On Inclusion: Video Analysis of Older Adult Interactions with a Multi-Modal Voice Assistant in a Public Setting

Andrea Cuadra
andreacuadra@stanford.edu
Stanford University
USA

Hyein Baek
Cornell Tech
USA

Deborah Estrin
Cornell Tech
USA

Malte Jung
Cornell University
USA

Nicola Dell
Cornell Tech
USA

ABSTRACT

Older adults around the world lack access to a wide range of potentially life-changing digital applications, services, and information that could be provided by voice assistants (such as Amazon's Alexa, Google's Assistant, or Apple's Siri). However, older adults' needs are underrepresented in the design of voice assistants. Because of this, we are missing opportunities for digital inclusion, and increasing risks of excluding older adults as these devices permeate public settings. In this work, we video record older adults (n=26) interacting with a multi-modal voice assistants while waiting in line at food pantries, and use Interaction Analysis to draw insights from these recordings. We find that by being agnostic to body language, audio-prosodic features, and other contextual factors, voice assistants fail to capture and react to some important aspects of interactions. We discuss design (e.g, interpreting users' posture as a cue to wake the device when they are leaning towards the device) and research (e.g., surveillance trade-offs) implications, and argue for the use of multi-modal inputs with attention to privacy. Designing and training voice assistants to take in and appropriately respond to non-verbal cues may increase their inclusivity, helping them fulfill important needs of our aging population.

CCS CONCEPTS

• **Human-centered computing**; • **Human computer interaction (HCI)**; • **Empirical studies in HCI**;

KEYWORDS

Older adults; video analysis; smart speaker; voice assistant; inclusive design; Alexa

ACM Reference Format:
Andrea Cuadra, Hyein Baek, Deborah Estrin, Malte Jung, and Nicola Dell. 2022. On Inclusion: Video Analysis of Older Adult Interactions with a Multi-Modal Voice Assistant in a Public Setting. In *International Conference on Information & Communication Technologies and Development 2022 (ICTD2022), June 27–29, 2022, Seattle, WA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3572334.3572371

ICTD2022, June 27–29, 2022, Seattle, WA, USA

ACM ISBN 978-1-4503-9787-2/22/06…$15.00
https://doi.org/10.1145/3572334.3572371

1 INTRODUCTION

Older adults around the world lack access to a wide range of potentially life-changing digital applications, services, and information that could be provided by voice assistant technology (such as Amazon's Alexa, Google's Assistant, or Apple's Siri). Voice assistants are a promising platform, because they are meant to be easy to use. People simply have to speak to them to get a response. This means that, for people who do not feel comfortable or are unable to use a computer or a smartphone, voice assistants could be a portal for accessing the benefits of the Internet [17]. Voice assistants can also be used to encourage healthy behaviors [23] and meet home health needs [45, 92]. In addition, research in healthcare is increasingly recognizing the importance of social connection as a health determinant [21, 86] and voice assistants can help enable socialization [67]. Moreover, as some have argued, accessibility came by accident [69], meaning that voice assistants are already widely used, which may reduce the stigma associated with devices meant to address age-related disabilities [32]. If inclusive design issues are not addressed with immediacy, we risk not only not serving marginalized groups in ways that could be greatly beneficial, but also excluding them from everyday digital activities. As hinted by Google's Duplex, a humanlike automated system that makes calls on behalf of its users [48], voice-based artificial intelligent interfaces will be permeating our lives, whether we acquire them ourselves or not. Thus, people who are underrepresented in the design and research process of these systems may have to deal with frustrating experiences with few avenues to opt out (e.g., if they are on the receiving end of a Duplex call).

In this work, we employ an inclusive design orientation, which brings the needs of the people at the margins to the design of mainstream products and services [28]. We do so by focusing on the needs and experiences of older adults, a globally marginalized group within the context of Information and Communication Technologies (ICT) [75], which despite its growing proportion in the population [20, 58] has been largely excluded from the design and research of now mainstream voice assistants [77]. We recruit diverse participants from historically underrepresented groups in the design of voice assistants to study their interactions with a smart speaker. The purpose of this paper is to study their needs not in relation, but rather in addition, to the needs of people already

addressed in existing designs. Older adults' exclusion, in particular the exclusion of those with intersecting marginalized identities, may contribute to why some older adults abandon voice assistants. Trajkova and Martin-Hammond [85]'s study of why older adults use, limit, and abandon voice assistants found that most participants stopped using these devices due to their difficulty finding valuable uses, beliefs about the lack of essential benefits of the voice assistant, or challenges with use in shared spaces [85].

One possible way for voice assistants to become more inclusive could be by obtaining information from non-verbal cues, as people do in human-human communication [22, 55]. Humans naturally react to other humans' body language, facial expressions, and acoustic-prosodic features (intonation, tone, and rhythm), often subconsciously. Ekman and Friesen [35] characterized the category of nonverbal acts that maintain and regulate the back-and-forth nature of speaking and listening as *regulators*. Regulator actions occur in the attentional periphery; people perform them without thought, but can recall and repeat them if asked [35]. Despite the human-likeness of voice assistants, non-verbal cues are, for the most part, currently being overlooked by voice assistants. We utilize a framework developed by Suchman that analyzes the information available to the user, the information available to the machine, and their intersection [81]. In this paper, we refer to the information that is not mutually available to both communication partners (i.e., the human and the voice assistant) as *the human-machine communication gap*.

Although the use of video analysis is common in industry [50], existing research on voice assistant usage by older adults predominantly relies on usage logs, interviews, or product reviews [14, 60, 64, 68, 72]. We analyze the image, audio, and human-machine communication gaps in video recorded interactions of 26 older adults, who are predominantly novice users of voice assistants, with an Amazon Echo Show 10. **In particular, we 1) seek to characterize challenges in interactions with voice assistants that may obstruct inclusion, and 2) identify alternate paths that may mitigate these challenges.**

We chose to conduct our study with older adults who are predominantly novice users of voice assistants for several reasons. Although experienced users may adapt their behaviors over time as they learn how voice assistants respond, the experiences of first-time users are extremely important in determining whether someone will deem it worthwhile to adopt the technology at all [46, 93]. This may be particularly true for older adults, who may be more hesitant to use new technologies. Furthermore, not developing expertise in the privacy of one's home may result in exclusion from everyday digital activities as these technologies permeate public spaces. These encounters could become embarrassing, scary, or frustrating for novices. Moreover, although some older adults may have caregivers who could help them to learn how to use the technologies, such assistance unnecessarily increases dependence. Hence, we studied the difficulties that novice older adult users encounter when interacting with voice assistants, and how we might make these technologies more usable to them.

Our findings reveal gaps in human-machine communication that often result in the voice assistant reacting inappropriately, interrupting the user, or not responding at all. We (1) describe human-machine communication gaps revealed by our data, differentiating information that was overlooked by the machine (e.g., interaction attempts, the presence of more than one user) from information that was overlooked by participants (e.g., the indication that the voice assistant was not actively listening, and technical terminology). We then (2) take a closer look at body language features of the interactions and categorize them into those that provide reliable signals (e.g., leaning forward and gaze), and those that are somewhat ambiguous (e.g., laughing). Finally, we (3) analyze audio-prosodic features, such as rhythm (e.g., interruptions during pauses in speech), and tone and intonation (e.g., associations between various tones and intonations and interaction outcomes). Together, our findings show that by being agnostic to body language, audio-prosodic features, and other contextual factors, voice assistants fail to capture and react to some important aspects of interactions. Designing and training machines that take in and appropriately respond to non-verbal cues might be a crucial step in building voice assistants that can fulfill important needs of our aging population.

We present design and research implications for the ICTD community. In terms of design implications, we provide recommendations addressing interaction errors that result from not being able to successfully wake the voice assistant, such as by relying on design paradigms that may be more familiar to older adults. We suggest ways in which automatic detection of non-verbal cues can be used to improve interactions with voice assistants, such as having the voice assistant analyze a user's posture to determine whether they are attempting to engage with the voice assistant. We then emphasize differences and complexities for adapting voice assistants' interactions to older adults' needs and abilities in the context of prior research about code switching and knowing the user [13, 27], and discuss several ethical design considerations. In terms of research implications, we surface questions surrounding how we might use recent technological advancements to recognize body language and audio-prosodic features, and discuss the societal implications and tradeoffs associated with higher levels of surveillance. Taken together, our contributions help to relieve some of the burdens placed on older adults to adapt to the constraints imposed by new technologies, allowing older adults to appropriately benefit from the technologies' promises and improving inclusion in everyday digital activities.

2 RELATED WORK

In this section, we situate our work by first explaining what voice assistants are, describing existing research on older adults' interactions with voice assistants, and finally elaborating on the importance of non-verbal communication in ICT research.

2.1 Voice assistants

Voice assistants are autonomous speech-first software agents that can perform tasks upon receiving spoken requests—they first transcribe spoken words to text, then derive meaning from the text, and last respond using speech, and/or another modality [74]. Because voice assistants are software agents, they must be embodied via hardware. The physical devices that house voice assistants may be principally designed for the voice assistant, like the Amazon Echo or Google Home smart speakers (which are frequently referred to by the name of the voice assistant they house—e.g. in

reference to the Amazon Echo, people tend to say “ask the Alexa for the weather"), or multi-use devices such as laptop computers or smartphones [2, 4, 8]. Additionally, there are many devices, such as light bulbs, electrical plugs, locks, or vacuum cleaners, that can be connected to a device that houses a voice assistant via supporting software applications, many times requiring another device such as a smartphone or tablet [1, 5, 6]. Moreover, an increasing number of products are being released with built-in voice assistants [7, 9]. The embodiment of a voice assistant can dictate its role and its importance in a particular setting. For example, recent work by Doyle et al. [33], found that a screenless smart speaker with Amazon Alexa was perceived as more emotive and engaging but less flexible and contextual than a smartphone-based Siri voice assistant.

2.1.1 Voice assistants in public settings. Even though voice assistants are currently mostly used in private spaces (e.g., homes and cars), they are becoming more common in more-public venues (e.g., hotels, schools, and stores) [73, 87]. In an ethnographically-oriented study published in 2017, Porcheron et al. [65] explored how groups of friends interacted with Siri at a coffee shop, identifying insights, such as that participants had to rely on the screen of their devices to share parts of interactions with each other. Similarly, Cowan et al. [29] studied infrequent users of voice assistants, finding that cultural norms affected some participants’ willingness to use Siri in public. During the same year that these studies were published, scholars from industry and academia met at CSCW to discuss the use, research, and design of conversational agents, such as voice assistants, in social and collaborative settings, raising the importance of this topic of research [63]. Since then, some have studied voice assistant interactions in multi-user home settings, questioning how conversational voice assistants truly are [64]. Despite general agreement on the importance of studying voice assistant use in public settings, to the best of our knowledge, no one has focused on studying voice assistant use by older adults in public, potentially excluding a growing segment of our population that could highly benefit from, or be excluded by, this technology.

2.2 Older adult interactions with voice assistants

Research on older adults interactions with voice assistants is scarce [77], leaving many open issues to be investigated [71]. A few studies investigating how older adults use voice assistants [18, 78] point to the promise of these devices for providing older adults with access to valuable information and services. For example, O’Brien et al. found that older adults use voice assistants for companionship (e.g., *“You can ask Alexa the same question 50 times and she won’t get irritated with you*”) or to notify their caregivers if there is an emergency [60]. Similarly, Hoon et al. analyzed usage logs, finding that older adults used the voice assistant more when compared to younger adults [61]. Analysis of the content of these interactions by Chung et al. revealed that older adults tended to personify the agent more by using polite words such as “grateful", viewing it more as a companion [26]. By contrast, younger adults tended to consider it as a tool by placing more importance on its convenience [26]. Masina et al. found that users with motor, linguistic, and cognitive impairments, which tend to increase with age, can effectively interact with voice assistants, as long as they are able to repeat full sentences and have a Mini-Mental State Examination score greater than 24 [51]. Moreover, in 2020, Pradhan et al. [68] conducted a general-use, three-week field deployment of the Amazon Echo Dot in the homes of seven older adults, and found consistent usage for finding online information. In another more-recent study, Kim and Choudhury [43] found that over time older adults felt less worried about making mistakes and enjoyed the digital companionship as they got used to using voice assistants.

Several efforts focus on helping voice assistants to reach, and become more useful for, older adults. Voice apps designed for older adults (e.g. Alexa’s Ask My Buddy[1], or Google Assistant’s Vigil Connect[2]) have many positive ratings on their respective app stores, an indication of their traction. Moreover, there are many startups emerging that specifically aim to meet the needs of older adults and their caregivers via voice assistants, including: Aiva Health[3], LifePod[4], and SoundMind[5]. These academic findings and industry efforts signal the promising role that voice assistants may play in the lives of older adults.

However, voice assistants are not yet easy to use or understand for many older adults. Trajkova and Martin-Hammond [85] cited difficulty in finding valuable uses, beliefs about the lack of essential benefits of the voice assistant, or challenges with use in shared spaces as key reasons for older adults abandoning the use of these devices. More specifically, some of Trajkova and Martin-Hammond [85]’s participants mentioned not wanting to bother others who they shared a living space with, concerns about being surveilled, stigma surrounding getting help from a technology for something they could do independently, and awkwardness or distaste for the speech modality. Despite the abandonment, participants in Trajkova and Martin-Hammond [85]’s study also saw the potential for voice assistants to support aging and independent living in the future. Our study complements these interview-based findings by directly looking at video recordings of older adults interactions with voice assistants.

2.3 The importance of non-verbal communication in ICT research

Non-verbal forms of communication have been deemed important in the ICT communities for a long time. In 1994, Nagao and Takeuchi acknowledged the multiplicity of communication channels that act on multiple modalities, and set out to study how humans would react to facial expressions from a machine in human-computer dialogue [55]. Shortly after, Reeves and Nass published The Media Equation, supporting the claim that we attribute characteristics to machines in the same way we do to humans [70]. In the same line of research, Cassell et al. analyzed human monologues and dialogues that suggested that postural shifts can be predicted as a function of discourse state in monologues, and discourse and conversation state in dialogues [22]. As a result, they designed an embodied conversational agent that could change its posture [22]. Moreover, Liebman and Gergle examined the role of nonverbal, paralinguistic cues in computer-mediated, text-based communication, such as

[1] www.amazon.com/Beach-Dev-Ask-My-Buddy/dp/B017YAF22Y
[2] https://assistant.google.com/services/a/id/581192e4fd1a63df/
[3] www.aivahealth.com
[4] www.lifepod.com
[5] www.soundmindinc.com

punctuation and emoticons, and found a positive causal relationship of conversation duration and cue use on perceived affinity, and that reciprocity may play a central role in supporting this effect [49].

Designing non-verbal expressions for voice agents impacts how humans react to them; research has found that matching the tonality of a voice assistant's speech to the mood of its human user results in better performance [39], that gender stereotypes carry over to gendered synthetic voices [56], and that we consider different voices from the same device to be different social actors [56], mimicking how we may distinguish different people talking on a telephone. Additionally, Jung et al. found that although robots that used backchanneling improved team functioning, backchanneling robots were perceived as less intelligent than those that did not use backchanneling [41].

Recently, Cuadra et al. found that that voice assistant self-repair greatly improves people's assessment of an intelligent voice assistant if a mistake has been made, but can degrade assessment if no correction is needed [31]. Even though Cuadra et al.'s findings rely on an assumption of error-recognition based on visual cues by the voice assistant, to the best of our knowledge, no work has successfully examined how voice assistants may interpret non-verbal expressions displayed by their users. This is despite a recent line of work studying the human-likeness of human-agent conversations. Motivated by key characteristics of human-human conversations that do not get captured by conversational agents, Clark et al. studied what features people value in conversation, calling for a redefinition of design parameters for conversational agent interaction [27]. They argue that participants describe the need for mutual understanding and common ground, trust, active listenership, and humor as crucial social features in human conversations, but in agent conversations these are described almost exclusively in transactional and utilitarian terms [27]. Beneteau et al., support this argument by recognizing that to improve communication repair strategies, knowledge of the context and the communication partner is extremely helpful, allowing digital home assistants to artificially code switch as needed [13]. The tension between the human-likeness of voice assistants, and their inability to meet the expectations that their appearance sets might contribute to the fluid movement between "human-like" and "object-like" categorizations displayed by older adults in Pradhan et al's study [67]. Taken together, these studies call for improvements in voice assistants' abilities to understand and react to non-verbal cues, especially because of their implied humanness.

The importance of context in human-machine interactions is well known in the ICT communities [31, 41, 80, 84]. Additionally, we know that behavioral responses to robots, from which context can be extracted, are in a large part non-verbal [44]. Research has also made technological strides in the last decade in sensing [3, 76, 79, 88] and computer vision [24, 38, 89]. With these considerations in mind, we set out to identify and valorize the visual and audio-prosodic elements present in older adult interactions with voice assistants.

3 APPROACH

We conducted an IRB-approved field study with older adults who visited senior centers, and video recorded their interactions with a voice assistant. We now provide a description of the settings in which we conducted our observations, details about the participants, and explain our methodological and analytic orientations.

3.1 Research Setting

We situate our study in senior centers, which can be categorized as "third place" settings. A "third place" setting is described by Oldenburg [62] as a place where one relaxes in public, encounters familiar faces and makes new acquaintances.[6] We chose this setting as way to capture the heterogeneity of the older adult population while also engaging with a central theme demarcating the ubicomp of the present, the "messiness of everyday life" [12]. Senior centers are community centers designed to make older adults feel supported, and happy—they bring older adults together for a variety of services and activities designed to enhance their quality of life [11]. Both of the senior centers in our study had computer labs with programming to teach older adults computer skills. According to the *National Council on Aging*, "Compared with their peers, senior center participants have higher levels of health, social interaction, and life satisfaction and lower levels of income."[7] To capture use in public, we set up research booths with a camera facing the participants (Figure 1) near food pantry lines—food pantries offer free groceries to members on a periodical basis—in two senior centers in a large U.S. city. Our "in the wild" [30] approach allowed us to capture public interactions with voice assistants that are becoming increasingly common in public places [87].

3.2 Recruitment and Participants

We approached older adults who visited the center and invited them to participate in our study. We explained the purpose of the study, what we were asking participants to do, and sought their permission to video capture their interactions with a voice assistant. Consent forms were available as physical copies placed on a table, and consent was obtained verbally. The researchers followed recommended health and safety protocols during the explorations to keep both participants and researchers safe during the COVID-19 pandemic.

In total, we recruited 26 participants (20 women), who were on average 73 years old. Table 1 summarizes participant demographics. Participants were visiting the senior center for the food pantry: some were picking up food and others were organizing the pantries. To pick up food, participants must attest to income levels below a certain threshold (typically less than $2200 per month if there is just one person in the household). Senior center staff reported that most members owned smartphones, echoing our participants' responses when asked about their current technology usage. Most participants (n=16, eight unreported) owned one or multiple computing devices, including smartphones, tablets, laptops, or desktop computers. They reported using their computing devices for a variety of reasons, including: information retrieval, messaging others, audio and video calls (including doctor appointments), reminders, social media, playing games, viewing or attending religious events, taking photos, playing music, writing, accessing specific websites, shopping online, and paying bills. All participants who owned and

[6] Oldenburg [62] calls the "first place" the home, and the "second place" the workplace.
[7] https://www.ncoa.org/article/get-the-facts-on-senior-centers

used a computing device had access to the Internet. Some participants (n=5) indicated using speech-to-text functionality of their phones, tablets, or computers, but none expressed knowing how to send voice notes, such as the ones supported by iMessage or Whatsapp. Most participants (n=18, eight unreported) were at least somewhat confident reading and writing; however, three participants expressed declining confidence due to age-related cognitive, motor, or visual impairments. Participants lived in their homes, predominantly with relatives. Most participants (n=19) reported never having used a voice assistant before. We considered participants novices if they reported having used a voice assistant before, but did not feel very confident in their abilities using it or whose interactions suggested novice-level expertise. Even though our counts (see Section 3.5) include interactions from users with some experience (e.g., P5 or P6), we only use one specific example from non-first-time users in our findings—P5 & P6 playing Trivia together—, which we call out as such. We included all participants in our interaction counts (including P5, our most experienced participant), because they are representative of the heterogeneity of the older adult population and the "messiness of everyday life" [12]. Additionally, most of our non first-time user participants were still novices.

Because of the in-the-wild nature of the study, some participants arrived in pairs and interacted with the device in pairs (three pairs, n=6), which we see as resembling how real-world interactions with voice assistants might take place (e.g., several people might be in the room where the voice assistant is installed). However, because we segmented the data for analysis, we were able to extract individual interactions from participants who arrived in pairs. In most cases, one participant spoke while the other listened. In rare cases, participants responded in unison, these segments were annotated accordingly. We kept an eye on potential influences paired individuals could have on each other, and made note of them in the findings. However, for the most part, since all participants interacted in public, they all knew they were being watched, providing something of a control for potential behavioral differences caused by The Hawthorne effect [52].[8]

3.3 Procedure

The booths included signs indicating we were conducting a research study, a voice assistant, and a camera from the perspective of the voice assistant. The voice assistant was placed on top of a table, and a chair was positioned nearby for participants to have the option to sit. We told participants that had never interacted with a smart speaker before that the device on the table was a smart speaker that responded to speech, and explained that they could initiate conversations with it by saying its name, Alexa, followed by a command. We temporarily muted the device to provide utterance examples such as, "Alexa, hello" or "Alexa, what's the weather".[9]

After receiving this guidance, participants were instructed to freely interact with the voice assistant, and we pointed at signs with utterance suggestions. These signs were posted on the wall behind the table the device was on. The messages on the signs suggested participants to say "Alexa, hello," "Alexa, what are the symptoms of COVID-19," "Alexa, what can you do," and "Alexa, what's the weather." The first author was available throughout all the sessions, usually sitting somewhere near the participant but outside the participant's field of view. The researcher occasionally provided support to participants, such as when a participant seemed stuck, was unable to wake the device, or looked at the researcher for guidance. Often, even if participants seemed to be getting frustrated, the researcher would simply suggest that they keep trying. For example, we did not intervene in the three occasions in which participants introduced themselves to Alexa, and Alexa initiated a voice training "setup" activity. But we did intervene if Alexa was not responding at all after several failed attempts, encouraging participants to speak louder, sometimes escalating the suggestion by telling participants to imagine they were upset at Alexa. Whenever possible, we asked participants what they had noticed or thought of the interaction, or if a request had gone as they expected.

Figure 1 depicts the setup used. We used an Amazon Echo Show 10 device, which has a 10-inch touchscreen, set to a default American female voice and wake word "Alexa". We chose this embodiment for our voice assistant research because the touch screen complements the audio modality, providing additional information for older adults who are prone to experiencing a variety of cognitive, audio, visual, and motor impairments. The voice assistant was configured to speak in either English or Spanish, depending on the language used to address it.

Study sessions lasted approximately 30 minutes including obtaining consent, the initial introduction, the interaction itself, and the post-interaction interviews. However, sessions in which too many interactions failed from the very beginning were much shorter, as participants did not wish to continue engaging. After capturing the interactions on video, we clipped the videos into smaller segments we called "chunks" (described in Section 3.5 along with the resulting dataset).

3.4 Video Analysis Methodology

We used video analysis methodology to carefully study the interactions of older adults with the multi-modal voice assistant. By observing these interactions, we hoped to uncover important insights for the design of voice assistants, which tend to be used in the private space of a home, where the visual elements of interactions are not usually captured. Many have employed video analysis methodology to capture patterns that would not be visible without video (e.g., by playing the video at slow or accelerated speeds), collect primary empirical data, and have more consistency and reliability in observations [40, 51, 81–83, 90].

Our work utilizes Interaction Analysis [40]. According to Jordan and Henderson [40], Interaction Analysis commits to grounding theories of knowledge and action in empirical evidence with the goal of identifying "regularities in the ways in which participants utilize the resources of the complex social and material world of actors and objects within which they operate." We chose this method because it would allow us to reconstruct the events, keep and replay the primary record, and capture the complexity of the data.

[8]The Hawthorne effect refers to a type of reactivity in which individuals modify an aspect of their behavior in response to their awareness of being observed.
[9]We skipped this step for non first-time users.

Participant demographics (n=26)	
Age	Avg: 73, Median: 74, SD: 7.74
Gender	Female: 20, Male: 6
Language used	English: 18, Spanish: 7, Korean: 1
Latinx	No: 15, Yes: 11
Race	Black: 10, White: 3, Native American: 2, Asian: 1, Other or mixed: 9, Declined to answer: 1
Prior experience with a smart speaker	First-time users: 19, Non first-time users: 7
Confidence using speech-based computing device (after interaction)	Very confident: 6, Somewhat confident: 1, Only a little confident: 1, Not at all confident: 10, Unreported: 8
Highest degree or level of school you have completed	Less than a high school diploma: 2, High school degree or equivalent: 8, Some college - no degree: 5, Bachelor's degree: 3, Master's degree : 1, Unreported: 8
Gross income ($)	<20k: 11, 20-40k: 6, 80-100k: 1, Unreported: 8
Living alone	No: 11, Yes: 7, Unreported: 8
Own and use at least one computing device	No: 2, Yes: 16, Unreported: 8
Frequency of use of computing devices to go online	Less than once a week: 2, About once a week: 1, About once a day: 4, Multiple times every day: 11, Unreported: 8
Confidence using computing device	Very confident: 6, Somewhat confident: 5, Only a little confident: 3 , Not at all confident: 4, Unreported: 8
WiFi at home	No: 7, Yes: 11, Unreported: 8
Confidence reading and writing	Very confident: 15, Somewhat confident: 3, Unreported: 8

Table 1: Demographic details of study participants.

3.5 Data Analysis

After we captured the videos, we employed a bottom-up approach to analyze the data. We began by watching all the videos and created a rough content log, as described by Suchman and Trigg [82]. Then, we watched all the videos at 2x and 4x speeds to see if anything stood out, as replaying the videos at different speeds can help to see patterns that were otherwise not noticeable [40]. By doing this, we were able to identify our a unit of analysis, which we called a *chunk*. We based this decision on previous work by Weingart et al., which identify "units" to be coded [91], and Jordan and Henderson, which relies on "ethnographic chunks" to break down large videos into smaller, more analyzable, segments [40]. The first chunk of a participant's interaction always started when a participant addressed Alexa for the first time, and ended when there was an interruption or the participant addressed Alexa again. Subsequent chunks started at the end of the previous chunk (see Figure 2 for an example). In parallel, we selected our own analytic foci: body language, audio-prosodic features (such as tone, intonation, and rhythm), and human-machine communication gaps. Jordan and Henderson define analytic foci as "ways of looking that are quite consistently employed in Interaction Analysis" [40].

- **Body language**: For every interaction, we carefully annotated all aspects that visually changed during an interaction including aspects relating to gaze, posture, and facial expressions. We then ascribed meaning to these, and evaluated whether they were signs that the conversation was going well or poorly. Based on these assessments, we extracted body language features, such as gaze and posture, that we were able to consistently identify and make inferences from, and ones that were more difficult to tell apart, such as laughter.
- **Audio-prosodic features**: We categorized every chunk based on its rhythm, tone, and intonation. For rhythm, we noticed that participants often did not pause or paused for too long to fit Alexa's listening window, so we labeled chunks accordingly. We open-coded the tone as Upset, Nervous, Friendly, Exaggerated, Indifferent, Excited, or Tired, and the intonation as Fall, Rise, Rise-Fall, or Same (Constant). Two authors watched multiple similar clips together, discussed possible descriptions for those clips, and subsequently agreed on the aforementioned codes for their tone and intonation. We then used these annotations to inform our inferences about participants' behaviors during interactions.
- **Human-machine communication gaps**: We noticed that the machine was missing a lot of important signals emerging from non-verbal communication. Thus, we employed Suchman's analytic framework [81] to compare and contrast the information available to both the older adults and the voice assistant, with the information only available to one or the other. An excerpt of what this analysis looked like is available in Figure 2, and immediately noticeable from these two chunks (amounting to only 32 seconds of interaction) is the quantity of user actions, in particular non-verbal, that were not interpreted by the voice assistant (everything not highlighted in yellow).

Chunks were labeled based on their outcome: "Success", "Failure", or "Ambiguous," and ones that were parts of interruptions were marked as "Help," or "Unrelated." Clipping, numbering, and appropriately labeling interruptions or interventions allowed us to exclude them from the analysis while still noting that an intervention had occurred right before a particular interaction. Doing so was important to maintain the order integrity of each participant's interactions at a higher-level while being able to code and analyze each chunk in detail. Our criteria were:

- **Failure:** The interaction failed. For example, if the participant did not succeed at "waking" the device, or if the device misheard a participant's request.
- **Success:** The interaction succeeded—a user made a request, and the machine responded appropriately. At some point, there was an agreement in understanding from the user and the machine. For example, the user asked for a joke, and the voice assistant told a joke.

Figure 1: A participant interacts with the voice assistant at a senior center. A camera behind the voice assistant records the participant. Wooden panels label the booth as a research study, and provide suggestions with example utterances to interact with the voice assistant.

- **Ambiguous:** These happened when the goal was not clear, or the success was only partial, so we were not able to classify them as successes or failures with certainty. For example, a chunk in which a participant said "*thank you*" to the machine was classified as ambiguous, because the machine did not do anything but no response was necessarily expected. The machine did not register that it was thanked, meaning that the perceived "interaction" was just a one-way communication. Also classified as ambiguous were chunks based on the context of previous interactions. For example, if a participant had been ignored for several chunks, and the machine finally responded to them but not with what they were asking. In this case, getting a response from the machine after having been repeatedly ignored was considered a partial achievement, rather than a clear failure.

Once we clipped all the interactions into chunks, we coded each chunk. In our dataset, each chunk has an index number, a participant number, a chunk number (starts at one for each participant, except for participants interacting in pairs), a duration, and an outcome. When the wake word was said, we labeled how it was pronounced, and the order in which it was pronounced. Audio prosodic features (rhythm, tone, and intonation) were marked for each chunk, and body language features were noted when they occurred. In some instances, we played the clip aloud in front of an active Alexa device to reconstruct the event, and verify that our codes were accurate. Similarly, we reviewed usage logs from the interactions to see how the voice assistant interpreted the information. Each chunk was initially coded by one researcher and then reviewed by another researcher. All disagreements in the codes were discussed until

agreement was reached. We do not report inter-rater reliability since all data was double-coded and disagreements were reconciled [53].

Summary of dataset: The dataset we gathered included 221 interaction chunks (56 minutes and 37 seconds of footage) from 26 participants. The longest interaction had 44 chunks, lasting 18 minutes after excluding "help" or "unrelated" clips. The shortest interaction had one chunk, lasting nine seconds. Out of these, 68 were labeled "Success", 92 "Failure", and 61 "Ambiguous". We excluded from the dataset 47 chunks in which participants were not interacting with the voice assistant, but marked their position to account for interventions and/or interruptions. Appendix A provides a summary of observed behaviors coded in our dataset.

4 FINDINGS

We start by (1) describing human-machine communication gaps revealed by our data, broken down into information that was overlooked by the machine (e.g., interaction attempts, the presence of more than one user) from information that was overlooked by participants (e.g., the indication that the voice assistant was not actively listening, and technical terminology). We then (2) take a closer look at body language features of the interactions and categorize them into those that provide reliable signals (e.g., leaning forward and gaze), and those that are somewhat ambiguous (e.g., laughing). Finally, we (3) analyze audio-prosodic features, such as rhythm (e.g., interruptions during pauses in speech), and tone and intonation (e.g., associations between various tones and intonations and interaction outcomes).

4.1 Human-machine communication gaps

Our analysis highlights gaps in communication between participants and the voice assistant that led to interaction challenges. In some cases, these gaps corroborate prior work reporting similar challenges [13, 27, 64, 65]. Our goal here, however, is to ultimately show how considering audience, body language and/or audio-prosodic features might help to overcome these gaps. At a high level, we found that older adults' challenges interacting with voice assistants were often due to a lack of conversational grounding, where the voice assistant did not understand older adults' expressions, and participants did not reliably understand the voice assistant's state. In this section, we describe data that was overlooked by the machine and data that was overlooked by our participants. Note, these are not necessarily a comprehensive list of all gaps, rather, they are examples that stood out to us as most relevant for inclusion.

4.1.1 Overlooked by the machine. This section describes information that was overlooked by the voice assistant.

Second by second interaction data. Our analysis reveals valuable information overlooked by the machine.Standard video is usually shot at 30 frames per second. If we look at just one frame for every second in only two interaction chunks, reducing the data to $1/30^{th}$ the size, and label a participant's gaze, posture, and facial expression we can make many inferences about an interaction (Figure 2). By adding context from previous interactions, what is displayed on the voice assistant's screen, and participants' audio-prosodic features, we can infer even more.

For example, leaning forward while directing their gaze at the voice assistant (t=1s–3s) can be used to infer that the participant is engaging with Alexa. The tilting from side to side while the voice assistant is speaking (t=9s–17s) can be used to infer that the participant is listening. The nod and eyebrow raise at second 20, four seconds after Alexa finished speaking, can be used to infer that something went well. Directing the gaze at the voice assistant's screen can be used to infer that the participant is reading content on the screen (t=21s–24s), especially if side-to-side eye movement is also detected. The content on the screen can be used to infer what the participant might refer to in their potentially upcoming utterance. The laughing before directing their gaze downwards and then to the left (t=28s–32s) can be used to infer that something went wrong. The rising intonation (t=24s–26s) can be used to infer that a question was asked. And so on. These second-by-second interactions are currently not taken into account by the voice assistant, resulting in numerous interaction problems that we discuss below.

Interaction attempts. Despite multiple cues from participants that signal interaction attempts, these attempts were frequently missed by the voice assistant, which is programmed to respond only when it has heard its wake word or is engaged in multi-turn interactions (e.g., while using the Trivia voice app). Participants frequently did not use the wake word (Alexa) adequately. The only way to appropriately wake the voice assistant with a voice command is by pronouncing the wake word in a specific way and before saying the request. The wake word was either mispronounced or omitted in 70% of the Failure chunks, more than a quarter of all interactions. A few alternative pronunciations of the wake word were used, including: *"Alexia"* (n=2), *"Alexis"* (n=2), *"Alessa"* (n=3), and *"Alexi"* (n=2); note, these are all from different participants, except one who called Alexa both, *"Alexia"* and *"Alexi"*. Another reason for failing to wake the voice assistant was not saying the wake word before the request—seven participants said the wake word last at least once, and 13 omitted saying the wake word when initiating an interaction at least once. These do not include omissions that were not clear failures, such as Ambiguous "thank you" chunks. These interactions might be improved via more intuitive ways to wake the voice assistant and understand when it is paying attention.

Interactions with its other modalities. Voice assistants with screens, such as the Amazon Echo Show used for this study, have displays that may provide suggestions for how to interact with the voice assistant. Many participants used the prompts displayed on the screen (generic prompts generated by the voice assistant, not the research signs we posted on the wall) to interact with the voice assistant, but not everyone understood that these were mere suggestions. By looking at the participants' gaze and side-to-side eye moment, a human can tell that the participant is reading prompts on the screen. However, Alexa did not do this. Furthermore, Alexa's responses were also agnostic to what it was showing on its display. For example, if a participant asked for a recipe based on what the voice assistant was showing on its screen, Alexa started offering new recipe options for that type of food, instead of showing the specific recipe that was requested in response to the display's content. P15 fell into a repetitive loop, treating the suggestions as instructions. These

Time	Chunk	Posture	Gaze	Facial Expression	Human words	Machine words	Tone	Intonation	Rhythm
0:00		Neutral	Towards Alexa						
0:01		Forward	Towards Alexa						
0:02		Forward	Towards Alexa		Alexa what's the		Neutral	Same	Just Right
0:03		Forward	Towards Alexa		weather outside?				
0:04		Forward	Right						
0:05		Forward	Right			In New			
0:06		Forward	Towards Alexa			York it's fifty			
0:07		Forward	Towards Alexa			five degrees			
0:08		Neutral	Towards Alexa			Farenheit with			
0:09		Tilted left	Towards Alexa			showers			
0:10	18 (Success)	Neutral	Towards Alexa			Today you can			
0:11		Tilted right	Towards Alexa			look for clouds and			
0:12		Tilted left	Towards Alexa			showers with			
0:13		Tilted left	Right			a high of sixty			
0:14		Neutral	Towards Alexa			four degrees and			
0:15		Tilted right	Towards Alexa			a low of fifty two			
0:16		Neutral	Towards Alexa			degrees			
0:17		Tilted left	Towards Alexa						
0:18		Neutral	Towards Alexa						
0:19		Tilted right	Towards Alexa						
0:20		Neutral	Towards Alexa	Nods & raises eyebrows					
0:21		Neutral	Towards Alexa	Reads screen					
0:22		Neutral	Towards Alexa	Reads screen					
0:23		Neutral	Towards Alexa	Reads screen					
0:24		Neutral	Towards Alexa	Reads screen	What's				
0:25		Neutral	Towards Alexa		the weather like in		Neutral	Rise	Just Right
0:26	19 (Failure)	Neutral	Towards Alexa		Paris				
0:27		Neutral	Towards Alexa						
0:28		Neutral	Right	Laughs					
0:29		Forward	Down						
0:30		Forward	Down						
0:31		Forward	Down						
0:32		Neutral	Left						

Figure 2: Annotated events over a period of 32 seconds (two chunks). A large part of the interaction richness, that could theoretically be made available and interpreted by the machine using existing computer vision and/or sensing technology, is unavailable to the the machine. The only part that is available and interpreted here is the text highlighted in yellow, "Alexa what's the weather outside?"

findings suggest opportunities to better establish conversational grounding by connecting interactions to displayed content. In her post-interaction interview, P15 said how displeased she was with the voice assistant and said she would never use one, also demonstrating the importance of these initial interactions for adoption.

The presence of more than one user. We found that participants interacting in pairs sometimes reacted, or were influenced, by each other's interactions. For example, P14 & P26 were a couple interacting together. In one of the initial interactions P14 greeted Alexa. Then P26 greeted it too, this time introducing herself, and P14 briefly directed his gaze at P26 while she spoke. P14 then introduced himself as well, as if copying P26. The introduction prompted Alexa to start a voice training to learn an individual person's voice. However, P14 & P26, who were interacting as a pair, responded to Alexa's commands in unison, undermining the purpose of the voice training. As another example, P5 & P6 (who were not first-time users) engaged in a game of trivia, and had to devise silent strategies to communicate with each other about which answer to select to avoid Alexa prematurely recording a response. This made the interaction burdensome. Alexa missed important interaction data: the number of people interacting with it and their interactions with each other. If it had not overlooked this data, it might specify who a request is directed at, avoiding confusion, or know that an utterance is not directed at it. Moreover, if an issue continued, for example if users continued to respond in unison during the voice training, Alexa could interrupt to repair the issue, such as by explaining why it is better for only one person to respond at a time.

Social norms. We also saw cases where the voice assistant did not respect social norms at play. For example, P26 introduced herself to ask for the temperature, *"Alexa, my name is [P26], and I want to ask you how the temperature will be today"* (P26). Alexa was not able to interpret P26's introduction as a formality preceding an actual request for the weather. Instead, Alexa interrupted P26 after she said her name and in the middle of the request for the temperature, setting off a voice-recognition training.

4.1.2 Overlooked by humans. This section describes information that was overlooked by participants.

The voice assistant's indication that it was not actively listening. Alexa's blue line, which is used to signal different states (such as listening, processing a request, responding, or not actively listening) via its light and color patterns[10], did not appear to be an intuitive indicator of the voice assistant's conversational state for older adults; many did not even notice it. Participants often talked to Alexa when it was not displaying the blue line that indicates it is actively listening. For example, if they had a follow-up question, they would just ask it without "waking" Alexa and waiting for visual confirmation that it was listening. For example, P13 asked Alexa, *"Alexa, tell me about exercises for back pain?"* Alexa responded. Then, without making sure Alexa was listening, the participant asked *"How about specifically for older adults?"* and Alexa did not respond.

Moreover, some participants talked to Alexa at length without ensuring Alexa was listening. For example, P22 relayed in one conversational turn,

> *"My name is P22, how are you doing over there? All I can say is this mask doesn't save me. What I s[sic], what I do, I can't breathe. I like to take it off most of the time because I can't breathe. I'm almost 65. I'm retired. I'm happy. That's all I can say. I'm happy. I'm retired. I should have retired earlier than 65. Having a great time here. Retiring is great. So all who don't enjoy, sorry. I'm enjoying retirement over 65. Who cares about Coronavirus? Doesn't bother me. I use my vitamins, my juice, all the vitamins, juice and healthy food as long as they're available at nice healthy restaurants. Have a great day. God bless!"* (P22)

P22 did not establish conversational grounding with Alexa, probably because he did not know to look for the blue line indicator. Alexa, in return, did not respond to P22.

Technical terminology. The use of technical terminology likely contributed to many misunderstandings. For example, Alexa would say, *"Once I learn your voice, I'll be able to call you by name, tell you apart from others who use the Echo device you speak to, and personalize your experience. First, you'll need your own profile. I can create one for you now."* The way a machine learns a voice is different than the way humans do, so these differences must be described to someone who does not know how machine learning works. Moreover, the next part of Alexa's explanation is even more confusing to someone who is new to these types of technologies, *"As part of learning your voice, I'll ask you to say four phrases to create a voice profile. Your voice will be stored in the cloud until you delete it in the app."* Creating a voice profile, storing it in the cloud, and deleting it from the app are all explanations that assume technical familiarity. The communication gap this created was demonstrated by participants' expectations that Alexa would fulfill the requests they were making through the voice training, and by their actions, such as when they repeated an utterance that was not intended for the training (see the last paragraph of this section).

Requests to use other technology. Frustration and blank expressions were also common when Alexa required participants to know how to use other platforms. For example, Alexa made a suggestion to ask about Amazon orders, *"I didn't get that. By the way, there's lots more to discover. For example, I can keep you up to date on Amazon orders."* When the participant followed through and asked about her stuff, Alexa responded, *"I didn't find any open orders for Participant. If you're waiting for a delayed package, you can check the status at the orders page on Amazon."* Checking the orders page on Amazon is unfeasible for many who might be relying on a voice assistant as their gateway to the Internet. The participant's reaction was to laugh in dismay, and choose to end the activity (i.e., leave). In another example, Alexa once again asked a participant to try activities that they were unable to try without having access to a smartphone app, *"Okay, here's Activity Book. To use Activity Book a parent needs to give permission. To do that, I sent some information to the home screen of your Alexa app."* In this case, the participant had a blank expression and tried something new. In another, more navigable example, Alexa asked a participant to *"please select a default browser."* To know what a "default browser" is requires technical familiarity, but at least in this case there were only two options to pick from, meaning that there was a way to select a browser even without knowing the differences between the options.

That voice training activities were setup activities to make voice profiles. As a result of the unstructured nature of the study, three participants (P14, P15, P26) ended up completing a voice training. Alexa would launched the training when participants introduced themselves. During this activity, participants repeated Alexa's commands, but did not grasp that this was for training the voice assistant. Alexa took control of the interaction, and participants diligently followed Alexa's instructions. A couple participating in this task together did not question the activity (at least in front of us) but one participant did seem increasingly frustrated about Alexa continuing to ask her to say things. Her dismay was betrayed by her upset laughter and raised eyebrows, gazes towards us to request help, and confused expression. At the end of these interactions, Alexa offered advice about how to help others, "a*nd if you'd like to help others get recognized on these devices, remind them to say, 'Alexa, learn my voice.'*" In all instances (P15, P14 & P26), despite just having trained Alexa to recognize their voices, participants responded, *"Alexa, learn my voice"*, suggesting that they had not understood the purpose of the activity.

4.2 Body language

In most cases discussed above, there were visual cues in participants' body language available that helped us, the researchers, diagnose conversational problems that could use repair. For example, had Alexa seen P22 looking and talking at it, it could have responded to him after he said *"God bless!"* Similarly, had Alexa noticed P26 was not done speaking after introducing herself, it could have waited to

[10]https://www.cnet.com/how-to/what-do-the-light-ring-colors-on-your-amazon-echo-mean/

respond. In this section, we take a closer look at the body language that was expressed in these interactions.

Leaning forward, gaze, and nodding. There often were clear visual indications of when a person wanted to interact with the voice assistant, such as leaning forward and looking at the voice assistant, but the voice assistant overlooked them. Most participants (n=17) leaned forward at least once. When leaning forward, participants also directed their gaze at the voice assistant. In total, we identified 77 instances of participants leaning forward. Moreover, we noticed this behavior in a picture from a separate study with participants that seemed younger by Porcheron et al. [65], where a participant leans forward to speak to Siri on an iPad (see the bottom right picture in page 214 of their paper). We noticed that after a failed interaction, predominantly Alexa not responding, participants would lean forward, closer to the voice assistant. Nearly half the participants (n=12) leaned forward as a form of conversation repair. After a successful repair attempt via leaning forward, some participants would continue to lean forward in subsequent interactions (e.g., P4, P9, P13). Once Alexa responded, the tendency was to return to their initial position, and lean forward again for the next request. We also noticed instances of leaning forward in which no error had happened, suggesting leaning forward also occurred as a form of heightened engagement. For example, several participants (n=7) leaned forward towards the voice assistant when it was speaking, possibly to hear better. Similarly, some participants (n=13) leaned forward when initiating an interaction, possibly to ensure that the voice assistant could hear them or to signal that they were speaking to it. Another consistent interaction was nodding, which signaled that a positive interaction had occurred, suggesting either a pleasant surprise, being impressed, agreement, or affirmation. Given these findings, leaning forward while directing ones gaze at the device may be an important body language feature to recognize as an alternate form of "waking" voice assistants. In addition, nodding slightly and briefly could be used to automatically mark interactions as successful, to train voice assistants and to avoid repair [31].

Other forms of body language were somewhat ambiguous. We found that gestures such laughing, raising eyebrows, furrowing eyebrows, waving hands, and looking away could signal positive and negative interactions alike. The differences in the gestures themselves were too subtle, sometimes unnoticeable to us, to rely on them alone. For example, P15 laughs when she is caught in the loop of asking the same question over and over again due to thinking that the interaction suggestions were instructions, suggesting frustration. By contrast, P18 laughs when Alexa finally responds to her, suggesting relief. Alone, these reactions can perhaps be too difficult to interpret, but when more data is available, inferences can be made with more certainty, and their presence can signal an interaction event worth analyzing. For example, we can infer the valence of these actions from understanding interaction context—P15's misunderstanding of what is happening, and P18's previously unrecognized attempts—that was available to us, the researchers, but not interpreted by the voice assistant.

Tone	n	Success	Failure or Ambiguous
Exaggerated	8	71%	29%
Excited	10	26%	74%
Friendly	11	59%	41%
Indifferent	5	36%	64%
Nervous	8	22%	78%
Neutral	42	53%	47%
Tired	1	100%	0%
Upset	3	100%	0%
Intonation			
Fall	14	32%	68%
Rise	14	53%	47%
Rise-Fall	9	40%	60%
Same	51	56%	44%

Table 2: Outcome percentages of chunks by tone and intonation. We excluded all chunks in which the wake word was pronounced differently, omitted (note, in some chunks omitted wake word interactions were still successful as they were follow-up interactions), or said after the command. This table displays the resulting 88 chunks (48 successful ones). Note, percentages are not exact portions of the total counts as group sizes were adjusted to calculate them.

4.3 Audio prosodic features

In this section, we take a closer look at the rhythm, tone, and intonation in participants' speech patterns during their interactions with the voice assistant.

Rhythm. We found that the voice assistant often did not pay attention to a participant's speaking rhythm. For example, by the eleventh chunk in her interactions with the voice assistant and after having asked for the weather in multiple cities, P15 seemed exhausted, and took a deep breath in the middle of her request. Taking a deep breath slowed down the rhythm of her speech, *"Alexia what's the weather (deep breath)...in Paris?"* As she started saying where (per the suggestion on the screen), Alexa interrupted with the local weather. The voice assistant could detect a user's speech rhythm, and give room for pauses when needed.

Tone and intonation. To better understand how tone and intonation were affecting chunk outcome, we counted their occurrence as shown in Table 2. As we can observe in the table, some *tones* tended to result in Successful outcomes (Exaggerated, Friendly, Neutral, Tired, and Upset), and others in Failed or Ambiguous ones (Excited, Indifferent, and Nervous). Similarly some *intonations* tended to result in Successful outcomes (Rise, and Same, or Constant), and others Failed or Ambiguous ones (Fall, and Rise-Fall). Though this analysis is preliminary, it suggests that tone and intonation may give us more context about interactions. Taking these factors into consideration could also help voice assistants recognize errors and subsequently perform self-repair.

As can be seen in Table 2, Friendly tones were more likely to be missed by the voice assistant than Exaggerated or Upset tones. However, our participants were very hesitant to speak to it in an impolite manner. Because of this trend, when participants had multiple failed interactions attempts, we encouraged them to speak more sternly. Alexa often did not respond to their soft, friendly

tones. If failures continued, we suggested that participants imagine they were upset at Alexa, and speak in an upset tone. Often, once they started to speak to Alexa as though they were angry, Alexa finally responded (see Figure 3). When we suggested P20 speak to Alexa as though she was reprimanding Alexa, she responded, as she nervously prepared to try to speak more strongly, *"yo no hablo tan duro,"* which is Spanish for, *"I don't speak so strongly."* After four failed attempts, she asked to stop the activity without ever "waking" the voice assistant.

5 DISCUSSION

Our findings show how older adults who are novice users may interact with voice assistants in public settings. As such, our work contributes to a small, but growing, body of research that examines human-voice assistant interactions in the field [64, 65, 67]. Our inclusive design approach [28] may help guide future research on voice assistants that are more suitable for older adults and, as a result, for other users as well. Currently, most interactions with smart speaker-based voice assistants happen in the home, but in the near future, voice assistant technology will likely be pervasive in a variety of public contexts (perhaps airline check-in counters, medical facilities, or shopping centers) [73, 87]. Moreover, if purchasing a voice assistant from an electronics store, customers are likely going to interact with them, in a public setting, before deciding if they will buy the device. If the issues our findings surface are not addressed, we may be making voice assistants, and the promises they present, unapproachable to a large and important segment of the global population, hindering adoption, and creating systematic exclusion as voice assistants permeate public spaces.

We divide our discussion into design and research implications. In the design implications section, we 1) provide recommendations addressing interaction errors that resulted from not being able to successfully wake the voice assistant, 2) suggest ways in which automatic detection of non-verbal cues can be used to improve interactions with voice assistants, 3) emphasize differences and complexities for adapting voice assistants' interactions to older adults' needs and abilities in the context of prior research about code switching and knowing the user [13, 27], and 4) close by raising ethical design considerations. In the research implications section, we surface questions surrounding how we might use recent technological advancements to recognize body language and audio-prosodic features, and discuss the societal implications surrounding surveillance tradeoffs.

5.1 Design implications

Our findings have important implications for the design of more intuitive multi-modal, speech-first interfaces for older adults. Voice assistant design could rely on more familiar interaction paradigms, and/or responsibly capture and analyze data from multiple inputs, to create more natural conversations. In this section, we discuss recommendations for improving voice assistants for older adults and raise concerns regarding doing so ethically.

Improving interactions surrounding waking the voice assistant. Although we gave participants clear instructions on how to initiate interactions, waking the voice assistant was one of the biggest interaction problems we observed. Though we focused on older adults, this finding may also provide some context surrounding the large number of voice assistant interactions that were not successful, or were wake-word only commands, that Ammari et al. [10] identified in the usage logs of younger participants (18–64 years old). This said, Lee et al.'s findings suggest people's first words in an interaction with a robotic agent can predict their schematic orientation to an agent, making it possible to design agents that adapt to individuals during interaction [47]. Conversational errors that prevent interactions from occurring in the first place can thus hinder human-computer cooperation. This issue could be addressed in several ways:

- **New mechanisms to indicate when the voice assistant is not actively listening:** The interaction design of voice assistants could make it more clear to older adults when it is not listening, as our participants overlooked the blue line indicator; for example, by completely shutting off the screen, or having an avatar that looks away. Along the same lines, more consistency could be enforced for "waking" mechanisms, so that the design does not confuse users by sometimes requiring the wake word and sometimes not requiring it (e.g., during multi-turn interactions, such as Trivia).
- **Relying on familiar interaction paradigms:** Other mechanisms to wake the voice assistant could be put in place, such as using physical form-factors. Form factors that could be explored in future studies could be using a (possibly wearable) button that, when touched, would wake the voice assistant. Alternatively, picking up a telephone to talk to the voice assistant might provide a more familiar way to activate the voice assistant.
- **Responding to body language:** Voice assistants could "wake" when a person lean towards them, or showed other signs of engagement. Someone calling a voice assistant's attention by making a sound or motion while looking at it could also wake the device.

Providing friendlier explanations for people who are less familiar with technology. Participants who did not understand how the voice assistant worked did not understand that it was "learning their voice" from making them repeat phrases. More explanations could be added for people that are unfamiliar with this technology. For some, using this voice assistant is a big technological leap, and having it use terms such as "the cloud" and the "Alexa app" without offering additional explanation could be off-putting. Integrating these explanations into the design of the voice assistants could help older adults use voice assistants without the need for additional training from others. These design considerations could help increase digital empowerment for older adults.

Relying on automatic detection of visual and audio-prosodic cues. Voice assistants could be designed to appropriately react to visual and audio-prosodic cues, gaining social intelligence. Some of this is already happening [66]. Ideas for how this might take shape include:

- **Mirroring and understanding the user:** Voice assistants could try to mirror certain characteristics in their users, such as the speed at which they are speaking, to adapt to a user's needs and abilities. Similarly, echoing Nass's research, they

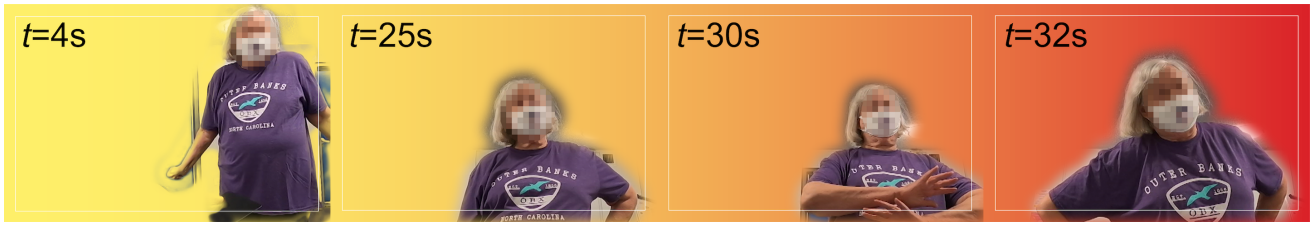

Figure 3: Four stills from a participant's video interacting with Alexa. The participant's tone is initially friendly and changes to upset throughout the interaction, indicated by the background color change from yellow to red. Alexa is unresponsive during the participant's "friendly" attempts, and responds when the participant uses an upset tone.

could mirror a person's mood [56] or tone, which could increase user satisfactions with the voice assistant. Additionally, voice assistants should be able to recognize different intonations, and use that information to respond appropriately (e.g., if the intonation conveys uncertainty, the voice assistant could reassure the user.)

- **Designing gaze intentionally:** Voice assistants with a movable screens[11] could be designed to make "eye contact" when addressed, or to turn their screens away when they are not paying attention. In prior work, McMillan et al. built a robot, Tama, that could detect the gaze of a user (instead of a wake word), and respond by moving an articulated "head" to achieve mutual gaze [54]. They found that gaze is a promising mechanism for augmenting or even replacing, the wake-work in initiating interaction with voice assistants [54].
- **Improving communication between different modalities:** Voice assistants could detect interactions with their different modalities, such as when users read the content being displayed on their screens, and consider that content in their responses. For example, users should be able to request more information about a recipe being displayed, such as by saying, "show me more details about the macaroni and cheese recipe you're displaying."
- **Differentiating between single-user and multi-user interactions:** Voice assistants could detect when there is more than one person involved in an interaction (e.g., by using voice recognition or computer vision), and adjust their reactions accordingly. For example, voice assistants could address users individually when needed, and react (or not) to users' interactions with each other when appropriate.

Adapting to users' needs and abilities. We also found that some participants said many words to the voice assistant during a single interaction, sometimes speaking for longer than the voice assistant could process. For example, P22, as described in our findings, told the assistant information about how he was doing, what he cared about, his views surrounding wearing masks, and more in only one conversational turn. In prior research, which did not include older adults, Beneteau et al. found that Alexa did not code switch with people of different ages to adapt its dialogue to the needs and abilities of the people it was interacting with [13]. They found that younger children tended to struggle more than older children and adults under 56, and provided an example of parents noticing their four-year-old would omit the wake word and get frustrated when the voice assistant did not respond back; the child would also use long sentences and often change topic before Alexa responded [13]. Beneteau et al.'s findings are echoed in the interactions we observed with older adults, where many older adults, similar to young children, addressed Alexa in the same way they would address a person. In response, voice assistants should be prepared to listen for longer to users who use more words in each conversational turn.

In this work, we find more evidence to support Beneteau et al.'s claim that "knowledge of the context and the communication partner is extremely helpful, allowing digital home assistants to artificially code switch as needed," and Clark et al.'s assertion that "there may be specific application areas where conversation may be appropriate if not essential between humans and agents, particularly in areas such as healthcare and wellbeing, where the nuances of contexts and demographics need to be considered" [27]. We contribute findings specific to the older adult demographic, which was not included in Beneteau et al.'s nor Clark et al.'s study [13]. Determining how to craft voice assistant dialogue for older adults would require further research, as it is unlikely that there will be a one-size-fit-all solution [32].

Ethical design. Despite their close resemblance to human voices, voice assistants are mere machines with many social deficits, making them unable to meet the expectations that they set. Our work highlights possible modifications to voice assistants—such as relying on visual cues to determine responses—that have the potential to make interactions more human-like. However, this comes with great responsibility. Human-likeness may affect our expectations of voice assistants [57], potentially increasing undue trust placed in them and encouraging stronger emotional connections. Design choices, such as using a robotic-sounding voice, may more accurately portray a voice assistant's true nature and prevent undue trust from being placed on it.

Feasibility. Our recommendations complement recent technological advances and work in progress. For example, in 2018, Kepuska and Bohouta [42] proposed developing a multi-input voice assistant that is able to interpret speech, video, images, and gestures from users. The system they proposed relies on piecing together various existing technologies, such as Kinects, cameras, APIs, and machine learning models [42]. More recently, Brunete et al. [19] developed a prototype for a robotic system to control a room that also relies on multiple inputs, including gaze, body language, and

[11] https://www.amazon.com/echo-show-10/dp/B07VHZ41L8

voice. Moreover, Nie et al. [59] recently developed a scheme to wake voice assistants without the need for a wake word by relying on other visual and audio cues. Large technology companies are also exploring how to include multi-channel inputs to improve human-machine conversations. For example, Amazon is using acoustic, linguistic, and visual cues to help Alexa interact more naturally [66]. Taken together, these advancements may make it possible to develop improved software agents.

5.2 Implications for voice assistant research

Through this research, we find that the most widely used research methodologies in the academic literature for studying older adult interactions with voice assistants (usage logs and interviews) are partial and incomplete, as many older adults are not able to even activate their voice assistants with consistent success. Because of this, analyses of usage logs collected in the privacy of the home may miss a large portion of failed interaction attempts. We therefore call for more research entailing interpreting body language and audio-prosodic features while honoring privacy expectations.

Interpreting body language. Video analysis revealed visual information that could be used to improve interactions. For example, posture shifts, such as leaning forward to be closer to the voice assistant, could be used for waking the voice assistant. In our research, posture shifts were an important component for indicating engagement. Additionally, following a participant's gaze was essential to understanding when a participant was reading or looking at something displayed on the voice assistant's screen, was distracted by something else occurring around them, or was requesting assistance. Given the advanced state-of-the-art of computer vision and sensing technologies [34, 36], it is important to study how these technological advancements may be used to recognize and interpret body language automatically in interactions with voice assistants.

Interpreting audio-prosodic features. Voice assistants tend to have human-sounding voices, and can be programmed to have prosodic variations. Alexa's friendly tone is a human-like conversational quality, which signals that it would be able to respond appropriately when spoken to as a human. However, in our research we noticed that Alexa's friendliness was a deceptive characteristic, at times, as Alexa had more difficulty responding to participants when they spoke to it in a friendly tone than when they approached it with an upset tone. Even though it was projecting friendliness, it did not understand friendliness when participants displayed it, resulting in inappropriate responses (or lack thereof). This calls for more research to interpret participant's audio-prosodic features, such as by using Amazon's Halo band that can measure tone of voice [3], to further understand how prosidic variations correlate to the voice assistant's responses.

Privacy considerations. Voice assistants that can "see" into our homes are already entering the market [37], but their societal implications are understudied. This is concerning because they could strengthen and continue to normalize technological surveillance [94]. It is important for interaction elements that could threaten our privacy to be considered and critiqued, as capturing and interpreting visual and audio-prosodic information requires potentially invasive data collection that comes with privacy and surveillance risks. Ensuring that computations happen on-device could be one way to limit the amount of data collected and stored. However, even then, having autonomous speakers with a camera consistently able to observe us could normalize surveillance by device and platform providers, as well as businesses, employers or remote family members. Concerns surrounding video surveillance of older adults are already being raised in the literature, and should be considered when adding mechanisms that could increase the risk of privacy violations [15, 25]. In addition, Bonilla and Martin-Hammond [16] found that knowledge of voice assistant privacy practices, data use and management are key concerns for older adults, and that many of their participants were unaware of existing resources available to mitigate such concerns. Future work is needed to explore not only the privacy and ethical implications of potentially intrusive technology, but also how vulnerable users may perceive and be affected by them.

5.3 Limitations

Our study has several limitations: it is a small scale, qualitative study conducted in an urban setting in the U.S. Moreover, we used a smart speaker-based voice assistant with a screen, so we do not know if our findings generalize to other voice assistants, such as screenless ones. Future research could investigate how these findings translate to voice assistants embodied in different devices. In addition, most participants were novice users of voice assistants, which may limit the generalizability of our findings. However, understanding novice user's interactions and struggles is necessary to promote adoption and prevent systematic exclusion of certain populations. Future research could explore interactions of users with varying levels of expertise and from different population segments. Participation was also limited to those who chose to participate; those that chose not to participate may have additional reasons for why they chose not to interact that our study did not surface. Another exciting area of future research would be to conduct video analyses of older adult interactions with voice assistants in different geographic locations and settings.

6 CONCLUSION

We used video analysis to characterize challenges with voice assistants' current design that may hinder older adults from benefiting from the promises the technology holds, or worse, exclude them from everyday activities as these technologies permeate public spaces. We described human-machine communication gaps revealed by our data, differentiating information that was overlooked by the machine (e.g., interaction attempts, the presence of more than one user) from information that was overlooked by participants (e.g., the blue line indicator, and technical terminology). We then examined body language features of the interactions and categorized them into those that provide reliable signals (e.g., leaning forward and gaze), and those that are somewhat ambiguous (e.g., laughing). Relatedly, we found that audio-prosodic features could also generate important information for reducing human-machine communication gaps, such as by identifying pauses from breathing or different tones and intonations. We discussed design implications for more intuitive interfaces for older adults, and conclude with a

call for more research entailing responsibly capturing and analyzing data from multiple inputs to create more natural conversations. Taken together, our findings help improve the inclusion of older adults in the design of voice assistants.

ACKNOWLEDGMENTS

This research was funded by NSF Awards #2026577 and #1700832. AC was additionally supported by a Digital Life Initiative Doctoral Fellowship. We sincerely thank all our study participants and our collaborators. Special thanks go out to Lisa Fernandez and the Roosevelt Island Senior Center, a Program of Carter Burden Network. We would also like to thank Cary Reid and Elaine Wethington for their mentorship and support, Kerstin Fischer and Qian Yang for discussing how to analyze video data with us, and Alexa Lempel, Jessie Taft, Jessica Bethune, and Paulina Cuadra for helping us proof-read earlier drafts of this paper.

REFERENCES

[1] [n.d.]. Alexa Connected Devices. Retrieved 2020-06-22 from https://developer.amazon.com/en-US/alexa/connected-devices.

[2] [n.d.]. Amazon Alexa Official Site: What Is Alexa? Retrieved 2020-06-22 from https://developer.amazon.com/en-US/alexa.

[3] [n.d.]. Amazon Halo - Health & wellness band. Retrieved 2021-01-27 from https://www.amazon.com/Amazon-Halo-Fitness-And-Health-Band/dp/B07QK955LS.

[4] [n.d.]. Google Assistant, your own personal Google. Retrieved 2020-06-22 from https://assistant.google.com/.

[5] [n.d.]. Google Nest, build your connected home - Google Store. Retrieved 2020-06-22 from https://store.google.com/us/category/connected_home.

[6] [n.d.]. iOS - Home - Apple. Retrieved 2020-06-22 from https://www.apple.com/ios/home/.

[7] [n.d.]. Nest Cam IQ indoor - Smart Security Camera - Google Store. Retrieved 2020-06-22 from https://store.google.com/us/product/nest_cam_iq.

[8] [n.d.]. Siri - Apple. Retrieved 2020-06-22 from https://www.apple.com/siri/.

[9] [n.d.]. Smart Video Calling with Alexa Built-in | Portal from Facebook. Retrieved 2020-06-22 from https://portal.facebook.com/.

[10] Tawfiq Ammari, Jofish Kaye, Janice Y Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput. Hum. Interact.* 26, 3 (2019), 17–1.

[11] Beverly Beisgen and Marilyn Kraitchman. 2003. *Senior centers: Opportunities for successful aging.* Springer Publishing Company.

[12] Genevieve Bell and Paul Dourish. 2007. Yesterday's tomorrows: notes on ubiquitous computing's dominant vision. *Personal and ubiquitous computing* 11, 2 (2007), 133–143.

[13] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[14] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24.

[15] Clara Berridge and Terrie Fox Wetle. 2020. Why older adults and their children disagree about in-home surveillance technology, sensors, and tracking. *The Gerontologist* 60, 5 (2020), 926–934.

[16] Karen Bonilla and Aqueasha Martin-Hammond. 2020. Older adults' perceptions of intelligent voice assistant privacy, transparency, and online privacy guidelines. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*.

[17] Robin Brewer. 2016. Connecting Older Adults through Voice-Based Interfaces. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion* (San Francisco, California, USA) *(CSCW '16 Companion)*. Association for Computing Machinery, New York, NY, USA, 131–134. https://doi.org/10.1145/2818052.2874350

[18] Robin Brewer. 2016. Connecting older adults through voice-based interfaces. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*. 131–134.

[19] Alberto Brunete, Ernesto Gambao, Miguel Hernando, and Raquel Cedazo. 2021. Smart Assistive Architecture for the Integration of IoT Devices, Robotic Systems, and Multimodal Interfaces in Healthcare Environments. *Sensors* 21, 6 (2021), 2212.

[20] US Census Bureau. 2018. Older People Projected to Outnumber Children. Retrieved 2021-09-01 from https://www.census.gov/newsroom/press-releases/2018/cb18-41-population-projections.html..

[21] Robin Caruso. 2019. CareMore Health: Addressing loneliness leads to lower rates of ED, hospital use. Retrieved 2021-02-04 from https://www.hfma.org/topics/operations-management/article/caremore-health--addressing-loneliness-leads-to-lower-rates-of-e.html..

[22] Justine Cassell, Yukiko I Nakano, Timothy W Bickmore, Candace L Sidner, and Charles Rich. 2001. Non-verbal cues for discourse structure. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. 114–123.

[23] Laura Castañón. 2020. The new fitness coach might not have a body - News @ Northeastern. Retrieved 2021-02-04 from https://news.northeastern.edu/2020/10/27/this-virtual-fitness-coach-could-be-just-as-effective-as-a-human-one/..

[24] Yuan Cheng, Yuchao Yang, Hai-Bao Chen, Ngai Wong, and Hao Yu. 2021. S3-Net: A Fast and Lightweight Video Scene Understanding Network by Single-Shot Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3329–3337.

[25] Jia-Luen Chua, Yoong Choon Chang, and Wee Keong Lim. 2015. A simple vision-based fall detection technique for indoor video surveillance. *Signal, Image and Video Processing* 9, 3 (2015), 623–633.

[26] Kyungjin Chung, Young Hoon Oh, and Da Young Ju. 2019. Elderly Users' Interaction with Conversational Agent. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 277–279.

[27] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What makes a good conversation? Challenges in designing truly conversational agents. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–12. https://doi.org/10.1145/3290605.3300705 arXiv:1901.06525

[28] P John Clarkson, Roger Coleman, Simeon Keates, and Cherie Lebbon. 2013. Inclusive design: Design for the whole population. (2013).

[29] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What can i help you with?" infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[30] Andy Crabtree, Steve Benford, Chris Greenhalgh, Paul Tennent, Matthew Chalmers, and Barry Brown. 2006. Supporting ethnographic studies of ubiquitous computing in the wild. In *Proceedings of the 6th conference on Designing Interactive systems*. 60–69.

[31] Andrea Cuadra, Shuran Li, Hansol Lee, Jason Cho, and Wendy Ju. 2021. My Bad! Repairing Intelligent Voice Assistant Errors Improves Interaction. In *Proceedings of the 2021 conference on Computer supported cooperative work*. ACM.

[32] Sara J Czaja, Walter R Boot, Neil Charness, and Wendy A Rogers. 2019. *Designing for older adults: Principles and creative human factors approaches*. CRC press.

[33] Philip R Doyle, Leigh Clark, and Benjamin R Cowan. 2021. What do we see in them? identifying dimensions of partner models for speech interfaces using a psycholexical approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[34] Sidney K D'mello and Arthur Graesser. 2010. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction* 20, 2 (2010), 147–187.

[35] Paul Ekman and Wallace V. Friesen. 2010. *The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding:*. De Gruyter Mouton, 57–106. https://doi.org/doi:10.1515/9783110880021.57

[36] Sean Andrist et al. 2018. Situated Interaction in the Open World: New Systems and Challenges. Retrieved 2021-01-27 from https://www.microsoft.com/en-us/research/uploads/prod/2018/09/2018-Andrist-Breakthroughs.pdf..

[37] Prakash Iyer. 2020. The science behind Echo Show 10. Retrieved 2021-02-08 from https://www.amazon.science/blog/the-science-behind-echo-show-10..

[38] Longlong Jing, Toufiq Parag, Zhe Wu, Yingli Tian, and Hongcheng Wang. 2021. VideoSSL: Semi-Supervised Learning for Video Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1110–1119.

[39] Ing-Marie Johnsson, Clifford Nass, Helen Harris, and Leila Takayama. 2005. Matching in-car voice with driver state: Impact on attitude and driving performance. (2005).

[40] Brigitte Jordan and Austin Henderson. 1995. Interaction analysis: Foundations and practice. *The journal of the learning sciences* 4, 1 (1995), 39–103.

[41] Malte F Jung, Jin Joo Lee, Nick DePalma, Sigurdur O Adalgeirsson, Pamela J Hinds, and Cynthia Breazeal. 2013. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 1555–1566.

[42] Veton Kepuska and Gamal Bohouta. 2018. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In *2018 IEEE 8th annual computing and communication workshop and conference (CCWC)*. IEEE, 99–103.

[43] Sunyoung Kim and Abhishek Choudhury. 2021. Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study. *Computers in Human Behavior* (2021), 106914.

[44] Dimosthenis Kontogiorgos, Andre Pereira, Boran Sahindal, Sanne van Waveren, and Joakim Gustafson. 2020. Behavioural responses to robot conversational failures. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 53–62.

[45] Małgorzata Kowalska, Aleksandra Gładyś, Barbara Kalańska-Łukasik, Monika Gruz-Kwapisz, Wojciech Wojakowski, and Tomasz Jadczyk. 2020. Readiness for Voice Technology in Patients With Cardiovascular Diseases: Cross-Sectional Study. *Journal of medical Internet research* 22, 12 (2020), e20456.

[46] Ewelina Lacka and Alain Chong. 2016. Usability perspective on social media sites' adoption in the B2B context. *Industrial Marketing Management* 54 (2016), 80–91.

[47] Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2010. Receptionist or information kiosk: how do people talk with a robot?. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 31–40.

[48] Yaniv Leviathan and Yossi Matias. 2018. Google Duplex: An AI system for accomplishing real-world tasks over the phone. (2018).

[49] Noah Liebman and Darren Gergle. 2016. It's (Not) Simply a Matter of Time: The Relationship Between CMC Cues and Interpersonal Affinity. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 570–581. https://doi.org/10.1145/2818048.2819945

[50] James Martin. 2010. What works and what doesn't inside Google's labs (photos) - CNET. Retrieved 2021-02-08 from https://www.cnet.com/pictures/what-works-and-what-doesnt-inside-googles-labs-photos/8/..

[51] Fabio Masina, Valeria Orso, Patrik Pluchino, Giulia Dainese, Stefania Volpato, Cristian Nelini, Daniela Mapelli, Anna Spagnolli, and Luciano Gamberini. 2020. Investigating the Accessibility of Voice Assistants With Impaired Users: Mixed Methods Study. *Journal of medical Internet research* 22, 9 (2020), e18431.

[52] Jim McCambridge, John Witton, and Diana R Elbourne. 2014. Systematic review of the Hawthorne effect: new concepts are needed to study research participation effects. *Journal of clinical epidemiology* 67, 3 (2014), 267–277.

[53] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[54] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: Tama–a Gaze Activated Smart-Speaker. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[55] Katashi Nagao and Akikazu Takeuchi. 1994. Speech dialogue with facial displays: Multimodal human-computer conversation. *arXiv preprint cmp-lg/9406002* (1994).

[56] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[57] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, MA.

[58] United Nations. 2019. World Population Ageing 2019. Retrieved 2022-01-20 from https://www.un.org/en/development/desa/population/publications/pdf/ageing/WorldPopulationAgeing2019-Highlights.pdf.

[59] Liqiang Nie, Mengzhao Jia, Xuemeng Song, Ganglu Wu, Harry Cheng, and Jian Gu. 2021. Multimodal Activation: Awakening Dialog Robots without Wake Words. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 491–500.

[60] Katherine O'Brien, Anna Liggett, Vanessa Ramirez-Zohfeld, Priya Sunkara, and Lee A Lindquist. 2020. Voice-Controlled Intelligent Personal Assistants to Support Aging in Place. *Journal of the American Geriatrics Society* 68, 1 (2020), 176–179.

[61] Young Hoon Oh, Kyungjin Chung, Da Young Ju, et al. 2020. Differences in Interactions with a Conversational Agent. *International Journal of Environmental Research and Public Health* 17, 9 (2020), 3189.

[62] Ray Oldenburg. 1989. *The great good place: Cafés, coffee shops, community centers, beauty parlors, general stores, bars, hangouts, and how they get you through the day*. Paragon House Publishers.

[63] Martin Porcheron, Joel E Fischer, Moira McGregor, Barry Brown, Ewa Luger, Heloisa Candello, and Kenton O'Hara. 2017. Talking with conversational agents in collaborative action. In *companion of the 2017 ACM conference on computer supported cooperative work and social computing*. 431–436.

[64] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

[65] Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2017. "Do Animals Have Accents?" Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 207–219.

[66] Arindam Mandal Pradeep Natarajan and Nikko Ström. 2020. New Alexa features: Natural turn-taking. Retrieved 2021-02-08 from https://www.amazon.science/blog/change-to-alexa-wake-word-process-adds-natural-turn-taking..

[67] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. "Phantom Friend" or "Just a Box with Information" Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[68] Alisha Pradhan, Amanda Lazar, and Leah Findlater. 2020. Use of intelligent voice assistants by older adults with low technology use. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–27.

[69] Alisha Pradhan, Kanika Mehta, and Leah Findlater. 2018. "Accessibility Came by Accident" Use of Voice-Controlled Intelligent Personal Assistants by People with Disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[70] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press.

[71] Sergio Sayago, Barbara Barbosa Neves, and Benjamin R Cowan. 2019. Voice assistants and older people: some open issues. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–3.

[72] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. "Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 857–868.

[73] Katie Seaborn, Norihisa P Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–43.

[74] Heung-Yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26.

[75] Jaisie Sin, Rachel L. Franz, Cosmin Munteanu, and Barbara Barbosa Neves. 2021. Digital Design Marginalization: New Perspectives on Designing Inclusive Interfaces. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 380, 11 pages. https://doi.org/10.1145/3411764.3445180

[76] George M Slavich, Sara Taylor, and Rosalind W Picard. 2019. Stress measurement using speech: Recent advancements, validation issues, and ethical and privacy considerations. *Stress* 22, 4 (2019), 408–413.

[77] Brodrick Stigall, Jenny Waycott, Steven Baker, and Kelly Caine. 2019. Older adults' perception and use of voice user interfaces: A preliminary review of the computing literature. *ACM International Conference Proceeding Series* (2019), 423–427. https://doi.org/10.1145/3369457.3369506

[78] Brodrick Stigall, Jenny Waycott, Steven Baker, and Kelly Caine. 2019. Older Adults' Perception and Use of Voice User Interfaces: A Preliminary Review of the Computing Literature. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 423–427.

[79] Erik Stone, Marjorie Skubic, Marilyn Rantz, Carmen Abbott, and Steve Miller. 2015. Average in-home gait speed: Investigation of a new metric for mobility and fall risk assessment of elders. *Gait & posture* 41, 1 (2015), 57–62.

[80] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 178–186.

[81] Lucy Suchman and Lucy A Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

[82] L Suchman and R Trigg. 1991. Understanding Practice: Video as a Medium for Reflection and Design. Design at Work: Cooperative Design of Computer Systems. M. Kyng.

[83] John C Tang. 1991. Findings from observational studies of collaborative work. *International Journal of Man-machine studies* 34, 2 (1991), 143–160.

[84] Margaret L Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A Christakis. 2020. Vulnerable robots positively shape human conversational dynamics in a human–robot team. *Proceedings of the National Academy of Sciences* 117, 12 (2020), 6370–6375.

[85] Milka Trajkova and Aqueasha Martin-Hammond. 2020. "Alexa is a Toy": Exploring Older Adults' Reasons for Using, Limiting, and Abandoning Echo. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[86] Jim Turner. 2020. Humana, Papa and Uber Health Join Industry Leaders to Remind Public They're "Far from Alone" in Combatting Loneliness and Social Isolation | Business Wire. Retrieved 2021-02-04 from https://www.businesswire.com/news/home/20200508005101/en/Humana-Papa-and-Uber-Health-Join-Industry-Leaders-to-Remind-Public-They%E2%80%99re-%E2%80%9CFar-from-Alone%E2%80%9D-in-Combatting-Loneliness-and-Social-Isolation..

[87] Joseph Turow. 2021. 3 An Operating System for Your Life. In *The Voice Catchers*. Yale University Press, 110–150.

[88] Akshith Ullal, Bo Yu Su, Moein Enayati, Marjorie Skubic, Laurel Despins, Mihail Popescu, and James Keller. 2020. Non-invasive monitoring of vital signs for older adults using recliner chairs. *Health and Technology* (2020), 1–16.

[89] Qifei Wang, Junjie Ke, Joshua Greaves, Grace Chu, Gabriel Bender, Luciano Sbaiz, Alec Go, Andrew Howard, Ming-Hsuan Yang, Jeff Gilbert, Peyman Milanfar, and Feng Yang. 2021. Multi-Path Neural Networks for On-Device Multi-Domain Visual Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3019–3028.

[90] Laurie Weingart, Philip Smith, and Mara Olekalns. 2004. Quantitative coding of negotiation behavior. *International negotiation* 9, 3 (2004), 441–456.
[91] Laurie R Weingart, Mara Olekalns, and Philip L Smith. 2004. Quantitative Coding of Negotiation Behavior. (2004), 441–455.
[92] Samuel Yang, Jennifer Lee, Emre Sezgin, Jeffrey Bridge, and Simon Lin. 2021. Clinical Advice by Voice Assistants on Postpartum Depression: Cross-Sectional Investigation Using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. *JMIR mHealth and uHealth* 9, 1 (2021), e24045.
[93] Po-Yin Yen and Suzanne Bakken. 2012. Review of health information technology usability study methodologies. *Journal of the American Medical Informatics Association* 19, 3 (2012), 413–422.
[94] Shoshana Zuboff. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power: Barack Obama's books of 2019.* Profile books.

A NUMBER OF INDIVIDUAL PARTICIPANTS EXHIBITING OBSERVED BEHAVIORS

Observation	No. of Participants (*N*=26)
Errors	
Omitted wake word when initiating an interaction	13
Mispronounced the wake word	8
Rhythm	
Did not pause	13
Paused for too long	5
Tone	
Neutral	19
Friendly	17
Upset	10
Excited	10
Nervous	9
Indifferent	9
Exaggerated	4
Tired	4
Intonation	
Constant	26
Fall-Rise	16
Rise	13
Fall	9
Body Language	
Leaned forward	17
Changed gaze to request input from others	15
Laughed	10
Raised eyebrows	9
Waved hand(s)	9
Looked away	8
Nodded	8
Furrowed eyebrows	7

Table 3: This table shows the number of individual participants (out of *N*=26) that displayed at least one instance of specific observations marked in our dataset.