, that may lack discovery aids but have other contextual benefits like hands-free operation.

Achieving one-shot interactions with smart space intelligent assistants is challenging because the vast majority of modern smart homes are built around what we call the "proper device name" paradigm. Under the proper device name paradigm, basic smart home interactions focus on individual devices. To invoke device functionality, users must know the proper name assigned to the device by the administrator during installation. This leaky system abstraction requires users to know *in advance* which devices are smart and what their proper names are in order to successfully interact with the smart space.

Computing systems that rely on proper names to reference data objects have long struggled with discoverability. Command lines, databases, and file systems all use names to specify actions or entities. In the 1980s, Furnas studied what he called the vocabulary problem in computing systems, and found that across many domains, the guessability of even the best proper name was very low [13]. He also identified a fundamental tension in proper names—the more guessable a name is, the more generic it is, meaning it could potentially apply to multiple objects. Furnas called this "the precision problem."

The precision problem means that simply providing users with a list of device names is not sufficient for achieving one-shot interactions in unfamiliar spaces. Given a list of names, users might take multiple attempts to guess which name corresponds to the desired device. Worse, if the appliance they want to control is not in fact a smart device, users may need to eliminate *all* the possible names through trial-and-error before determining that the device must not be connected! This means that in addition to providing the list of device names, we also need to map the names to their real-world devices.

Different computing systems have developed different solutions to the vocabulary problem, but most involve some kind of visual aid or pointing-based selection. One such solution is *autocomplete*, which provides real-time visual feedback to users as they type. Autocomplete has been used to help users discover available proper names of tables, fields, and operations in databases [32], and proper names of actions and objects in command lines [28, 31]. These days, popular operating systems such as Windows and MacOS provide a search bar augmented with autocomplete suggestions for finding files by name. *Augmented reality* is another method that has been used to provide textual or visual feedback specifically for physical objects [12, 38, 40, 50].

Upon the introduction of graphical user interfaces, another solution was to capture *gesture* via the cursor, which allowed users to select an object by pointing to it rather than having to specify a proper name. These days users can launch applications by clicking icons rather than typing names on the command line, and select files by clicking through directory hierarchies in a file explorer. These objects are usually also labeled with their proper names, but users only need to recognize the name of a desired object, rather than recall it. More recently, smartphone cameras have been used to capture user gaze as a gesture for indicating real world objects [29].

In this work, we draw on this rich history to design a discovery-centric smart space intelligent assistant interface that uses photo-taking as a selection gesture, and uses augmented reality and autocomplete to provide users with feedback that helps them learn a transferable mental map of smart device names, locations, and functionality while accomplishing their goals.

We unite these concepts into a cohesive whole by framing communication with the intelligent assistant as a familiar messaging interaction. We chose to model ARticulate's workflow after Snapchat, a popular ephemeral photo messaging app with 293 million daily active users worldwide [43]. The app opens into a live camera view that allows users to snap a quick photo of their immediate surroundings, caption it, and send it to a contact.

Inspired by this workflow, we design a messaging app called ARticulate, where the live camera view reveals smart devices and their names using augmented reality. ARticulate users can take a picture of the devices, caption the image with the aid of autocomplete suggestions that are tailored to the devices in the image, and then send the captioned image to the smart space intelligent assistant.

We evaluate the ARticulate design by running user studies with nine participants. While users of the baseline voice interface struggle to operate an unfamiliar smart space, ARticulate users are able to achieve the goal of one-shot interactions in unfamiliar smart spaces, as well as zero attempted interactions with unconnected devices. Additionally, we find that the knowledge learned during use of this messaging interface is able to translate to a voice interface afterwards, unlocking a previously inaccessible part of the smart space ecosystem after only a few interactions. By successfully designing a smart space interface that prioritizes the ability to easily operate unfamiliar smart spaces, we greatly improve the usability of existing systems and remove a significant obstacle to widespread adoption of smart space technologies.

2 BACKGROUND AND RELATED WORK

While there has been an extensive amount of work on designing usable smart home interfaces in situations where users are familiar with the underlying capabilities, less work has been done to study the scenario where people walk into an *unfamiliar* smart space and must discover what capabilities are available and how to use them.

2.1 Smart Space Discoverability

In the paper "Let There be Light," Brumitt and Cadiz compare five different interfaces for controlling lights to determine which modality would be best for intelligent home environments [5]. In particular, the researchers were interested in seeing how people would try to interact with the system upon first entering the room and which interface they preferred. This closely resembles our intended use case of unfamiliar smart spaces.

The voice interface was rated the most-liked, the easiest to use, and was by far the most popular choice for the final task where participants could pick any interface. However, this is likely because the voice interface was Wizard-of-Oz and understood advanced input like pronouns and contextual descriptions. For example, users would say commands like "turn on this light" without verbally specifying which light, or "turn on the left light" with the implied frame of reference facing the television. A voice system that only had access to the text of the user's speech would have struggle to select the correct light. However, the authors observed that *in 91% of the tasks, the participants looked at the device they were referring to.* This means that gaze direction can be used to resolve most ambiguous device references, without the users having to know any proper device names.

The authors cautioned against relying on proper device names, observing that labeling all the lights with proper names would "[require] a person to learn a naming scheme for all the lights in the house, and visitors will be faced with the problem of trying to guess the labels selected by the owner." Unfortunately, modern smart homes have coalesced around the proper device name approach. The study did not evaluate how the usability of voice interfaces would be impacted if the system only understood proper names, as in current systems.

2.2 Potential Discovery Mechanisms for Intelligent Assistants

Many researchers have noted that discovery is a significant obstacle to voice interface usability. The 1990s saw a rising interest in telephone-based automated voice assistants for managing calendars, booking travel, purchasing goods, playing music, and so on [22, 53]. While early work focused on improving the natural language capabilities of the systems, researchers found that users would adapt their speech to what they thought the system's capabilities were anyway (often incorrectly) [23]. These initial findings launched extensive work on helping users more accurately model language interface functionality and limitations, such as through examples and tutorials [21, 54]. Voice-only discovery approaches focused on prompts that expanded or condensed options and provided instructions at appropriate times [53]. However, prompt-based approaches tend to take worst-case time for initial interactions [3, 8], which in smart spaces would impose a potentially prohibitive overhead cost to interacting with each new space. In response, further work has explored conveying discovery information through a high-bandwidth visual channel [47]. These visual approaches to discovery for speech interfaces have included multimodal interfaces, autocomplete, and augmented reality.

2.2.1 Multimodal Interfaces. Some voice interfaces address discovery by co-existing with another interface that shows hints about voice commands for the various functions [53]. One example is VoiceNavigator, a mobile voice user interface (M-VUI) that enables hands-free operation of a mobile phone for those with motor impairments [10]. To help users understand 1) what actions VoiceNavigator could perform on the phone, and 2) how to invoke the actions with speech, users initially used the phone's direct manipulation interface, which showed users the equivalent language-based command for any task they performed that could also be done with speech. VoiceNavigator also exposed contextual help menus that showed users *all* available options, if desired. Though we also use a graphical interface to help users learn linguistic information, we forgo a direct manipulation

approach in favor of maintaining an intelligent assistant interaction framing that could more easily transfer to voice interactions.

2.2.2 Using Autocomplete for Discovery. Autocomplete is a list of possible valid user input text that updates dynamically based on what the user has typed, and which is visually co-located with the input widget. Autocomplete became mainstream when Google introduced it for Search in 2004 [6], though it has been used in many other domains, such as browser URLs, Unix commands, and email [18, 28].

Autocomplete has also been used to aid interactions with smart space intelligent assistants, such as Google Assistant [15], which powers Google Home [16]. Though users can talk to Google Assistant through smart speakers or their phone, users also have the option of typing to it through the phone app, accompanied by autocomplete suggestions. For example, the suggestions for "turn" include "turn on the lights."

However, these suggestions are for improving input speed and effort, and do not support smart space discovery. Google explicitly states that their autocomplete provides predictions of what the user is *already* planning to type, and is not designed to help the user discover new interactions with devices that are actually in the room [17].

Though less common, autocomplete has also been used to support more exploratory workflows, where the technique is sometimes called "autosuggest" [48]. Prior work has used autocomplete to help users interactively explore an unfamiliar database's schema, data, and query language while constructing a single query [32]. This problem is similar to smart space discovery, as to formulate a query against an unfamiliar database, users must learn 1) what is available in terms of tables, fields, and data, and 2) how to invoke queries on the available entities using the query language. Constructing a query for an unfamiliar database normally requires running many additional supporting queries to understand the structure of the database first. Autocomplete reduces the number of queries users must run to achieve their goal by helping users learn about the database *while* crafting the query.

Another example of discovery-oriented autocomplete is code completion in IDEs. When writing code, developers need to use the exact proper name of a class or function. Many IDEs provide autocomplete suggestions that show a comprehensive list of all the valid classes or functions that could occupy the current position in the expression, sometimes with accompanying descriptions from documentation. In addition to aiding in the recall of proper names already known to the developer, these suggestions allow developers to look through the complete list to discover and invoke new functionality [46].

2.2.3 Mobile Augmented Reality for Discovery. Augmented reality, first proposed by Ivan Sutherland in 1968, is a way of visually embedding virtual objects in the physical world [45], including informative readouts about physical objects [12, 38, 40, 50]. WorldGaze is an augmented reality smartphone app that shows labels over real-world objects that the intelligent assistant knows about, which can include businesses and products, but also smart devices. WorldGaze determines which object a user is looking at while speaking to the assistant by detecting the orientation of the user's head in the front-facing camera and combining it with the rear-facing camera [29]. The gaze information allows users to use pronouns or omit names entirely while speaking.

While WorldGaze appears to address a similar problem as ARticulate on the surface, there are a number of differences. Most significantly, WorldGaze does not expose the functionality of the devices, such as whether the lights support dimming or different colors, and what phrases the smart space intelligent assistant (if present) is guaranteed to understand. ARticulate does this using autocomplete. WorldGaze also does not prioritize teaching knowledge that will transfer to other interfaces. The augmented reality annotations did not provide the proper names of the devices, but rather the general class of device, such as "light" or "speaker." In a future where every smart space interface supports name-free interaction this may suffice, but while we live in a world where some interfaces are built around proper device names, there is an advantage to teaching users the proper names.

2.3 Gesture-based Smart Space Interactions

Gestures are physical movements that express meaning without using words, like pointing, gaze, and even pose within the environment. However, many gesture-based systems leave smart space discovery issues unaddressed.

2.3.1 Gestures as an Alternative to Proper Device Names. Some gesture-based systems have explored how to indicate smart devices without using words at all. For example, WorldCursor is a wand that lets users point at lights and turn them on or off with a button press [49]. However, many more complicated gesture-based systems essentially create non-verbal "proper names" for actions and side-step the question of device selection [27, 33]. This body of work runs into a vocabulary problem similar to that of speech interfaces. Free-hand elicitation studies find high levels of disagreement between people's gestures, meaning that gestures are not generally guessable [52]. For any given system there must therefore be some way for users to discover what specific gestures it understands and what they mean, just like with voice systems.

Gesture-based systems also do not help users determine which devices are smart before trying to interact, nor do they provide hints as to the available functionality of each appliance. For example, Ubiwheel allowed a simple controller to operate on many devices and functions by changing its behavior based on where it was placed [41], but users must still know in advance which appliances are smart and which regions control which functions.

2.3.2 Camera-based Gestures in Smart Spaces. Using smartphone cameras to select real-world objects is a specific form of gesture that in some sense captures user gaze, which Brumitt and Cadiz showed was an intuitive way to indicate smart space devices. Several recent works have examined this interaction, including WorldGaze (discussed above) and Snap-to-it [11]. Like ARticulate, Snap-to-it also uses photo-taking as a way to refer to smart devices. Snap-to-it allows users to access a graphical user interface for a smart device by snapping a picture of it. The system uses computer vision to identify which of the known smart devices is in the photograph, and then provides the user with an interface for that device. While the core interaction is appealing, the design of Snap-to-it does not support discovery in an unfamiliar smart space at all. Users have to know (or at least suspect) that a device is interactive in order to take a picture of it, but the system does not provide any hints. To aid discovery during the study, the researchers posted fliers stating that there was a smart device in the area. Snap-to-it also does not teach users information, such as proper names, that would enable the use of other interfaces. Finally, Snap-to-it lacks an intelligent assistant interaction framing. Our approach addresses all of these shortcomings, helping users autonomously discover and use capabilities in an unfamiliar smart space.

3 DESIGN GOALS

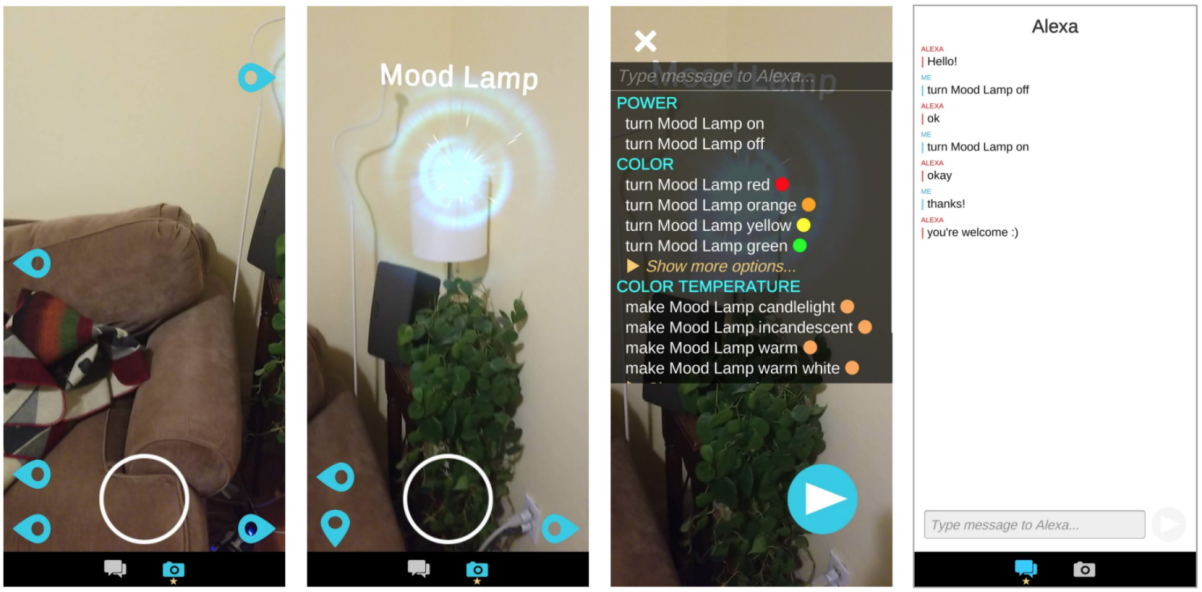
An interface design that helps passenger users independently discover and use the capabilities of an unfamiliar smart space should achieve the following four design goals:

Goal 1: One-shot interactions with connected appliances in unfamiliar smart spaces. Users should be able to enter a smart space, discover available functionality, and invoke the desired functionality in the first attempt.

Goal 2: No attempted interactions with unconnected appliances. As a part of discovering what functionality is available, users should also learn what functionality *is not* available. Users should be able to determine when a device is not a smart device without making any failed interaction attempts.

Goal 3: An assistant-oriented interaction framing. While we want to enable one-shot interactions in unfamiliar smart spaces, we want to do it specifically while preserving an assistant-oriented interaction metaphor. Centering the design on intelligent assistants means that the user should feel as though they are interacting primarily with the assistant, not the devices.

Goal 4: Learning transferable information that enables the use of other smart space interfaces. If the user could build a map of the overall capabilities of the space beyond those involved in their specific tasks, the knowledge would unlock the ability to use other interfaces for future tasks, even if those interfaces lack discovery aids.



(a) (b) (c) (d)

Fig. 2. ARticulate interface. ARticulate is a smartphone app for communicating with smart space assistants inspired by the Snapchat messaging application, with key modifications to support smart home discovery. Off-screen markers that slide around the edge of the screen indicate the presence of nearby smart devices (a). Glowing animated orbs with labels indicate the location and proper names of on-screen smart devices (b). Users can take a picture and caption it with a message for the assistant. Autocomplete suggestions (c) aid in caption composition, providing insight into what the agent understands and what capabilities the device has. Suggestions are included for all devices in the photo. A chat history window (d) allows users to see a record of what captions were sent and the assistant's responses, and also provides users with a way to type directly to the assistant without the necessity of photos.

4 AUGMENTED REALITY MESSAGING FOR SMART SPACE ASSISTANTS

We designed an app called ARticulate, a Snapchat-like messaging app for communicating with smart space assistants that introduces augmented reality and autocomplete to support discovery of smart space devices and their capabilities. Figure 2 provides an overview of the main user interaction. Each of the major design points is highlighted below.

4.1 Situational Photo Messaging

To draw on users' prior experiences with messaging and ensure that the design had street-tested usability, we wanted to base our solution on technology that people already use to communicate with each other. We chose to model ARticulate's workflow after Snapchat, a popular messaging app with 293 million daily active users worldwide [43]. In the United States, 65% of adults aged 18-29 use Snapchat [7].

Snapchat's main usage centers around photo messages, which are pictures of the users' immediate environment with an optional caption, sent to specific contacts. Snapchat emphasizes a style of messaging where photos are valued for their situational relevance, making Snapchat an archetype of photo-centric messaging that puts a strong emphasis on the immediate physical context of the sender [2, 24, 34, 39].

ARticulate users take photo messages of smart devices, caption them, and send them to an intelligent assistant. The idea is that there is one assistant for the room, and the assistant that the user is messaging is the same one that they might, for example, talk to through a smart speaker. Our goal is not to replace voice, but rather to complement it, and even enable its use by supporting discovery.

In ARticulate, taking a photo acts as the means of selecting the desired device *without needing to know the name in advance*. While technically the assistant does not need the photo itself to correctly perform the tasks in our study, the interaction metaphor of "sending" the assistant a picture of the physical context along with the message is useful for framing the communication about the user's immediate context. This lays the groundwork for future interactions where the user's intent is expressed in a combination of information in both the photo and the caption. For example, while ARticulate's autocomplete suggestions for captions are unambiguous standalone phrases that will also work with smart speakers, our implementation also allows users to omit device names from captions and rely solely on the photo context for scope. Concretely, users can snap a photo, caption it with "red," and any device in the photo that supports color settings will be set to red. We did not evaluate this interaction mode in our study and our users did not discover it, but it illustrates the potential role of photos in communication. Sending intelligent assistants photos with captions could potentially become a staple mode of communicating with assistants about context-specific topics.

4.2 Augmented Reality

ARticulate allows the user to use the live camera view to look around the smart room and discover the location and names of smart devices (as well as which devices are not smart by their lack of a label). Off-screen indicators point in the direction of the various smart devices in the room, and when the device is in view of the camera, the fact that it is a connected device is indicated by a glowing orb and its name placed over its real-world location.

4.3 Autocomplete

Autocomplete suggestions help users caption a photo with a message that the assistant will understand. The suggestions span the set of available interactions for the devices in the photo. When the picture is sent, the user is taken to the chat history window, which shows the sent messages as well as the assistant's responses.

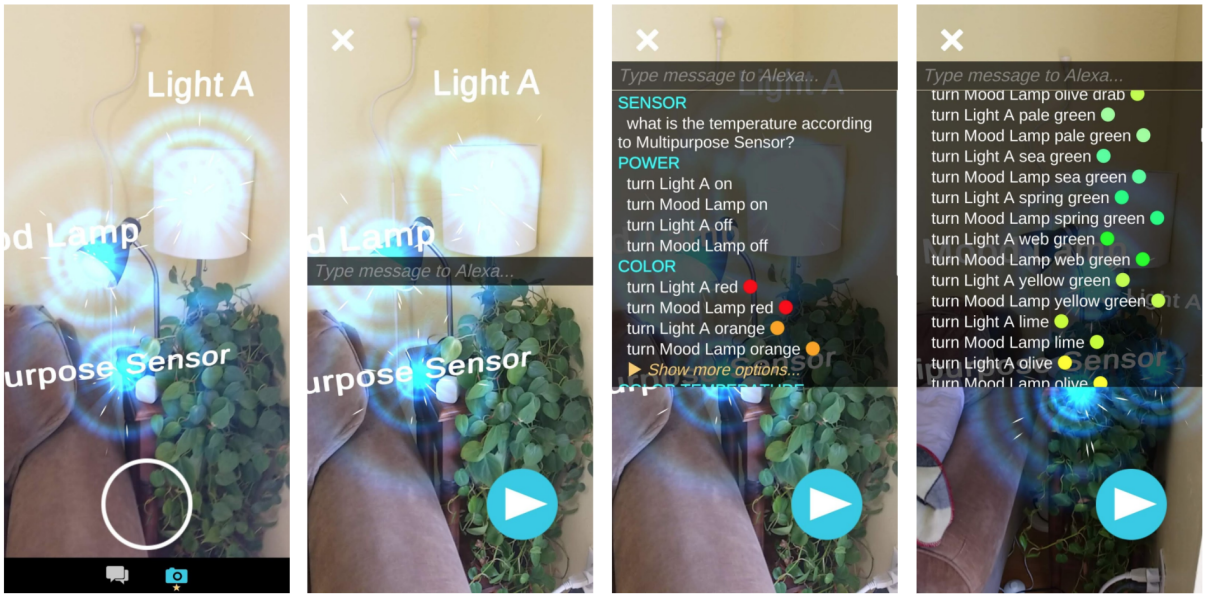
5 IMPLEMENTING THE ARTICULATE AUGMENTED REALITY MESSAGING APP

To evaluate whether this design works for our stated goals, we built an Android app that implements all the key features described above. We also built a smart home testbed for the ARticulate app to interact with, and a chatbot for users to message. We discuss the details of our implementation in the next sections.

5.1 "Alexa" Chatbot

In our implementation, we chose to use Alexa as the underlying smart space assistant, as Alexa-enabled devices such as the Amazon Echo and Echo Dot are among the most popular smart speakers at this time. However, Alexa does not provide an API that allows developers to send text-based utterances. This means that we had to take extra steps to preserve the illusion that the Alexa that the user communicated with over chat was the same as the Alexa that the user spoke to using the smart speaker.

We wrote a custom chatbot that closely imitated Alexa's responses and behavior. Manipulation of smart devices was done through cloud access to the smart home platform we set up. However, this chatbot was not a perfect clone. Alexa can understand and respond to non-smart home interactions, such as answering questions by looking up information online, which we did not attempt to duplicate. We also did not duplicate functionality like setting timers, playing music, or responding to queries about the weather. These choices were intentional as we were focused on use of the devices in our smart home testbed.



(a) (b) (c) (d)

Fig. 3. Autocomplete suggestions for multiple devices. ARticulate users can view (a) and photograph (b) multiple smart devices at once. In this example there are three devices: Mood Lamp, Light A, and Multipurpose Sensor. The autocomplete design has several features to help structure the large number of resulting suggestions (c). Suggestions are categorized into "capabilities" like *power* and *color*. When multiple devices are in a photo, the autocomplete suggestions are interleaved by capability. Each capability section hides excess suggestions with a "Show more options" drop-down. This was especially critical for the color suggestions, as Alexa understands over 145 different color names—too many to casually scroll over. An illustration of the variety of color suggestions that are revealed after tapping "Show more options" is shown in (d). These design choices are intended to help users quickly understand the range of what the available devices can do with minimal scrolling and reading.

5.2 Autocomplete

Even though the assistant understood many paraphrases of each command, autocomplete suggestions were only provided for "canonical" phrasings like "Turn Mood Lamp on," and "Turn Floor Lamp red." This was due to the overhead of providing autocomplete suggestions for every possible paraphrase that the system understands.

As shown in Figure 3, the autocomplete suggestions showed entire suggestions as full sentences rather than building up word-by-word. Suggestions would be shown if the input the user typed (e.g. "blue") appeared anywhere in the suggestion. Users could tap on suggestions to place them into the user input box.

To help users organize the large number of suggestions into broad categories of available actions, suggestions were divided into sections based on "capability" like *power*, *color*, *color temperature*, *brightness*, *lock*, and *sensor*. These capabilities were derived from the device schemas provided by the platform we used to control the devices, and thus were programmatically generated from the system's API. When multiple devices were in one photo, the autocomplete suggestions for the different devices were interleaved by capability.

To allow users to quickly scroll over the available types of capabilities and build a broad mental map of available actions, each capability section hid excess suggestions with a "Show more options" button. This was especially

critical for the color suggestions, for example, as Alexa understands over 145 different color names—too many to casually scroll over.

5.3 Augmented Reality

The ARticulate app was implemented for Android in Unity using Google's ARCore library to provide the underlying augmented reality functionality. ARCore is an entirely vision-based method for augmented reality. To display the locations and names of smart devices, the app must be able to access a scan of the room with the device locations specified. We envision that in real-world settings, a scan could potentially be made once during the initial system setup as a part of configuring the device, and shared with other users who later enter the space. For ARticulate, we provided a scanning feature that allowed users to create a scan of a smart space and place device markers, stored locally on the phone. This feature was used exclusively by the study team during setup.

6 EVALUATING ARTICULATE IN UNFAMILIAR SPACES

To evaluate whether ARticulate achieves the four design goals we laid out in Section 3, we devised an IRB-approved study protocol where we sent a smart home testbed directly to users' homes and asked them to perform tasks using both voice and ARticulate. This allowed us to observe participants performing tasks in an unfamiliar smart space with a mix of connected and unconnected devices. As a baseline, we examined how hard it was to accomplish the tasks using state-of-the-art voice assistant interfaces. We then examined whether the ARticulate design improved the ability of users to accomplish the tasks and reduced the number of failed interaction attempts. We also examined whether use of ARticulate enabled successful use of the voice interface afterwards, including for tasks that the user had not done before.

6.1 COVID-19 Pandemic Challenges

Designing the protocol for evaluating ARticulate was unusually challenging due to the COVID-19 global pandemic. During this time, most citizens in the authors' country were under some form of shelter-in-place order that prohibited them from leaving their homes except for essential reasons, and research facilities and university campuses were closed. This placed major constraints on the study protocol design.

Since indoor gatherings of strangers were not permitted under the COVID-19 lockdown, we made a portable testbed and sent the testbed to each participant. However, the participants needed an on-site facilitator who could set up the testbed for them, so that they could approach the system with no knowledge. To ensure that there was an on-site person already in the participant's "germ bubble" who could set up the system, we added our entire research group to the official IRB study team and had every member undergo human subjects research training. The recruitment pool was therefore limited to people who lived or were otherwise already within contagious contact of our lab members. This is a form of snowball sampling, which is subject to biases that we will discuss during the overview of our participants in Section 7. This approach also imposed additional requirements onto the design of the testbed system to be easy to install, configure, and operate remotely.

6.2 Recruitment

The principal audience for this test were people who might need to operate an unfamiliar smart space. Examples include people likely to be guests in a smart home, inhabitants of a smart home managed by someone else, lodgers in a smart hotel room, workers in a smart office space, and visitors in a smart conference room. Since this includes many adults, we defined our recruiting criteria as people over 18, with a home internet connection with WiFi, and fluency in the English language. As mentioned above, due to COVID-19 restrictions, recruitment used a form of snowball sampling that drew from residents and close contacts of members of the study team.

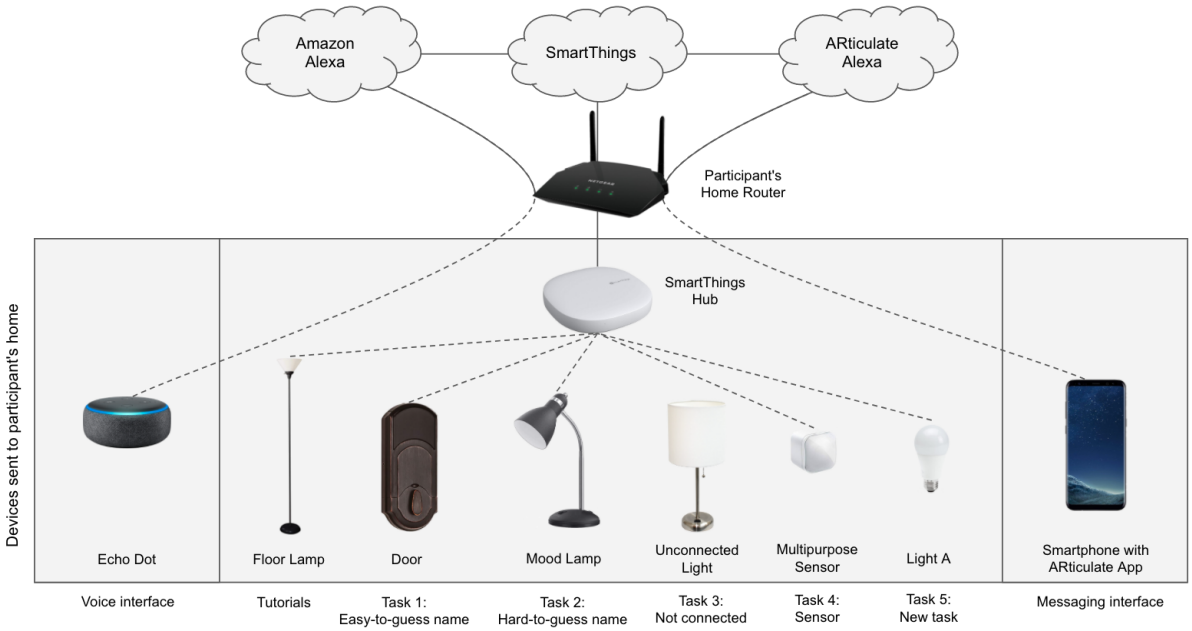


Fig. 4. ARticulate testbed. To evaluate ARticulate during the COVID-19 pandemic, we created a testbed kit that we could send to each participant's home. Each participant lived with a study team member who could set up the testbed and facilitate the in-person needs of the study, such as running the video call and using a laser pointer during task instructions to indicate relevant devices to the participant without using language. The smart home testbed itself consisted of six devices, each of which corresponded to a different task that tested some aspect of the research question. The "smart" appliances connected wirelessly to the commercial SmartThings platform [42] through a hub. Each participant used two different interfaces to communicate with the assistant at different phases of the study: a voice interface via the Amazon Echo Dot [1], and the ARticulate messaging app preloaded on a smartphone. The cloud servers powering these interfaces interpreted the participant's utterances and actuated or queried the smart devices through the SmartThings API.

6.3 Smart Home Testbed

To evaluate ARticulate's performance in an unfamiliar smart space, we sent an entirely self-contained set of testbed devices to participants' homes, ensuring that while the space itself was familiar to the users, the smart system setup was novel. Familiarity with the normal room and novelty of the testbed devices did pose an issue for evaluating discovery, however. It would be reasonable for users to assume that every new fixture or device added to the room for the experiment has smart capabilities, while every appliance that is normally present is not part of the smart room testbed. If these assumptions were true, it would give them an unrealistic advantage in evaluating whether or not a device is smart.

To counteract this, one of the fixtures we added was a lamp that was not connected to the smart space at all. Also, one of the lights we included, Light A, was installed in an existing fixture that is normally in the room and is normally not smart. Knowing whether or not a device was normally in the room therefore provided users with no additional information about whether or not it was smart.

The complete configuration of the testbed is shown in Figure 4. We chose this set of smart devices to correspond with particular tasks, each of which spoke to a particular research question. Devices in our smart space testbed consisted of three smart LED bulbs that all supported power, color, brightness, and white temperature, one smart

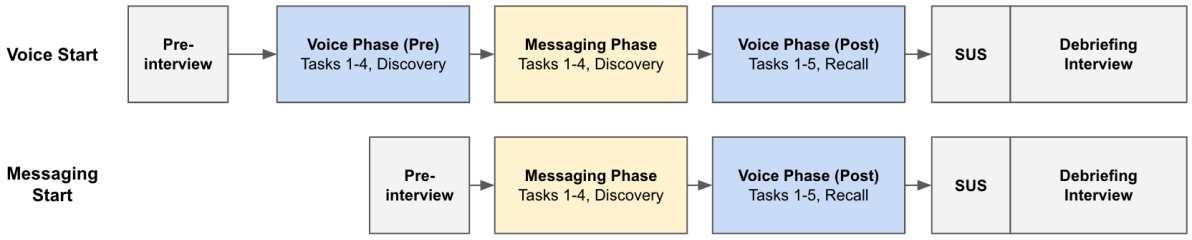


Fig. 5. User study protocol. Participants were assigned one of two possible session types, Voice Start or Messaging Start. The session type determined whether the user started the interactive tasks using voice or ARticulate. "Voice Start" participants experienced all three interaction phases starting with a voice phase. The "Messaging Start" treatment skips the first voice phase entirely to show that performance during the messaging phase is not due to learning effects. In both the first voice phase and the messaging phase, the tasks are the same (Tasks 1-4), and are followed by a three-minute self-directed discovery period. However, note that the final voice phase contains an extra task (Task 5: New Task) that is not posed anywhere else. The final voice phase also ends with users recalling what they have learned about the devices in the room, rather than discovery. All sessions begin with a pre-interview about user experiences with prior technologies, and end with a System Usability Scale (SUS) questionnaire and debriefing interview.

door lock that could be remotely locked and unlocked, and an environmental sensor that measured temperature, motion, and light (though Alexa only exposes temperature).

6.4 Test Preparation

In advance of each user evaluation, we used a no-contact method to deliver the evaluation kit containing the smart home testbed to the participant's home. Over a video call, we worked with the on-site facilitator to set up the smart home system and perform the augmented reality scan and labeling of the room. With the potential participant, we held an informed consent discussion where we reviewed the goals and overall structure of the session, and if they chose to sign the consent form, we continued with the rest of the session.

6.5 Session Structure

The overall session took anywhere from an hour to two hours, depending on the session type they were assigned. Figure 5 illustrates the main components of the study protocol. First, all sessions began with a short pre-interview where we asked the participants about their prior experiences with related technologies that might inform how they approached the systems in the study. Then we began the interactive tasks, which were broken up into two or three phases depending on the type of session. After the interactions, the participant filled out a System Usability Scale (SUS) questionnaire [4] about their experiences, which we used to guide a final debriefing interview where we asked the participant detailed questions about their experience and thoughts.

6.5.1 Pre-interview. All sessions began with a short interview about the user's prior experience with related technologies, to understand any biases that might influence users' interactions. The main topics covered were intelligent assistants (including smart speaker usage like Echo and Google Home, but also usage outside of a smart home context, such as on mobile devices), smart rooms (including those controlled by tablets or phone apps), and the Snapchat messaging application from which ARticulate drew inspiration. This interview usually took only a few minutes.

6.5.2 Interactive Phase Structure. There were two different session types that participants could be assigned to that determined what the interactive portion of the session looked like. The interactive portion was broken up

into two or three phrases depending on session type, with each phase specifying which interface the participant used. Figure 5 outlines the general structure of the two different session types. The two session types were as follows:

- **Voice start**. In this treatment, the participant went through three phases, first a voice-only phase where the participants were asked to perform several specific tasks and to try to discover capabilities of the room using only a voice interface to Alexa. This phase established a baseline of how difficult the tasks were to complete using state-of-the-art voice interfaces to smart home assistants. In the second phase, the participants were introduced to the ARticulate messaging app and performed the same directed tasks and free-form discovery, but with ARticulate instead of voice. In the third phase, the participants were returned to the voice interface to see if their ability to use the voice interface to perform tasks had improved after using ARticulate. There was also one participant assigned a **Voice start, no sensor** variation. This was the same as the normal voice start, except that during the first voice phase the sensor task was omitted. This was to see whether the participant would notice the sensor and attempt to interact with it using voice during free-form discovery.
- **Messaging start**. In this session type, participants were introduced immediately to the ARticulate messaging app. This was to rule out any learning effects from earlier phases on the performance of the app. In the final phase, users interacted with the Alexa voice interface to see if they were able to successfully use the voice interface more effectively than those who started with the voice interface.

6.5.3 Interactive Tasks. In each phase, we asked users to perform the same four or five tasks using the interface for that phase. In these tasks, the users were asked to interact with certain devices. However, due to the linguistic nature of the tasks, we wanted to avoid priming users with any language about the device itself. The on-site facilitator would state the task instructions, such as "get Alexa to turn on this device for you. Please narrate your thoughts out loud as you do so," while using a laser pointer (included in the test kit) to indicate which device they were referring to. The interactive tasks were as follows:

- Task 1: Easy-to-guess name. For this task, the instructions were to "get Alexa to unlock this device for you." This could be done with the command "unlock the door." Unlocking the door also required a numeric code, which we provided users with in advance and do not count against interactions, since we are merely interested in whether the user can guess the easy name. This task provided a baseline, as we expected users to be able to operate this device correctly in one guess using either voice or the messaging app.
- Task 2: Hard-to-guess name. For this task, users needed to turn on the Mood Lamp. Mood Lamp is a device with a hard-to-guess name, as most people might call this a desk lamp or small lamp or table lamp. We would expect this task to be difficult using the voice interface compared to using ARticulate.
- Task 3: Not connected. In this task, the user was asked to turn on an unconnected lamp. Since this light is just a normal light bulb, the task is not possible. This represents situations where a user is in a smart space where only some devices are smart. The user may want to perform a task that is in fact not possible because the device they want to operate is not connected. Throughout our study, users were informed that some of the tasks we will ask them to do may be impossible, and that they must figure out whether that is the case. Before each phase we reminded users that they could either complete the task or decide it was impossible and move on. We expected this to be extremely difficult to determine using the voice interface, because there is no way for users to distinguish between cases where the device is not connected and cases where the user simply has not guessed the right device name. We expected this task to be easy with ARticulate, since the unconnected light lacked a glowing AR marker.
- Task 4: Sensor. In this task, the facilitator stated that "there is a device in this room that measures something. Get Alexa to tell you what the current value is." This could be accomplished by asking Alexa "What is the temperature according to the Multipurpose Sensor?" We expected this task to be quite difficult with voice,

Treatment	Participant ID	Gender	Age	Assistants (Voice)	Assistants (Text)	Smart Speaker Use	Smart Home Use	Unfamiliar Smart Spaces	Snapchat
Voice start	1	F	41	Yes	No	Yes	Yes	Yes, hotels	No
	2	F	31	Infrequently	No	Infrequently	Yes	No	Yes
	3	M	24	Yes	No	Yes	Yes	No	Yes
	4	F	26	Yes	No	Observer	Observer	No	No
Voice start, no sensor	5	M	21	Yes	No	Yes	Yes	No	Yes
Messaging start	6	F	24	Yes	No	Yes	Yes	No	No
	7	F	28	Yes	No	Yes	Yes	No	Yes
	8	F	28	Yes	No	Observer	Observer	No	Yes
	9	F	28	Infrequently	No	No	Observer	No	Yes

Fig. 6. Participant demographics and background. The columns show participant demographics as well as whether the participant has used various technologies. The "Assistants (Voice)" column includes both smartphone and smart speaker use, so it is a superset of the "Smart Speaker Use" column. "Observer" is used to indicate when someone does not use a technology themselves, but frequently observes another use it, such as a roommate, partner, or close friend. Strikingly, though speaking out loud to assistants was very common, no one remembered typing messages to an assistant, though both of the main smartphone assistants (Google Assistant and Siri) support it. Also, while all of the participants frequently visited or lived in a home with smart space technologies, almost no participants had been in an unfamiliar smart space set up by someone else that they needed to figure out how to use.

as small wireless sensors are unobtrusive by design, and their affordances are not visually obvious, making the need for discovery aids particularly acute.

- Task 5: New task. This final task was only performed in the final voice phase, where participants had to use voice without any discovery aid except what they remembered learning from ARticulate. In this task we asked them to turn on Light A, which is a device that we had never previously asked the participant to interact with in other tasks. We used this task to evaluate whether after using ARticulate, a user could use voice to interact with a device that they had not interacted with before.

In addition to the tasks, at the end of the first voice phase and the messaging phase, users were given three minutes for self-directed discovery where their goal was to learn as much as possible about the room using the interface for that phase. At the end of the final voice phase, there was a recall section where we asked users to share as many device names and capabilities that they remember learning about the room.

6.5.4 SUS Questionnaire and Debriefing Interview. The System Usability Scale (SUS) is a commonly used usability instrument that asks users to answer 10 standard questions on the Likert scale related to usability issues. The total SUS score is out of 100. We instructed users that the SUS questions only applied to the ARticulate interface, not to the smart home itself. We primarily used the SUS as a prompt to guide user thinking through the post-interaction debriefing interview. In addition to going over the user's SUS questionnaire answers and discussing why they answered the way they did, we also asked users questions like "Under what circumstances would you use this? Why?" and "If you had three wishes to make this better for you, what would they be?" and "Which features of the app did you find the most helpful while accomplishing the tasks?"

7 PARTICIPANT OVERVIEW

We recruited nine participants. Four participants went through all three phases (the "Voice Start" treatment), one went through all three phases but was not given the sensor task in the first voice phase ("Voice Start, No Sensor"), and four skipped the first voice phase and started immediately with ARticulate ("Messaging Start"). The demographics and background experience with related technologies of the participants are shown in Figure 6.

			Participant ID								
Interface	**Task name**	**Ideal**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Voice	T1: Easy-to-guess name (Door)	**1**	1	1	1	1	1				
	T2: Hard-to-guess name (Mood Lamp)	**1**	1	4	4	6†	3*				
	T3: Not connected	**0**	28	8	4	2	3				
	T4: Sensor (Multipurpose Sensor)	**1**	9*	1†	4*	4*					
Messaging App	T1: Easy-to-guess name (Door)	**1**	1	1	1	1	1	1	1	1	1
	T2: Hard-to-guess name (Mood Lamp)	**1**	1	1	1	1	1	1	1	2*	1
	T3: Not connected	**0**	0	0	0	1	0	0	1	0	0
	T4: Sensor (Multipurpose Sensor)	**1**	1	1	1	1	1	1	2	1	1
Voice (Post-App)	T1: Easy-to-guess name (Door)	**1**	1	1	1	1	1	1	1	1	1
	T2: Hard-to-guess name (Mood Lamp)	**1**	1	1	1	1	1	1	1	1	1
	T3: Not connected	**0**	1	0	0	0	0	0	0	3	1
	T5: New Task (Light A)	**1**	1‡	1‡	1	1‡	1	1	1	1‡	1*
	T4: Sensor (Multipurpose Sensor)	**1**	2†	1†	3*	4*	3†	1*	3	1	2*

* Did not accomplish a task that was possible
† Accomplished task, but without using name
‡ Accomplished task, but with incorrect name

Fig. 7. Number of interactions per task for each participant. An interaction is a message or utterance directed to Alexa.

The majority of study participants were women. This is the result of our snowball recruiting method that enrolled co-habitants, particularly partners, from our mostly male research group. Past studies have shown that while most current smart home "maintainers" responsible for installing, setting up, and debugging the system are men, many of the passenger users of the resulting system are women [14, 25, 26, 30, 35, 51]. These women must discover and use the available functionality without knowing the system configuration. This means at least in the short term, women would be the ones most likely to use and benefit from a discovery-oriented interface like ARticulate, so our participants resemble a likely user population.

Though most participants did not have technical backgrounds, all participants lived in or had been in smart homes before, and the majority had operated a smart home using a voice assistant or closely observed others who did. This is likely a result of the snowball recruitment method imposed by the COVID-19 restrictions and may not be reflective of the general population. However, since we are interested in how our system would work in a future where smart spaces are more ubiquitous, it was useful that our participants were already familiar with the domain and did not need to learn about smart homes while also trying to help us evaluate ARticulate.

8 FINDINGS

We analyzed whether or not users were able to accomplish tasks and the number of interaction attempts each user made for each task. The results are shown in Figure 7. The interaction counts reveal that, as expected, users struggled operating the unfamiliar smart space using voice. However, when using ARticulate, in the majority of cases users accomplished their goals in a single interaction, and made zero attempts to interact with the unconnected device. We also found that some of the knowledge learned through ARticulate was able to transfer to voice, allowing users the option of using the hands-free modality after only a few tasks with ARticulate.

8.1 Voice Findings

Voice users required many attempts to accomplish goals in an unfamiliar space. Figure 7 shows that, as expected, participants were able to perform the easy-to-guess Task 1 (Door) in one attempt, but for the remaining tasks they were not. For Task 2 with the hard-to-guess device name (Mood Lamp), three of the five participants were ultimately able to achieve the goal. At first it may seem surprising that they could guess the name Mood Lamp at all. However, this is because Alexa helps the users with prompts, such as "Did you mean Mood Lamp?" (the right answer) in response to "Turn on the desk lamp" (a wrong guess). Participant 4 technically achieved the Task 2 goal of turning on Mood Lamp by asking Alexa to turn on all the lights.

8.1.1 The "Not Connected" Scenario Was Particularly Challenging for Voice. For Task 3, we asked participants to turn on a normal lightbulb not connected to the smart system. This represents a typical situation where not every device in the space is smart, and the user's goal cannot be achieved through the system because the device is a normal manual appliance. We mentioned repeatedly in the instructions that some tasks would be impossible and users could decide to move on at any time. However, with the voice interface, there was no way for participants to easily determine whether the task was impossible or whether they had just not guessed the right name yet. Consequently, users made many attempts to turn on a light that could not be turned on, with the highest count belonging to Participant 1, who made 28 attempts before giving up. Three out of the five participants gave up after asking Alexa to "turn on all the lights" and discovering that the unconnected lamp did not turn on.

8.1.2 Almost No Voice-start Participant Accomplished the Sensing Task. While actuators (smart or not) are often designed to advertise their functionality, the intentional unobtrusiveness of sensors posed unique challenges to voice users. Participant 1 incredulously summarized the requirements of Task 4: "So, I would have to identify what this device is, then figure out its name, and then use that name for the measurement reading." Participant 4 said, "I'm personally at a loss, because there are so many different ways to measure something, and not knowing what she's measuring makes it hard for me to ask, and she also can't tell me what she's measuring."

In real-world situations, users may not even realize there is a sensor to try to read. Participant 2 said, "Until the app showed the glowing dot [in Phase 2], I had no idea that [Multipurpose Sensor] was even there." Participant 4 said, "I think in real life if I were in an AirBnB, I would not have asked again. I would have been like, 'oh, I guess there's no sensor.'" We omitted the sensor task entirely in Participant 5's voice phase in order to see whether he would attempt to interact with it or ask Alexa about sensors during self-directed discovery with the voice interface. He did not. During the debriefing, he reported noticing the small white cube when he entered the room, but it did not occur to him to try to interact. The unobtrusiveness of the sensor even concerned some participants. After hearing the sensor task instructions, Participant 4 shared, "My first thought is that I'm freaked out, because I don't like being measured."

The task was further complicated by limitations in Alexa's natural language understanding for sensor interactions. The required phrases were complex, and some variations would work, while others would not. For example, "What is the temperature according to Multipurpose Sensor?" worked, but "Tell me the temperature according to Multipurpose Sensor" did not work, much to the surprise and frustration of Participant 4, who thought she had remembered it exactly during the final post-app voice phase. Alexa also understood some interactions that omit the sensor name, but not others. "What is the temperature inside?" worked, but "What is the temperature in the room?" did not, though multiple participants tried it. This issue with the complexity and specificity of the phrases, on top of the unobtrusive and opaque nature of sensors, made the sensor task difficult for voice users.

8.1.3 Participants Employed Several Strategies to Deal with Failed Guesses. A remarkably common strategy when users struggled throughout the first voice phase was to ask Alexa for a list of connected devices. Four out of five participants requested a list of devices, usually multiple times. Participant 4 was particularly distressed by the end of the voice phase: "I wish I could somehow ask to understand what all the devices are that she has

because I think that would help me [...] I'm frustrated because I want this list!" During self-directed discovery in the voice phase, Participant 3 made eight separate attempts to ask Alexa for a list or count of the connected devices. Surprisingly, considering it is such a common strategy for users, Alexa does not support providing a verbal list of connected devices. Nevertheless, just providing a list of connected device names without grounding them to their locations would not be a sufficient solution—many proper names could plausibly refer to multiple different devices, and more importantly, a list of smart device names does not help users quickly discover when an appliance they want to control is *not* smart and connected.

Another strategy relied on Alexa's "Did you mean *<device name>*?" hints. Participant 1 intentionally used this feature during Task 3, explaining, "It's not 'the reading lamp,' but I'm going to try 'the reading lamp' just in case she gives me the name of this device." However, sometimes Alexa would suggest the wrong light. Remarkably, users would occasionally reject Alexa's correct suggestions thinking they were incorrect.

Another common strategy after several failed guesses was to bypass the voice interface completely and physically reach for the device for closer inspection. Four out of five participants touched or asked to touch the physical fixture, either to look for clues or because they thought the device might be broken. Participant 4 said, "The other thing that I'm thinking about is that maybe there are physical signs that something is a smart thing. So I'm going to look at this floor lamp [the tutorial device] now. [...] Can I take this lightbulb out?" During Task 2, Participant 5 toggled the power switch on the Mood Lamp to make sure it was properly connected after failing to guess the name multiple times. Participant 5 was not the only user who thought a device was broken. During Task 3, Participant 1 said: "So there's something wrong with your lamp, you guys, it's like broken or something, or it has a very special name that's not lamp, and it's not light." In real-world conditions, it is likely that users would simply abandon the smart space platform rather than make so many attempts. As Participant 4 said, "I think what I would do if I needed to turn on this lamp would be to turn it on myself."

8.2 ARticulate Findings

ARticulate enabled one-shot interactions and identifying unconnected devices in unfamiliar spaces. Participants accomplished the possible tasks in a single attempt, with only one exception that we discuss below. Using ARticulate, participants also recognized right away that the unconnected device was not connected, and largely did not attempt to interact with it at all. We can see that the majority of users had an ideal zero interactions for Task 3 using the messaging app. The scores from the System Usability Scale questionnaire were very good overall, with a mean score of 83 and a median of 85.

8.2.1 Participants Would Use ARticulate in Unfamiliar Spaces. The lowest-scoring SUS question on average was Question 1, "I think that I would like to use this system frequently." However, when asked when they would use something like ARticulate, participants identified situations when they were in an unfamiliar smart space. As Participant 4 said, "I feel like once I knew the name of everything I would probably prefer to use voice personally, but [...] if I had this system, it would be the absolute first thing I did in any room that was a smart room. [...] every time I had a new room I would definitely use it."

8.2.2 Participants Were Not Sure When They Had Seen All Devices. Participants expressed appreciation for both the augmented reality markers and the autocomplete suggestions and found them useful for accomplishing the tasks. However, several participants stated that they still wanted a list of names or at least a count of devices. We observed many times during self-directed discovery that users would slowly pan around, moving back and forth to count the teardrop markers and map them to devices. Participant 4 expressed, "Something I want to do now is make sure I really have seen all of the devices in the room. I wish there was a number or something that could tell me. So the best I can do is look at these teardrops."

8.2.3 All Participants Recognized That the Unconnected Light Was Not Operable. The two participants that made attempts on the unconnected device verbalized beforehand that they did not think it was connected. Participant 4 said, "This is not a smart device. I can see now, using my special smart device window that there's nothing I can do to get her to turn that on for me, so I'm not going to try. Well, [...] now that I have her on text, I wanna see if I can get her to tell me what all the lamps are, to just confirm." She typed "List all lamps in room," and received the standard "I'm sorry, I don't know that one" response from Alexa, at which point she moved on.

Participant 7 similarly observed, "It doesn't have a thingy for it [...] so I think it's not set up?" However, she took a picture just in case. Unfortunately, the default behavior when there are no devices in the photo is to give suggestions for *all* devices that Alexa knows about. The idea behind this was to give users access to suggestions if they were inconveniently located for a photo. However, this design choice was confusing. Participant 7 typed the suggestion "Turn Light A on" since she had not interacted with Light A yet. When a different light turned on, she concluded, "I kinda think this one is just not set up to work with this device."

8.2.4 Multiple Devices Caused a Misunderstanding. In Task 2, Participant 8 was not able to figure out how to turn on the Mood Lamp. This is because she took a photo of Mood Lamp that also had Light A in the upper right corner, but the label for Light A was washed out by light. Even though the photo said "Mood Lamp" in big letters over the target device, and the suggestions interleaved "Mood Lamp" and "Light A," the Light A suggestion was the first one in the list so the user selected it. When it did not work, she then tried typing the same Light A command directly into the chat, which also did not work, so she decided to move on.

Participant 8 realized during self-directed discovery at the end of the tasks that Light A was another light and that the original was called Mood Lamp. However, she still thought that she had only received Light A suggestions previously: "That's interesting, when I tried to change Mood Lamp before, what was coming up was Light A." During the debriefing, after being shown how suggestions for multiple devices are interleaved under the capabilities, Participant 8 expressed an expectation for devices as the main subheadings, rather than capabilities: "Oh, I guess I didn't see a different hierarchy for that. So I guess for me that was a little confusing. [...] I would just assume a different heading with a different spacing under it for like, Lamp A all the way to the left, and then power slightly in, and then Floor Lamp..."

8.3 Transferring Knowledge from ARticulate to Voice

Knowledge learned from ARticulate enabled successful voice interactions, though sensors remained a challenge. After using the ARticulate app, we asked participants to set the phone aside and perform tasks using the voice interface to Alexa. After using the app and switching to voice, people knew what smart (and not smart) actuators were in the room and what their names were, including a light that they had never interacted with in previous tasks and only discovered incidentally. We can see from the large number of ideal scores for the post-app voice phase in Figure 7 that information was obtained for Tasks 1-3. We also asked users a new actuator task, Task 5, where users had to figure out to turn on Light A, even though we had not asked users to interact with Light A previously. We can see that though users guessed *Lamp* A about half the time, in keeping with the Floor Lamp and Mood Lamp naming pattern, eight out of nine users were able to accomplish the new task using voice.

However, users still had difficulty with the sensor task. Even after successfully using ARticulate, users struggled to remember how to obtain the sensor reading in the final voice phase. They knew it was possible (P5: "I know it's capable because I did it through the app"), and they knew it was a specific phrase (P7: "There was a very specific phrase that was used before"), but participants struggled to remember the details.

Two participants remembered both temperature and Multipurpose Sensor and were able to achieve the goal. Two other participants, Participant 6 and Participant 4, also remembered both pieces of information, but were not able to achieve the goal. Participant 4 said "Tell me the temperature according to Multipurpose Sensor" instead of "What is the temperature according to Multipurpose Sensor," and was frustrated when it did not work. Participant

6 said “What is the temperature...in Multipurpose Sensor?” but Alexa interrupted during the pause and gave her the weather forecast. However, the participant thought Alexa had heard the entire command, so she gave up and did not try again. Of the remaining participants, three obtained the reading by saying something about the temperature “inside” while the remaining two were not able to recall or guess a valid phrase.

Participants appeared particularly resistant to the idea that they needed a proper name for the sensor in order to get a temperature reading. The idea that only the attribute (temperature) should be sufficient was so persistent that even when using ARticulate, Participant 7 manually typed “what is the temperature” when the suggestion showed "what is the temperature according to Multipurpose Sensor." Participant 1 did not think that even temperature should have to be specified, since there was only the one sensor and the one attribute it measured: “For the sensor, if it’s the only sensor that is available in the room, maybe just say ‘A[lexa], give me a sensor reading’ versus give me a *temperature* sensor reading.”

8.4 Preserving the Intelligent Assistant Framing

Throughout the sessions, participants consistently anthropomorphized Alexa, and seemed to believe that the devices were mediated through her. For example, instead of wondering aloud “What is that device called?” Participant 2 said, “What is A[lexa] going to think that is called?” However, when it came to the ARticulate app specifically, a few participants suggested that the interaction with Alexa felt like an unnecessary layer of indirection. Participant 4 put the reason into words:

> “I didn’t know how to relate to Alexa as a human. [...] I tend to say please and thank you to Siri because she seems like a person. This app was so not – because I couldn’t hear her voice anymore—it was a little bit like, are you a robot, are you a person? [...] If I’m supposed to take pictures of things, and it’ll give me a list of commands—which was like the core of the app for me, not the texting so much as the list of commands, because I didn’t have the list beforehand—having the whole idea of there even being an Alexa seemed silly to me. If it’s just specific commands, why can’t I just press a button?”

In other words, our autocomplete suggestions were being perceived more like a menu than a true autocomplete to aid in a conversation with Alexa. You can see in Figure 2 and Figure 3 that our autocomplete design covered the screen with rigidly repeated template text, and also did not provide suggestions for any paraphrases of these “canonical” forms, even though Alexa would understand them. Even just starting a phrase with “please” like one might do conversationally would prevent suggestions from showing. In this way, the suggestions behaved much more like a menu than an autocomplete widget helping a natural conversation. Participants even tended to scroll and select, rather than type. We discuss possible directions for addressing this in Section 9.

9 IMPLICATIONS

Based on our findings, there are three main implications for readers to take away. The first is that designs like ARticulate can help users operate unfamiliar smart spaces. The second is that we need to further explore the tensions between centering intelligent assistants versus centering menus in our designs for discovery. The third is that intelligent assistant developers should revisit the natural language interface design for sensor interactions.

9.1 Bridging the Gap in Unfamiliar Smart Spaces

While the results revealed opportunities for minor design improvements, overall ARticulate’s approach enabled users to have successful one-shot natural language interactions with intelligent assistants in an unfamiliar smart spaces. Additionally, users could successfully operate the room using voice afterwards, even to use capabilities that they had not previously interacted with using the messaging app. This means that not only does ARticulate help users accomplish their goals in unfamiliar smart spaces, which no other interface has prioritized, it also

unlocks other "proper device name" interfaces and gives users the option of selecting whichever interface best meets their needs for future tasks. Designs like this enable ubiquitous intelligent assistant-operated smart spaces.

An additional unexpected takeaway was the potential for discovery-oriented designs to help with privacy concerns, particularly around sensors. Having the ability to immediately locate and identify sensors in a room using an app like ARticulate may help empower occupants by informing them of potential privacy issues.

So long as smart space interactions continue to be designed around a proper device name paradigm, the ecosystem of smart space interfaces will likely benefit from having discovery-oriented designs that combine 1) a name-free means of discovering and referring to devices, with 2) aids that help teach the proper device names for use with other interfaces. Once users can quickly discover and interact with technologies in an unfamiliar smart space, these technologies can scale and become truly ubiquitous.

9.2 We Should Explore Different Designs around Assistants and Menus

In the device-oriented paradigm that systems are currently programmed around, it is hard to know whether it makes more sense to take the ARticulate design in the direction of a more realistic conversational autocomplete, or in the direction of a better menu.

A more conversational autocomplete that helps make the language interactions feel more natural would show fewer words at a time, and therefore we would need an alternative way to provide users with an overview of the possible set of device capabilities (e.g., color, brightness, etc.) One option might be that in addition to the device name, devices could display little augmented reality "capability icons" that indicate whether they support color, brightness, and so on. Then the user knows when they snap the picture not only that the device is smart, but broadly what it is capable of, and can use the more minimal (but also more powerful) autocomplete to simply get the proper names of states (like "lavender blush") as they type.

An improved suggestion design that leans more in the direction of a tappable menu would follow the advice of several participants and start suggestions for a new photo with a drill-down list of all the device names in the picture. Users would then select a device name to reveal another drill-down list, this time of the device's capabilities, that users would again expand to show the individual canonical utterances. To help build language skills and make assistant interactions feel more natural, upon selecting an utterance, the suggestions could display a list of equivalent paraphrases that the assistant would also understand. While this device-oriented menu direction would no doubt be effective in the short term, it does not seem future-proof given the likelihood of assistants gaining smart space understanding and skills that transcend individual device commands. More work is needed to explore the trade-offs between these different directions.

9.3 We Should Revisit the Natural Language Approach to Sensors

The experiments with ARticulate revealed a potentially deeper issue with the underlying natural language understanding system itself. It would appear from our results that sensors require a different natural language approach than actuators. Sensor interactions may more intuitively follow a data-oriented, rather than a device-oriented, conceptual model. Ideally, users would not need to specify a specific device at all to receive useful information. While for large deployments this may require sophisticated aggregation techniques or understanding of deployment context, for simple deployments with only a single device that measures a particular attribute, it should be trivial for Alexa to infer which device should be queried to answer the user—the only one available.

Prior work has shown that moving towards a more data-oriented approach will also change the kinds of interactions that people will expect to have with the assistant [9]. Users will often not be interested directly in what the data from a particular device is, but rather in higher-level concepts like "when someone enters the room" or "when the room is empty," or "how much water have I used this month?" Such a system will be much more valuable to users, but the assistant will need to be redesigned to be able to handle these kinds of interactions.

10 LIMITATIONS

While we feel our results show that this is a promising direction, our study has a number of limitations, including bias in participant recruiting, the challenges of vision-based tracking, and potential scalability issues. We discuss the implications of these limitations below.

10.1 Biased Participant Pool

Due to our method of recruiting under COVID-19 limitations, the participants were close to the research team. This could be responsible for the observation that all participants had either used smart home devices or been an observer of someone using them. As mentioned, our system is designed for a future where smart spaces are available and most people would have had some prior interaction with them, so the fact that our participants were non-technical but had high exposure to smart spaces is representative of our expected users. However, there are other important ways in which the participants lacked diversity. Participant ages fell within a fairly narrow distribution over the 20s, with only two participants not in their 20s. There were no participants older than 41. Elderly people are a significant user demographic that has been shown in prior smart home work to have unique interaction desires and constraints [36, 37]. Further work would need to be done to determine whether this kind of interaction would be well-suited to the needs of elderly users.

10.2 Vision-based Tracking Challenges for Augmented Reality

Difficulties during our study highlighted the limitations of using a vision-based method of implementing augmented reality. Initial scans of the room could take anywhere from 15 minutes to hours depending on the lighting conditions and the visual complexity of surfaces in the room. Twice, users had to wait for us to redo the scan completely or deal with particularly fragile localization due to changing light conditions. This was a problem when performing scans during sunset, where every fifteen minutes the lighting conditions changed enough to completely break the localization system. Turning lights on and off during the session also occasionally changed the lighting conditions enough for the system to lose tracking. We ended up addressing these problems by running systems during daylight hours or at night, and by doing scans in rooms with permanently good lighting that would not be significantly affected by the testbed devices.

Despite the deployment difficulties that the augmented reality system posed during the study, we still believe that augmented reality feedback is a feasible avenue to pursue long term. Flagship smartphone manufacturers like Apple and Samsung have released smartphone models with ultra-wideband (UWB) radios that support high-resolution device localization and orientation, and LiDAR for creating depth scans of the environment. This additional instrumentation has the potential to significantly increase the robustness of location tracking for mobile augmented reality going forward [44].

10.3 Scalability

The study did not evaluate how well the system scales as the number of rooms and devices grows. Scaling could potentially present issues, such as a burdensome setup process and visual occlusion. For example, our setup process required the experimenter to make a scan of the room and place device labels, which could take between 15 minutes to several hours depending on the environment. While the speed of augmented reality scans may improve with better tracking technology as discussed above, configuring each room will still require some manual work. However, someone will need to manually install the devices anyway, so if the scanning time can be brought down (or eliminated completely using automatic localization of devices), adding an scan and label step to the already burdensome overhead of device installation still seems feasible at scale.

The study evaluation only used six devices in one smart room (including the Echo Dot). While this is fairly consistent with current smart spaces [20], future smart spaces could potentially involve many more devices.

ARticulate users can deal with clutter by using the camera to filter spatially, and also by typing the name of the desired device into the message bar to bring up only autocomplete results for that device. However, ARticulate would need to add additional layout logic to handle potential occlusion of multiple nearby labels.

11 FUTURE WORK

Future work remains in extending our approach beyond simple smart device interactions to more complex ones, such as those involving device groups or preset scenes. These are common features of smart spaces which do not neatly correspond to a single physical locus. Additionally, as the Internet of Things grows, more work should be done exploring this approach in smart spaces with a greater density of smart devices.

It is an open question whether this approach would be useful for automation. Smart home automation routines are usually refined over a long period of time, so creating or modifying such routines does not obviously fit the pattern of the first few interactions in an unfamiliar smart space. However, it could be that users in an unfamiliar space may nevertheless wish to discover what automation routines are running, especially to explain some surprising action that has just taken place. Assistants enabled with photo messages can potentially help users understand the automatic activity of particular devices.

Finally, while this work has focused on making it easier for users to learn the proper names of devices, we feel this is a crutch supporting a less than optimal situation. While real-world systems are likely going to continue using proper names, we encourage moving away from this approach. The proper device name paradigm is a leaky abstraction that has come up from the system design and become a part of interfaces like smart home apps and intelligent assistants, even though it is not how people intuitively invoke functionality. We should move towards multi-modal, contextually aware interactions with smart space assistants that are able to combine words with other information like gaze direction, pointing, or contextual information to resolve entities. We also need to move towards supporting interoperability and applications that are more centered around high-level goals and behaviors and less tied to individual devices. We should also support more assistant initiative for asking users to demonstrate commands, so that the assistant can learn new names and phrases. This would allow intelligent assistants to tailor their understanding to humans rather than the other way around.

12 CONCLUSION

In this work, we introduced ARticulate, an app for messaging intelligent assistants that helps users achieve one-shot interactions with intelligent assistants in unfamiliar smart spaces where each smart device has a proper name that users must reference. We used photo messaging enhanced with augmented reality labels for smart devices and autocomplete-enabled captions as mechanisms to help aid user discovery. To evaluate this design, we ran nine user evaluations in a smart home testbed. Presented with an unfamiliar smart space, ARticulate users were able to discover available functionality, including what was *not* available, and accomplish goals with a single interaction attempt. The knowledge users gained from use of ARticulate also translated into the ability to use the voice interface afterwards without a discovery aid. ARticulate is the first interface that makes unfamiliar smart spaces usable by overcoming the challenges posed by the proper device name paradigm. With interfaces like ARticulate, anyone who knows how to message can walk into any smart space and start using it with confidence.

ACKNOWLEDGMENTS

Many thanks to the members of Lab 11 who helped facilitate at-home experiments during the pandemic. Thank you also to the reviewers who made this work better. This work was supported in part by the CONIX Research Center, one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

REFERENCES

[1] Amazon.com, Inc. 2021. Echo Dot (3rd Gen) - Smart speaker with Alexa - Charcoal. https://web.archive.org/web/20210805031908/https://www.amazon.com/Echo-Dot/dp/B07FZ8S74R Accessed: 2021-08-10.

[2] Joseph B Bayer, Nicole B Ellison, Sarita Y Schoenebeck, and Emily B Falk. 2016. Sharing the small moments: Ephemeral social interaction on Snapchat. *Information, Communication & Society* 19, 7 (2016), 956–977.

[3] James H Bradford. 1995. The human factors of speech-based interfaces: a research agenda. *ACM SIGCHI Bulletin* 27, 2 (1995), 61–67.

[4] John Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* 189 (1996).

[5] Barry Brumitt and Jonathan J Cadiz. 2001. "Let there be light": Examining interfaces for homes of the future. In *Human Computer Interaction. INTERACT'01. IFIP TC. 13 International Conference on Human Computer Interaction. IOS Press, Amsterdam, Netherlands; 2001; xxvii+ 897 pp.* 375–82.

[6] Fei Cai and Maarten De Rijke. 2016. A survey of query auto completion in information retrieval. *Foundations and Trends in Information Retrieval* 10, 4 (2016), 273–363.

[7] Pew Research Center. 2021. Social Media Use in 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021 Accessed: 2021-08-10.

[8] Fang Chen. 2006. *Designing human interface in speech technology.* Springer Science & Business Media.

[9] Meghan Clark, Mark W. Newman, and Prabal Dutta. 2017. Devices and data and agents, oh my: How smart home abstractions prime end-User mental models. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3, Article 44 (Sept. 2017), 44:1–44:26 pages.

[10] Eric Corbett and Astrid Weber. 2016. What can I say? Addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services.* 72–82.

[11] Adrian A de Freitas, Michael Nebeling, Xiang 'Anthony' Chen, Junrui Yang, Akshaye Shreenithi Kirupa Karthikeyan Ranithangam, and Anind K Dey. 2016. Snap-to-it: A user-inspired platform for opportunistic device interactions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 5909–5920.

[12] Steven Feiner, Blair MacIntyre, Tobias Höllerer, and Anthony Webster. 1997. A touring machine: Prototyping 3D mobile augmented reality systems for exploring the urban environment. *Personal Technologies* 1, 4 (1997), 208–217.

[13] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. 1987. The vocabulary problem in human-system communication. *Commun. ACM* 30, 11 (1987), 964–971.

[14] Christine Geeng and Franziska Roesner. 2019. Who's in control? Interactions in multi-user smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* 1–13.

[15] Google. 2019. *Google Assistant.* https://assistant.google.com Accessed: 2019-02-14.

[16] Google. 2019. *Google Home.* https://store.google.com/us/product/google_home Accessed: 2019-09-19.

[17] Google. 2019. *How Google autocomplete works in Search.* https://www.blog.google/products/search/how-google-autocomplete-works-search Accessed: 2019-02-15.

[18] Korinna Grabski and Tobias Scheffer. 2004. Sentence completion. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 433–439.

[19] Amy He. 2020. *Amazon maintains convincing lead in US smart speaker market.* https://www.emarketer.com/content/amazon-maintains-convincing-lead-in-us-smart-speaker-market Accessed: 2021-07-06.

[20] Danny Yuxing Huang, Noah Apthorpe, Frank Li, Gunes Acar, and Nick Feamster. 2020. Iot inspector: Crowdsourcing labeled network traffic from smart home devices at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 2 (2020), 1–21.

[21] Candace A Kamm, Diane J Litman, and Marilyn A Walker. 1998. From novice to expert: The effect of tutorials on user expertise with spoken dialogue systems. In *Fifth International Conference on Spoken Language Processing.* Citeseer.

[22] Demetrios Karis and Kathryn M. Dobroth. 1991. Automating services with speech recognition over the public switched telephone network: Human factors considerations. *IEEE Journal on Selected Areas in Communications* 9, 4 (1991), 574–585.

[23] Laurent Karsenty. 2002. Shifting the design philosophy of spoken natural language dialogue: From invisible to transparent systems. *International Journal of Speech Technology* 5, 2 (2002), 147–157.

[24] Jette Kofoed and Malene Charlotte Larsen. 2016. A snap of intimacy: Photo-sharing practices among young people on social media. *First Monday* (2016).

[25] Vinay Koshy, Joon Sung Sung Park, Ti-Chung Cheng, and Karrie Karahalios. 2021. "We just use what they give us": Understanding passenger user perspectives in smart homes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–14.

[26] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31.

[27] Yubing Li, Kun Zhao, Meichen Duan, Wei Shi, Liangliang Lin, Xinyi Cao, Yang Liu, and Jizhong Zhao. 2020. Control Your Home With a Smartwatch. *IEEE Access* 8 (2020), 131601–131613.

[28] Axel Liljencrantz. 2005. *Friendly Interactive Shell.* https://fishshell.com Accessed: 2019-09-20.
[29] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing mobile voice assistants with WorldGaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–10.
[30] Sarah Mennicken and Elaine M Huang. 2012. Hacking the natural habitat: An in-the-wild study of smart homes, their development, and the people who live in them. In *International Conference on Pervasive Computing*. Springer, 143–160.
[31] Robert C Miller, Victoria H Chou, Michael Bernstein, Greg Little, Max Van Kleek, David Karger, and MC Schraefel. 2008. Inky: A sloppy command line for the web with rich visual feedback. In *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*. 131–140.
[32] Arnab Nandi and HV Jagadish. 2011. Guided interaction: Rethinking the query-result paradigm. *Proceedings of the VLDB Endowment* 4, 12 (2011), 1466–1469.
[33] Robert Neßelrath, Chensheng Lu, Christian H Schulz, Jochen Frey, and Jan Alexandersson. 2011. A gesture based system for context–sensitive interaction with smart homes. In *Ambient Assisted Living*. Springer, 209–219.
[34] Jenni Niemelä-Nyrhinen and Janne Seppänen. 2020. Visual communion: The photographic image as phatic communication. *New Media & Society* 22, 6 (2020), 1043–1057.
[35] Sophie Nyborg. 2015. Pilot users and their families: Inventing flexible practices in the smart grid. *Science & Technology Studies* (2015).
[36] Debajyoti Pal, Tuul Triyason, and Suree Funikul. 2017. Smart homes and quality of life for the elderly: a systematic review. In *2017 IEEE International Symposium on Multimedia (ISM)*. IEEE, 413–419.
[37] François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon. 2013. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. *Personal and Ubiquitous Computing* 17, 1 (2013), 127–144.
[38] Jun Rekimoto and Katashi Nagao. 1995. The world through the computer: Computer-augmented interaction with real world environments. In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*. 29–36.
[39] Franziska Roesner, Brian T Gill, and Tadayoshi Kohno. 2014. Sex, lies, or kittens? Investigating the use of Snapchat's self-destructing messages. In *International Conference on Financial Cryptography and Data Security*. Springer, 64–76.
[40] Eric Rose, David Breen, Klaus H Ahlers, Chris Crampton, Mihran Tuceryan, Ross Whitaker, and Douglas Greer. 1995. Annotating real-world objects using augmented reality. In *Computer Graphics*. Elsevier, 357–370.
[41] Frode Eika Sandnes, Jo Herstad, Andrea Marie Stangeland, and Fausto Orsi Medola. 2017. UbiWheel: a simple context-aware universal control concept for smart home appliances that encourages active living. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 1–6.
[42] SmartThings, Inc. 2021. SmartThings. https://web.archive.org/web/20210727135832/https://www.smartthings.com/ Accessed: 2021-08-10.
[43] Snap, Inc. 2021. SEC Quarterly Report Form 10-Q. https://investor.snap.com/financials/sec-filings/default.aspx Accessed: 2021-08-10.
[44] Yang Song, Mingyang Guan, Wee Peng Tay, Choi Look Law, and Changyun Wen. 2019. UWB/LiDAR Fusion For Cooperative Range-Only SLAM. In *2019 International Conference on Robotics and Automation (ICRA)*. 6568–6574.
[45] Ivan E Sutherland. 1968. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, Fall Joint Computer Conference, Part I*. 757–764.
[46] Arto Vihavainen, Juha Helminen, and Petri Ihantola. 2014. How novices tackle their first lines of code in an IDE: Analysis of programming session traces. In *Proceedings of the 14th Koli Calling International Conference on Computing Education Research*. ACM.
[47] Ryen W White. 2018. Skill discovery in virtual assistants. *Commun. ACM* 61, 11 (2018), 106–113.
[48] Ryen W White and Gary Marchionini. 2007. Examining the effectiveness of real-time query expansion. *Information Processing & Management* 43, 3 (2007), 685–704.
[49] Andrew Wilson and Hubert Pham. 2003. Pointing in intelligent environments with the WorldCursor. In *Interact*. Citeseer.
[50] Jason Wither, Stephen DiVerdi, and Tobias Höllerer. 2009. Annotation in outdoor augmented reality. *Computers & Graphics* 33, 6 (2009), 679–689.
[51] Jong-bum Woo and Youn-kyung Lim. 2015. User experience in do-it-yourself-style smart homes. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 779–790.
[52] Huiyue Wu, Shaoke Zhang, Jiayi Liu, Jiali Qiu, and Xiaolong Zhang. 2019. The gesture disagreement problem in free-hand gesture interaction. *International Journal of Human–Computer Interaction* 35, 12 (2019), 1102–1114.
[53] Nicole Yankelovich. 1996. How do users know what to say? *Interactions* 3, 6 (1996), 32–43.
[54] Elizabeth Zoltan-Ford. 1991. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies* 34, 4 (1991), 527–547.