What Could Possibly Go Wrong When Interacting with Proactive Smart Speakers? A Case Study Using an ESM Application

Jing Wei
jing.wei@student.unimelb.edu.au
University of Melbourne
Melbourne, Australia

Benjamin Tag
benjamin.tag@unimelb.edu.au
University of Melbourne
Melbourne, Australia

Johanne R. Trippas
johanne.trippas@unimelb.edu.au
University of Melbourne
Melbourne, Australia

Tilman Dingler
tilman.dingler@unimelb.edu.au
University of Melbourne
Melbourne, Australia

Vassilis Kostakos
vassilis.kostakos@unimelb.edu.au
University of Melbourne
Melbourne, Australia

ABSTRACT

Voice user interfaces (VUIs) have made their way into people's daily lives, from voice assistants to smart speakers. Although VUIs typically just react to direct user commands, increasingly, they incorporate elements of proactive behaviors. In particular, proactive smart speakers have the potential for many applications, ranging from healthcare to entertainment; however, their usability in everyday life is subject to interaction errors. To systematically investigate the nature of errors, we designed a voice-based Experience Sampling Method (ESM) application to run on proactive speakers. We captured 1,213 user interactions in a 3-week field deployment in 13 participants' homes. Through auxiliary audio recordings and logs, we identify substantial interaction errors and strategies that users apply to overcome those errors. We further analyze the interaction timings and provide insights into the time cost of errors. We find that, even for answering simple ESMs, interaction errors occur frequently and can hamper the usability of proactive speakers and user experience. Our work also identifies multiple facets of VUIs that can be improved in terms of the timing of speech.

CCS CONCEPTS

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

KEYWORDS

Voice user interface, smart speakers, voice assistants, Google Home, user experience, interaction error

ACM Reference Format:
Jing Wei, Benjamin Tag, Johanne R. Trippas, Tilman Dingler, and Vassilis Kostakos. 2022. What Could Possibly Go Wrong When Interacting with Proactive Smart Speakers? A Case Study Using an ESM Application. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3491102.3517432

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517432

1 INTRODUCTION

Voice user interfaces (VUIs) enable users to interact with computer systems through speech, facilitating touch-free and eyes-free interactions [22, 46]. Smart speakers are one of the major platforms that provide VUIs. It is estimated that the global smart speaker market will reach $19.91 billion in 2024 [3]. In addition to one-liner commands (e.g., "Hey Google, play music on Spotify"), various VUI applications are becoming available for the public [9], such as Amazon Alexa skills for exercise [1, 43] and Google actions for mental healthcare [2, 17]. VUIs that interact with users in multi-turn conversations are expected to proliferate in the future, supporting tasks such as booking a restaurant, searching for complex information needs, or providing personal health information [9, 10, 16, 52, 57, 67].

Recently, researchers have been considering ways to incorporate proactivity (i.e., acting in anticipation of future problems, needs, or changes[1]) into smart speakers [15, 35, 69]. Proactive speakers are envisioned to have the capability of initiating conversations and actively engaging users with speech [69]. Enabling smart speakers to be proactive can open up a wide range of applications, such as just-in-time health interventions [19], suggesting search results [4, 67], and capturing data through conversations [14, 28]. For example, a system called TandemTrack developed by Luo et al. [43] allows users to receive proactive voice reminders and self-report exercise data to Amazon Echo. However, through a field deployment, they identified that frequent voice recognition errors hampered the system's usability. While reporting data through voice was convenient, some users' self-reports could not be accurately recognized and could trigger random responses. Similarly, other studies also report frequent recognition and interaction errors of smart speakers [49, 57]. In particular, Kumar et al. [36] suggest that Alexa has an accuracy rate of only 68.9% on single-word recognition, and many transcription errors occur unpredictably (i.e., Alexa transcribes a distinct input word differently). This is concerning as frequent

[1] https://www.merriam-webster.com/dictionary/proactive

recognition and interaction errors may make users reluctant to explore new functionalities or even give up using VUIs [18, 42].

Previous studies have focused on interaction errors with VUIs [18, 36, 49]. For example, Myers et al. [49] developed a customized VUI calendar application that allowed users to add and delete calendar events via voice commands. They found that 52.1% of errors were natural language processing (NLP) errors. Cho and Rader [18] studied how novice users handled communication breakdowns with screenless Google Home. While many researchers conduct controlled studies [18, 30, 49], we have limited knowledge of what interaction errors could happen in the wild. Especially for proactive speakers, they are prone to inappropriately initiating conversations (e.g., noisy background [57], the user is not nearby [43]) as research on opportune moment predictions of VUIs is still at its infancy [15, 32, 33]. Both Alexa and Google Home have been found to perform poorly in a controlled environment [36, 71]. Proactive speakers may struggle even more to accurately perform speech recognition, the base of all voice interactions, in everyday life.

Meanwhile, some field user studies that implement VUI applications [28, 43] have observed interaction and transcription errors (e.g., music playing is triggered instead of the intended application) but failed to delve deeper into the root cause of errors and how users handle them. With the great potential of proactive speakers, investigating how they perform in the wild is critical. To systematically investigate interaction and transcription errors, we develop an interactive proactive speaker prototype based on Google Home and build a voice-based Experience Sampling Method (ESM) [11, 38] VUI application to run on the prototype. The ESM includes three 5-point numerical questions and one open-ended question. Our ESM, therefore, collects pre-defined and free-form answers, which allows us to quantifiably measure error rates of both kinds. Additionally, ESM represents a group of potential VUI applications, such as research data collection tools and self-tracking/reflection systems that can run on proactive speakers. The results of our study also help us understand the usability of VUI as a new modality for data capture in the wild and provide insight into future VUI applications.

Compared to previous studies on interaction errors and communication breakdowns [8, 30, 49, 57], our research aims to investigate interaction errors in the wild and quantify those errors through a case study of a voice ESM application. We deployed our proactive speaker prototype with the voice ESM application in 13 participants' homes for three weeks and collected over 1,000 audio recordings and 30,000 logs. We find that, although our proactive speaker only needs to recognize limited voice inputs (i.e., number 1 to 5), there exist substantial recognition errors that lead to non-smooth interactions with extended completion time and unexpected termination errors. The primary contributions of our work are: 1) extending prior work by quantifying different types of interaction errors and summarizing user strategies for resolving errors with the voice ESM application, and 2) synthesizing design implications for future proactive VUI applications and methodological insights into researching smart speakers in the wild.

2 RELATED WORK

In this section, we first review studies that implement proactive VUIs. Then we focus on existing work investigating VUI errors and user tactics or strategies to resolve errors. Lastly, we cover prior research on measuring interactions with smart speakers.

2.1 Proactive VUIs

Existing smart speakers can send voice reminders,[2] however, they only allow time-based scheduling of reminders. Poorly-timed voice reminders may be entirely missed [43] or may cause interruptions when users are engaged with a different task. Therefore, recent studies have tried to investigate which moments are suitable to initiate proactive conversations [15]. As there is no platform to deploy random (non-time-based) voice reminders, researchers have adopted laptops and smartphones to simulate speakers. Komori et al. [35] used a laptop to serve as a proactive VUI and a Kinect v2 to track the activity transition of users. They conducted a study with three participants who lived alone, and the proactive VUI system intermittently asked participants "Do you have a minute?" Participants were required to indicate their availability via finger gestures towards the camera. This way of interaction requires no audio input from participants. In a 2020 study by Cha et al. [15], a proactive speaker prototype was developed based on a combination of a smartphone and a commercial Bluetooth speaker. A voice-based ESM was scheduled and delivered by the smartphone and publicly displayed through the Bluetooth speaker. The system would ask participants "Is now a good time to talk?" and participants verbally provided their availability and the contextual reasoning. This system recorded user responses for one minute, and researchers later manually transcribed the audio data.

The systems mentioned above were developed to study the timing of proactive voice interruptions. Therefore, their interrupting tasks did not trigger multi-turn conversations between users and speakers. They also relied on post-hoc manual transcription to prevent interaction errors in-situ [15, 32, 33]. As we hypothesize that interaction errors might occur and extend interaction durations significantly [57], it is important to make proactive speakers interactive in real-time. From "fake" speakers to usable speakers in real life, we believe that proactive speaker prototypes should be built upon existing speech recognition technologies. Henceforth, we develop proactive speaker prototypes based on the popular off-the-shelf Google Home, which uses real-time speech recognition to interact with users.

2.2 Errors in VUIs

Users adopt various techniques for overcoming different errors and communication breakdowns with VUIs [8, 26, 49]. For instance, Jiang et al. [30] conducted a study on voice search where participants used the Google search app on an iPad to observe how the system actually transcribed their voice input. They found that when facing errors in recognized voice queries, users would reformulate their queries with phonetic and lexical changes based on the voice-to-text transcription they saw.

Besides techniques such as phonetic and lexical changes, other tactics such as hyperarticulation or simplification are observed [49]. For example, Myers et al. [49] developed a customized VUI calendar application to investigate how users overcome VUI problems. From their study, they suggested four major types of interaction errors:

[2] https://developer.amazon.com/en-US/docs/alexa/smapi/proactive-events-api.html

unfamiliar intent, NLP error, failed feedback, and system error. The NLP error is the most common type of error, and they identified that users adopted strategies such as hyperarticulation, simplification, or restarting. However, this study was a lab study. Furthermore, participants were provided with a GUI where they could see if errors occurred during the interaction.

VUIs deployed on touchscreen devices (e.g., smartphones) allow users to see on-screen clues to resolve interaction errors as voice-to-text transcription is often displayed [40, 57]. However, for screenless VUIs, such as most smart speakers, users can only rely on the VUI to remedy errors. In the work by Beneteau et al. [8], Alexa Echo Dots were deployed in 10 different family homes, and 59 conversational breakdowns were analyzed. The study found that family members used speech and language repair strategies, such as over-articulation, increased volume, and repetition, to recover from those conversational breakdowns. Family members also collaborated on repairing the conversations with Alexa. For example, family members expanded or shortened the voice commands spoken by another member that Alexa did not understand. Cho and Rader [18] found that screenless Google Home's fallback response "Sorry, I'm not sure how to help" did not help users reformulate new commands and resolve interaction errors. They recognized that new users of smart speakers might give up trying new commands and features if users constantly encounter interaction errors. Further, as suggested by Porcheron et al. [57], no-response from the speaker also indicates interaction failures, yet it provides no mechanism for further interaction. All these studies investigated errors to a broad extent: their participants explored smart speakers and encountered errors in various domains. It is difficult for researchers to fully understand why some errors occur and how to optimize the speaker or VUI applications to reduce the error rate. Through a systematic evaluation, Kumar et al. [36] find that homophones (e.g., "sail" as "sell"), compound words (e.g., "outdoors" as "out door"), and words with phonetic confusions (e.g., "wet" as "what") are consistently misrecognized by Alexa, while words with two or three syllables (e.g., "forecast") can almost be correctly recognized every time. This finding suggests that VUI developers can use more error-robust words when constructing dialogues and voice commands for their applications.

Compared to previous works on VUI errors, we aim to systematically measure error occurrence and delve into the root causes of interaction errors with an ESM application in realistic home settings.

2.3 Measuring Interactions with Smart Speakers

Interaction logs and audio recordings are frequently used to investigate smart speaker interactions [5, 9, 18]. For example, Bentley et al. [9] acquired Google Home interaction logs from their participants' personal Google accounts. The timestamps and command strings in these logs were used to categorize the application domains (e.g., information, music, home automation) and analyze temporal usage patterns of commands. Similarly, Ammari et al. [5] extracted command texts, timestamps, and device names from the interaction logs of 82 Amazon Alexa and 88 Google Home devices. They conducted both qualitative and quantitative analyses to understand how people use smart speakers at home. Analyzing logs is one way to study the conversations between users and smart speakers; however, its drawback is that fine-grained interaction details, such as voice volume and environment noises, cannot be fully learned through logs. If errors occur, logs may not help explain errors as users' speech may not be correctly transcribed [12].

Audio recordings provide additional user interaction data, such as environmental noise or speech loudness. These audios can often provide richer qualitative data to understand what is happening in the background when users are interacting with the smart speaker. Ideally, video capture can offer very rich data [23], but it can be too intrusive [66]. Porcheron et al. [57] raised the methodological issue of conducting studies on smart speakers in peoples' private homes. Henceforth, instead of using video, they collected user interaction data with Amazon Echo with a Conditional Voice Recorder. It is activated every time the wake-up word "Alexa" is detected and then captures a 1-minute audio recording. The audio recordings captured rich interaction data and allowed conversation analysis.

To conclude, popular commercial speakers, including Google Home and Amazon Echo, remain relatively closed-source [36]. Researchers need to either use real-time audio recording to capture interactions or rely on the voice history logs provided by the service providers (Google/Amazon) to refer to the interactions. Cho and Rader [18] captured both the participant speech and the Google Home transcriptions (from the logs) to investigate interactions. Drawn on the existing approaches, we capture both audio recordings and interaction logs of our custom action. We, therefore, have first-hand evidence of user behaviors when they face interaction errors and speaker event logs simultaneously. Our rich dataset allows us to examine and unveil causes of interaction errors, study user strategies, and perform temporal analysis of interactions with proactive speakers.

3 METHOD

We developed a proactive speaker[3] to explore how users overcome interaction errors in the wild. The proactive speaker was based on the off-the-shelf Google Home and bespoke hardware. We built a custom Google action called "Be Proactive" to implement a voice-based ESM that enquires users about their cognitive contexts. We can investigate how users interact with proactive speakers and explore the usability of proactive speakers in the wild with a field study.

3.1 Hardware Prototype

We developed an external apparatus to unnoticeably invoke a Google Home through the playback of pre-recorded voice commands, starting with "Hey Google, talk to Be Proactive". We attached a pair of earphones on two visible microphones onto Google Home, portsfitted through a 3D-printed "tiara" powered by a Raspberry Pi 3B+ as seen in Figure 1. We stored pre-recorded voice commands in the Raspberry Pi, and those commands were played whenever we wanted the speaker to be proactive. As earphones were very closely attached to Google Home, the voice commands

[3] No commercially available proactive speaker was available at the time of our investigation.

Figure 1: Our proactive speaker prototype.

were inaudible to nearby users. The net result is that Google Home begins talking to users without seemingly being invoked by them.

Pre-recorded voice invocation commands were issued by a Python script triggered on a semi-random schedule (the default was 9 AM to 10 PM, which participants could alter). The Raspberry Pi was also integrated with a light sensor and a USB microphone. To better understand the interaction between users and proactive speakers in the wild, we used the USB microphone to record the real-time interaction between users and the speaker. The audio recordings allowed us to study the in-depth voice interaction, including environmental contexts, rich accounts of errors, and user behaviors.

3.2 Be Proactive Development

Experience sampling is a widely used research methodology to collect user data, including people's thoughts, feelings, behaviors, and environments [11]. Conventional ESMs usually use the pen and paper technique or GUIs on smart devices [38, 68]. Proactive VUIs provide a different modality to initiate ESM enquiries and record user reports [15, 44]. Thus, ESMs promise to be a good VUI application to run on proactive speakers. We can investigate how users interact with proactive speakers and answer ESM questions through multiple conversation turns.

We used Dialogflow Essentials [4] to build Be Proactive. Once being invoked by the Raspberry Pi's script, it can proactively start a multi-turn voice-based ESM that enquires about user's availability, boredom, mood, and ongoing activity. These data are all commonly asked questions in other ESM studies [55, 56]. More specifically, the four questions of Be Proactive are:

> Question 1 - Rate your availability on a scale of 1 to 5.
> Question 2 - Rate your boredom level on a scale of 1 to 5.
> Question 3 - Rate your current mood on a scale of 1 to 5.
> Question 4 - What are you currently doing?

We designed the ESM questions to be easy to answer and analyze. For Questions 1 to 3, users need to respond with a number between 1 and 5. For those questions, only when a numerical answer is given and successfully recognized by the speaker, the next intent (question) will continue to be asked. In the meantime, user response and its timestamp will automatically be stored in Firebase Cloud storage. Close-ended questions with 5-point numerical scales allow us to conduct quantifiable analysis of recognition errors of numbers. For Question 4, users can respond with any free-form input. As long as the speaker recognizes *anything*, the intent will be completed, and the speaker will end the interaction with "Thank you for your time". Open-ended questions are commonly used in ESM studies to obtain rich data [58]. We can investigate whether Google Home can reliably transcribe longer sentences through the implementation of Question 4.

The invocation command for Be Proactive is "Hey Google talk to Be Proactive". To ensure successful external activation, one of our researchers pre-recorded the voice commands. The audio was stored in the Raspberry Pi to trigger Google Home to invoke Be Proactive at the scheduled time proactively.

3.3 Field Study

We advertised our study through our university platform. In total, we recruited 16 participants through online advertising, who were either full-time or part-time students. Three participants dropped out within the first few days due to technical reasons (e.g., frequent Wi-Fi disconnections) or task burden (e.g., interfering with daily study time). Hence, we deployed our proactive speakers in 13 participants' homes and collected data over three weeks. All recruited participants were proficient in English and had prior experience with Google speakers (Google Home or Google Mini) that ranged from 1 month to 3 years before the commencement of this study. Our participants were between 19 and 38 years old (M = 26.6, SD = 4.6), and the gender split was balanced (46.2% female, 53.8% male). Before the start of the experiment, we conducted a Zoom orientation session to teach participants how to set up the Raspberry Pi and instruct them how to answer the voice ESM. We scheduled the Raspberry Pi to trigger the Be Proactive ESM approximately every hour between 9 AM to 10 PM (with an 18-min randomized jitter). Participants could also choose to customize their prompt schedule to better suit their lifestyle. Four participants (participants P03, P08, P09, and P11) specified their preferred scheduling time of the speaker ESM prompts. For P08 and P09, the speakers only prompted from 10 AM to 10 PM and from 5 PM to 11 PM, respectively; for P03 and P11, the ESMs were triggered from 8 AM to 11 PM and 9 AM to 12 AM, respectively. The remaining participants received the ESMs according to our specified default schedule. Completing one ESM took about 40 seconds if no errors occurred. However, the answering time varied as users made mistakes or the speaker might have made recognition issues; therefore, we programmed the Raspberry Pi to record a 90-second audio snippet whenever it was scheduled to invoke Be Proactive. This audio recording behavior was clearly communicated to participants in the consent statement that they signed. At the end of the three-week study, we conducted an exit interview with participants over Zoom. All the participants were compensated with a $50 gift card for completing the whole study. Ethics approval was obtained from the university's human research ethics committee.

[4] https://dialogflow.cloud.google.com/

3.4 Data Analysis

In total, we received 1,213 ESM entries from 13 participants. Respectively, we have collected around 30 hours of audio recordings and roughly 30,000 interaction logs. Since this paper is about studying interaction errors, we will only focus on analyzing the rich interaction data provided by the audio recordings and logs and will not analyze the meaning of those ESM entries. To study the interaction between users and proactive speakers and investigate errors and their causes, we processed the data we collected in three steps: (1) transcription error identification and correction, (2) audio recording labeling, and (3) temporal analysis.

3.4.1 Transcription Error Identification and Correction. We first identify transcription errors in the collected data. The data is stored in-situ during the study in Firebase Cloud storage and is subsequently downloaded by the researchers. We label all user responses as errors that are not 1, 2, 3, 4, or 5 for availability, boredom and mood. We consider responses to the final question (What are you currently doing?) to be erroneous if they are inexplicable. For erroneous answers, we refer to the respective audio recordings and manually transcribe user responses. And we also try to find explanations for erroneous answers based on the rich audio data.

3.4.2 Audio Recording Labeling. By default, Google Home has two system errors: *no-input* errors and *no-match* errors. If the speaker does not receive a response, it issues the *no-input* error message - Sorry, I couldn't hear what you just said and repeats the question on a second attempt. The speaker may also issue the *no-match* error message - Please answer numbers, or you can come closer to me to ask users to respond correctly while not repeating the original question. This error message could either be triggered when the user response is not correct (e.g., when a user does not respond with a number between 1 and 5) or when it is not correctly transcribed. If no valid response is received after three attempts, the speaker will stop talking and abandon the survey.

When we were correcting erroneous transcriptions, we noticed that even for fully completed ESMs, many system errors occurred during the respective interactions. Considering the frequent occurrence of system errors and their high time cost, we then decided to manually listen to all 1,213 audio recordings to label how many attempts it took for the speaker to successfully receive an answer to each question. To be precise, if the user answers a question without triggering any recognition errors, we label this question as successfully answered with one attempt; if the user needs to re-answer a question by triggering one system error, we label this question as successfully answered with two attempts. In rare cases, when different error types happen consecutively in response to one question, the user can trigger error messages more than three times: they can try to answer a question at the third attempt and then trigger a recognition error again. Since this scenario rarely happens, we group those instances together with those where participants answered with three attempts.

During the manual inspection and labeling process of all audio recordings, we made observations of user strategies to overcome interaction errors. In particular, we examined user strategies that previous studies have reported, such as repetition, hyperarticulation, or simplification [29, 30, 49, 60]. We noticed that the strategy - new utterance/answer - was used by many participants when they were waiting for the speaker to respond or facing recognition errors. For other VUI applications [30, 58], that may be an effective way to resolve recognition errors; however, in the case of data collection, this behavior significantly lowers data quality. We, therefore, also labeled whether a participant changed their initial answer for each question.

3.4.3 Temporal Analysis. Many previous studies that implement custom VUI applications have not measured the detailed timing of interactions [28, 43]. Current smart speakers rely on the cloud to analyze user speech and generate corresponding responses, which can introduce latency of one second or longer [47, 57]. If recognition errors occur, users need to further repeat or reformulate their speech. With the network delays and error (and its recovery) time adding up, user burden can be significantly increased [13, 64]. Since we collect event logs generated during users' interactions with Bella Proactive, we extract timestamps from each log. We then conduct a temporal analysis of our data to quantify the precise timing of interactions.

As described in Section 3.2, every intent enabled with webhooks generates event logs on Firebase. Each intent during the interaction runs as a cloud function, and Firebase generates 6 event logs: (1) Function execution started, (2) Request, (3) Headers, (4) Conversation, (5) Response, and (6) Function execution finished. The *Response* log is generated to flag the speaker's response to user input (i.e., next question in our case). We can extract the timestamp of the *Response* log and consider this timestamp (CR_{A4_2}) to represent the time when the next question is issued by Firebase. When each answer is recorded in the Firebase Cloud storage, a corresponding timestamp (CR_{A4_2}) is also recorded. Therefore, we calculate the time difference between the two timestamps ($CR_{A4_2} - CR_{A4_1}$) and consider this time gap to represent the participant's answering time for each question. We denote the answering time for each question in Figure 2. Based on the labels we have, we can calculate the answering time for each question. The answering time includes: the time (i) for the speaker to receive a response from Google Cloud Function, (ii) it spends prompting the question, (iii) that the user answers the question, and (iv) to transcribe the answer by Google Assistant and log it on Firebase. If errors occur in the interaction, then the answering time will increase due to the error message and question re-prompting time. Combined with the labels of attempts (see Section 3.4.2), the answering time data enables us to understand how long users take to successfully finish each question (i.e., time on task) and if errors occur what the time costs of errors are.

As shown in Figure 2, we also denote the time gap between the answer recording time (CR_{A4_2}) and the next question/prompt response time ($CR_{A4_{2+1}}$). This time gap is generated by two sequential, although not directly causal, events that occur in the cloud during the interaction, which can indicate the time for backend processing and routing of webhooks. This metric can examine whether there are fluctuations in the backend system processing and network latency. We further infer the network and system delay by calculating the activation time of our proactive speakers. The Raspberry Pi generates a timestamp when it runs the activation job to trigger the speaker. Before the speaker is activated and issues Question 1 from the ESM, a Google Cloud Function will handle the request from

Figure 2: An example of the conversation flow between the Raspberry Pi, Google Home, and a participant. The solid-lined arrow denotes the time when Raspberry Pi plays the pre-recorded voice triggers. The double-lined arrow denotes the time (CR_{A4B}) when Google Cloud Function issues the ResponseLog to questions. The dotted-lined arrow (CR_{A42}) denotes the time when a user response is recorded in the Firebase Cloud storage. The circle denotes the point in time when the smart speaker starts prompting, for which we have no timestamp.

Google Assistant and generate timestamped logs on Firebase. We extract the timestamp of ResponseTime from all the logs and consider this timestamp to reflect the point in time when the Cloud Function finishes processing the intent matching and produces responses. Therefore, the time difference between the activation job start time and the Cloud Function Response time is a reflection of both network conditions and cloud service processing time. By analyzing these two temporal metrics, we gain insights into system delays that can actually impact user experience [59] but are invisible to both users and developers.

4 RESULTS

4.1 Data Overview

4.1.1 ESM Question Completion Rate. From a total of 3,447 issued ESM prompts, we collected 1,213 ESM responses. The response rate to ESM prompts in our study is 35.2%. Among those are 1,213 availability scales (100% completion rate), 1,110 boredom scales (91.5% completion rate), 1,041 mood scales (85.8% completion rate), and 1,036 engaged activities (85.4% completion rate) reported by participants. In the following section, we first present the results of the temporal analysis; then, we summarize interaction and transcription errors and their respective root causes, we present our observation of user strategies in the end.

4.1.2 Interaction Time. Analyzing the logs on Firebase, we first calculate the average answering time for each question of the ESM. As aforementioned, users need to answer the same question again if system errors occur. For each question, depending on how many attempts the speaker takes to record a valid answer, we calculate the average answering time based on the number of attempts respectively. The time needed to answer each question is shown in Table 1.

4.1.3 System Processing and Network Latency. During any interaction, Google Home needs to remotely recognize user inputs and generate responses [3], which can sometimes introduce invisible delays. We note that this type of delay has been mentioned in other work [57, 59, 70], yet it is rarely quantifiably measured [47, 50]. Although we cannot access the internal routing of Google Home and its cloud service, we infer the backend system processing and network latency from two metrics: 1) the time gap between the answer recording and the question issuing, and 2) the activation time of proactive speakers.

With four ESM questions and one ending phrase, four between-question time gaps are calculated based on the logging (see 3.4.3). For each question, the time gap distributions are presented in Figure 4 with most between-question delays below 0.25 seconds. A previous study suggests that a 1-second delay can indicate a problem in both face-to-face and telephone conversations [63]. Although conversations with speakers are slower than human-to-human conversations, an additional 1 second may also deteriorate the user experience. Therefore, we group delays longer than 1 second together. It is worthy to note that between the activity response recorded time and the ending phrase, there are 19 instances that are over 1 second. The distribution of the activation time is shown in Figure 3, which is a bimodal distribution, with one peak at around 4 seconds and another peak at around 7 seconds. Furthermore, while most activation time instances are below 10 seconds, several instances appear to be over 15 seconds ($n = 11$). We note that these values do not include interaction or recognition errors: they are just the time needed to invoke the Proactive. Our results suggest that for a small number of interactions, the invisible delays are not trivial.

Table 1: The total time (in seconds) taken by the speaker to recognize and record participants' answers for each question. The values are: mean (standard deviation, percentage of all responses).

Question	All	1 Attempt	2 Attempts	3 Attempts
Q1. Availability	12.65 (7.50, 100%)	9.71 (3.28, 80.19%)	21.53 (3.87, 15.17%)	37.07 (7.34, 4.02%)
Q2. Boredom	9.99 (6.45, 100%))	8.17 (2.4, 88.9%)	21.01 (5.64, 8.57%)	36.77 (9.62, 2.53%)
Q3. Mood	9.48 (5.45, 100%)	7.91 (2.03, 88.49%)	20.08 (5.1, 10.54%)	38.09 (12.45, 0.97%)
Q4. Activity	8.74 (4.77, 100%)	7.69 (2.49, 93.11%)	20.78 (2.42, 5.91%)	35.48 (1.37, 0.98%)

Figure 3: Activation time is the time needed for routing and processing on the cloud, before the speaker initiates a conversation.

Figure 4: Back-end routing and processing time needed between two subsequent questions. Delays longer than 1 second are grouped into one bin, denoted by the red bar.

4.2 Interaction Error Categories

When a user responds to the ESM via voice, interaction errors can happen. Since we have auxiliary audio recordings for every ESM answering session, we are able to investigate the reasons behind uncompleted ESM surveys or seemingly erroneous responses. In this section, we summarize di erent types of interaction errors that occurred in the eld study.

4.2.1 System Default Errors. We nd that the default no-match errors and no-input error are quite common in our dataset. Generally, if users do not answer the question prompted by the proactive speaker, the no-input error will be triggered; if users give non-number answers to the rst three questions of the ESM, the no-match error will be triggered. However, it is important to note that these errors are subject to what the Google Home actually hears . For example, if the environment is noisy, the speaker may hear nothing and produce the no-input error despite that the user actually answers correctly. Also, no-match error can be produced if the speaker falsely transcribes the user's numerical input as a non-numerical input (e.g., 2 as pool). In fact, given the auxiliary audio recordings, we nd that participants rarely gave false answers or did not respond to the speaker once they committed to the ESM. Yet, we discover a substantial amount of system errors. Among 1,036 fully answered ESMs (1,213 ESMs collected), only 62.8% of those ESMs were completed without triggering any system errors, 23.6% of them were answered with one error triggered, and the rest (13.6%) were all answered with two or more errors triggered. The percentage of successfully recorded answers with di erent attempts for each question is shown in Table 1. As can be seen, fewer answers are recorded with three attempts in later questions. For example, for Questions 3 and 4, less than 1% of answers were recorded with three attempts. Comparing the answering time with one attempt and two attempts, we can see that the ESM completion time will be increased by at least 10 seconds if one error occurs.

4.2.2 Interaction Termination Errors. As aforementioned, if a no-match or a no-input fallback error message is prompted three times, the speaker will play the nal error message Sorry, I can't help and terminate the interaction. In our study, such an early termination

of the interaction can result in incomplete ESM surveys. Since this type of early termination is caused by consecutive system errors, we categorize it as an Accumulative Termination (AT). Before inspecting audio recordings, we originally assumed most incomplete ESMs were prematurely terminated because of consecutive system errors. However, we find another type of error actually caused more early terminations of the ESM, which we refer it as Sudden Termination (ST). To answer the first three questions of the ESM survey, participants were required to give numerical answers. We observe that numbers 1, 2, 3, 4, 5 are sometimes wrongly recognized as several other words. For example, we find that 2 can be recognized as words with similar pronunciations, including pool, tattoo, true, 3 can be recognized as train, and 5 can be recognized as bye. By default, a wrongly recognized response (e.g., 2 recognized as pool) should trigger the no-match error; however, we notice that the wrongly recognized words (e.g., pool) would invoke Google's default Map searching function. Instead of triggering the fallback error message, the speaker terminates the ESM application and proceeds to announce nearby pool addresses. For the case of a wrongly recognized number 5 as bye, the end of conversation is directly triggered, the speaker subsequently says goodbye (or other farewell messages) to users, and the ESM is then terminated. Lastly, we find network or timeout issues during the cloud connection would also cause sudden termination of the ESM. Usually, the speaker would prompt a timeout error message "There was a glitch, try again in a few seconds" then stop the ESM entirely.

With our manual inspection of the audio recordings, we calculate the frequency of ATs and STs in our dataset. We find that there are 30 instances when participants attempted to answer the ESM (these instances were not included in Table 2), but the Proactive was prematurely ended due to ATs ($= 6$) and STs ($= 24$). In other words, our participants tried to answer more ESMs, yet they were stopped by termination errors, which could only be found by manually inspecting the audio recordings. Among 177 recorded yet incomplete ESM responses, 14.4% of terminations are caused by ATs, and 85.6% are caused by STs. The prevalence of STs is quite unexpected. For each participant, the termination error percentages are presented in Table 2. It is worthy to note that STs occur pretty often for almost half of the participants, and ATs occur relatively more frequently for P08 and P09. Additionally, we observe that many participants changed their initial answers during the course of responding to the proactive speakers. Therefore, we also include how often people changed their answer across multiple attempts at the same question in Table 2.

4.3 Data Entry Error Categories

Prior studies suggest that a successfully recorded answer may be incorrectly transcribed and need manual correction [43]. Here, we present the summary of transcription errors in our dataset.

4.3.1 Erroneous Numerical Answers. For the first three questions, we consider any recorded answers that are out of the range of 1 to 5 to be erroneous; for the last question, which accepts free-form text, we consider any illogical/inexplicable answers to be erroneous. We manually identify transcription errors from recorded answers and then use the auxiliary audio recordings as the ground truth to correct answers that contain errors. Furthermore, we infer potential causes that lead to transcription errors. Examples of erroneous transcriptions can be found in Table 3.

For numerical answers, the percentage of transcription errors is quite low: seven erroneous availability answers (0.2%), eight erroneous boredom answers (0.7%), and three erroneous mood answers (0.3%). We identify three types of transcription errors. The first type is double response, and this error is more common than other errors and occurs ten times in total. Double responses occur because participants repeat their answers while waiting for the next question. The end result is that the speaker captures both utterances and treats them as a single response. We notice that the time gap between two repeated answers ranges from 1 second to 7 seconds.

The second error type is wrong transcription, and it occurs three times. This is a situation where the participant's response is indeed valid, but the speaker's transcription is inaccurate. As shown in Table 3, one error instance is that one participant answered "ah... 4" and we assume that the speaker misinterpreted that as "one fourth" and recorded "0.25" on Firebase.

Lastly, the third error type is a false positive error. This error occurs when the speaker captures answers from a source other than the participant. For instance, in one case, the response was captured from a video that the participant was playing at the time when the question was asked.

4.3.2 Erroneous Open-ended Answers. Question 4 has the highest 1-attempt answer recording rate, but transcription errors are much more prevalent in its free-form answers. For a total of 1,036 recorded activity answers, we find that 24.7% need to be manually corrected. We also identify four types of transcription errors: partially missing (2.4%), partially incorrect (12.0%), totally incorrect (9.3%), and extra information (1.0%). For each type, one transcription error example is presented in Table 3. Ultimately, transcription errors are directly caused by incorrect speech recognition. However, partially missing and extra information errors can also be accounted for factors such as environmental noise, Google Home's timeout window (5 seconds) of listening for inputs, and participants trying to repeat their responses (to get their speech recognized). We originally assumed that VUIs are more advantageous in administering open-ended questions compared to GUIs [24] as speaking is faster and more effortless than typing [61]. However, as transcription errors are prevalent, researchers should weigh the benefits of using open-ended questions against the high cost of manual correction efforts. For different transcription errors, the recovery difficulty also varies. In our dataset, answers categorized as totally incorrect are unlikely to be recovered through solely manual speculation (e.g., "eating" transcribed as "Ethan"). Such correction must rely on using the auxiliary audio recording. For answers categorized as partially incorrect or partially missing (e.g., "I'm joking email" can be corrected as "I'm checking email"), it was still possible to recover those answers without referring to auxiliary audio recordings (assuming the examiner is familiar with speech recognition limitations, e.g., "joking" is phonetically similar to "checking"). Lastly, we find that while longer free-form answers are more likely to contain transcription errors, they are easier to recover as they usually have more semantic contexts [7].

Table 2: The frequency of different types of error occurrence in logged 1,213 ESMs for each participant. AT: Accumulative Termination; SE: Sudden Termination; Changes: responses where answers where changed between attempts.

Participant ID	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	P13	Total
ESM responses	153	113	113	88	62	88	153	26	39	60	28	200	90	1213
AT (%)	3.9	0	0	1.1	3.2	5.7	1.3	7.7	12.8	0	0	1.0	1.1	2.1
ST (%)	1.3	2.7	5.3	6.8	8.1	26.1	0.7	23.1	25.6	43.3	7.1	17.5	26.7	12.3
Changes (%)	11.8	0.9	6.2	2.3	3.2	17.0	3.3	11.5	15.4	3.3	3.6	22.5	3.3	9.1

Table 3: Transcription error examples.

Error Type	Real Answer	Transcribed Answer
Numeric Response: Double response	2 (3-sec gap) 2	22
Numeric Response: Wrong transcription	ah. . . 4	0.25
Numeric Response: False positive error	(noise)	0
Open-Ended Response: Totally incorrect	have some meeting	zombie eating
Open-Ended Response: Partially missing	I'm reading some essays	reading some
Open-Ended Response: Partially incorrect	watering my plants	watering my friends
Open-Ended Response: Extra information	watching videos	watching watching videos

4.4 User Strategies

As aforementioned, 14.6% of the collected ESM responses are incomplete due to premature termination, and among complete ESM responses, 37.2% of those ESM interactions encountered at least one system error. As the frequency of interaction error occurrences is not trivial, we have observed and categorized four strategies adopted by our participants to resolve errors through manual inspection of the auxiliary audio recordings: (i) raising voice and approaching, (ii) repeating, (iii) phonetic and lexical changes, and (iv) help from others. Some of our participants' exit interview data further confirm their use of these strategies. We show the breakdown of user strategies adopted by different participants in Table 4.

Table 4: User strategies adopted by different participants.

Strategy	Participants
Raising Voice and Approaching	P01, P02, P07, P09, P12, P13
Repetition	P01, P02, P03, P04, P05, P07, P12, P13
Lexical Changes (Close-ended)	P01, P03, P04, P05, P06, P07, P08, P09, P10, P11, P12
Lexical Changes (Open-ended)	P01, P02, P03, P07, P08, P09, P13
Help from Others	P08, P09, P10, P13

4.4.1 Raising Voice and Approaching. By inspecting the auxiliary audio recordings, we are able to hear users' voices and responses. We find that when facing recognition errors or delays in response from the speaker, among 13 participants, 6 participants used raising voice and approaching to resolve interaction errors. Usually, when participants' first response was not recognized and triggered a *no-match*/*no-input* error, they would raise their voice in each subsequent attempt [8]. Similar to raising voice, another commonly used strategy is approaching the speaker or looking at the speaker [37, 39, 54]. While it remained unknown where our participants looked when interacting, the USB microphone captured (seemingly) door opening/closing sound and footstep sound. We, therefore, infer that some participants would enter the room where they installed the proactive speaker to answer the ESM better. Alternatively, they would try to answer the ESM initially in-situ and get closer to the speaker if encountering recognition errors during the interaction. In one audio recording of P02, the participant initially answered the last question and headed out of the bedroom (where he put the proactive speaker). However, as he did not receive the ending prompt after going out, he returned to the bedroom to repeat his answer and then left the bedroom the second time. Previous studies have suggested that this kind of user strategy is the most common way to resolve errors [42, 49]. However, less than half of our participants were found to use these two strategies. One reason could be that, compared to controlled lab studies [49], the interaction with proactive speakers in the wild can occur when users are not nearby. Users may raise their voices when responding, but such behavior could fail to be captured by the USB microphone due to the distance or background noise.

4.4.2 Repetition. Previous studies suggest that repetition is used when users sense extended delays when waiting for voice agents' responses [14, 57]. Such strategy - repetition - is also observed in our audio recordings. In total, we find 8 participants having the habit of repeating their responses when facing silence from the speaker. While some participants could patiently wait until the speaker prompts the next question, others tended to repeat their answers if they received no quick response from the speaker. This may suggest that participants have different sensitivities to delays, and people who like to repeat may treat the silence as an indication of interaction failures [57, 58]. Past works suggest that repetition is usually accompanied with rephrasing as users may assume the delay is a sign of incorrect requests or the *no-match* error [42, 49]. In our case, participants were well aware of the requests they

can use (at least for the numerical scale questions), and hence the no-match error is not possible. Thus, the repetition can suggest that participants treat the long no-response delay as an indication of a no-input error, i.e., their responses are not "heard" by the speaker. However, from the data collection point of view, this strategy can give a rise to more double-response transcription errors of numbers (= = 11) (see Table 3).

4.4.3 Phonetic and Lexical Changes. Previous work has found that users would reformulate their voice query both lexically and phonetically when responding to input errors [30, 42]. In our study, we also observe participants adjusting their answers when facing system errors. For numerical answers, two phonetic changes are observed from the auxiliary audio recordings: slow down and change pronunciation. Such phonetic changes are further con rmed by the exit interview. During the interview, a few participants mentioned that the speaker had an issue recognizing their answers, and they had to adopt a more standard pronunciation of English.

Other than adjusting the pronunciation, the audio recording allows us to nd that some participants even changed their initial answers when receiving no response from the speaker or encountering default system errors. As can been seen in Table 2, changing answers is quite prevalent. In total, 110 ESMs contain changed answers, which is 9.1% of the entire dataset. For the three 5-point close-ended questions, all participants, except P02, changed their initial answers during the course of responding. Especially for participants P01, P06, and P12, their ESM responses include many recorded answers that di er from their initial answers. This nding again echos prior work that demonstrates how users would reformulate requests to get the speaker to work [57]; however, in the context of voice ESM, the reformulation equals to changed answers. One interesting instance is from an audio clip of P12: when facing no response from the speaker, this participant tried to repeat his initial answer rst, and later literally uttered all the numbers (1, 2, 3, 4, 5) to get the speaker to recognize his answer. In other words, the participant's main goal is to get one recognizable answer recorded rather than giving the genuine answer.

For the open-ended Question 4, participants can give any free-form answers. Therefore, as long as the no-input error is not triggered, the speaker can record participants' answers. While there was more room for participants to reformulate their answers, very few recorded Q4 answers were changed answers. For a total of 1,036 recorded activity answers, our manual inspection nds that 18 recorded answers (1.7%) di er from participants' initial answers, and 7 participants have used lexical change as a way to resolve interaction errors. Based on lexical query reformulation patterns pointed out in previous studies and our own observations, we summarize four reformulation patterns: addition [30], removal [30], word substitution [29, 65], and total change. Compared to previous studies, the only new reformulation is the total change, which refers to giving an entirely new answer on the second or third attempt. This reformulation, similar to the change of numerical answers, while it resolves the interaction error, can negatively impact the data quality. We exemplify the lexical reformulations of Q4 answers in Table 5. In contrast to participants who shortened their speech to resolve errors in other studies [31, 54], our participants sometimes expanded their responses to Question 4.

4.4.4 Help from Others. One unexpected strategy is found to be the help from others. Besides three participants, all others live with cohabitants (family members or roommates) at home. In particular, the smart speaker was shared among the family for some participants. Our proactive speaker cannot perform person identi cation, and anyone can answer the ESM survey. When manually inspecting audio recordings, we notice that some ESM surveys were (partially) completed by the participant's partner or child. We also observe that this "replacement" in answering questions happened in di erent ways. For example, if the participant was not available for the ESM, their partner sometimes answered the entire ESM. Alternatively, the participant could answer the ESM initially, and if system errors occurred during the ESM session, their partners sometimes would try to answer the ESM with a louder voice or a di erent tone. Existing literature also describes similar collaborative e orts in rephrasing voice requests and commands to resolve interaction errors [8, 57, 58].

5 DISCUSSION

We agree with previous studies [43, 44] that smart speakers can be a promising new platform that provides a new modality (voice) to capture data. After all, our proactive speakers have successfully collected 1,036 complete ESMs from 13 participants under varying environmental conditions. However, our results suggest that collecting data through smart speakers can be challenging as the interactions are sometimes non-smooth and overly extended due to interaction errors. Also, the data quality can be compromised due to interaction errors (e.g., transcription errors) and the resulting user strategies. In the following, we discuss our nding, focusing on both technical and social aspects. We also discuss implications for future VUI application designs and research around smart speakers.

5.1 Technical Issues

Currently, smart speakers' speech recognition and response generation are all processed over the cloud. Therefore, network latency can happen [47] and sometimes introduce signi cant delays in interactions [57]. Timing is vital in the interactions with smart speakers [6], and the "silence" of speakers is considered as an indication of errors [25, 57]. In our study, we show that the activation time of "Be Proactive" is a bimodal distribution with one peak at around 4 seconds and another peak at around 7 seconds (see Figure 3). We speculate that the bimodal distribution is caused by the user input being sent to servers in two di erent locations. To end-users, the extra 3 seconds needed to activate a custom VUI application can lead to confusion and impatience [48]. Further, if the network delay happens during an interaction, the smart speaker's "silence" can result in some users continuously repeating an answer. Repetition can further cause the speaker to continue listening to the new input and sending this new information to the cloud for processing, which generally worsens the delay. Therefore, to reduce latency, it seems important to deploy the VUI application server close to the location of the targeted users [48].

Before manually inspecting audio data, we assumed incomplete ESMs were caused by consecutive recognition errors, i.e., AT. However, we also identi ed another termination error - ST, caused by

Table 5: Lexical reformulation observed in Q4 answers.

Strategy	Original Answer	Changed Answer	Explanation
Addition	studying	studying for tomorrow's exam	added more details
Removal	studying and listening to music	studying	removed listening to music
Word substitution	eating	having snacks	expanded the phrase
Total change	studying	chatting with friends	changed the entire activity

the erroneous triggering of the map search function or other applications. For example, when 2 is recognized as pool , instead of prompting the no-match error message, the speaker would announce nearby pool locations. It appeared that the map search function (and other application activations) were treated as a global intent phrase by Google Home (i.e., it can be triggered at any point of a conversation with the speaker), which overrode the no-match response. Interestingly, STs occurred much more frequently than ATs. We nd that the amount of STs increased for later participants. In this study, P01 was the rst to participate from late September to mid-October 2020, and for the remaining 12 participants, three groups of four people participated in similar overlapping time periods, e.g., P02 to P05 participated from late October to mid-November. We assume that Google updated their algorithms during our study and the new intent matching mechanism caused the high occurrence of STs. Unfortunately, Google's algorithms and their updates are invisible to us [18, 57]. One way to mitigate the impact of such errors is to improve the design of VUI applications by selecting more error-robust words and phrases, which will be discussed later.

5.2 Interaction Errors and User Behaviors

In our dataset, we collected 1,213 ESMs, and 14.4% of them were actually incomplete due to ATs and STs. Even for 1,036 complete ESMs, only 62.8% of them were completed without any interaction errors. In Table 2, we show the occurrence of STs and ATs for each participant. Overall, ATs occur infrequently (2.1%). Very few ESMs are incomplete due to system errors being consecutively triggered more than three times. On the other hand, STs appear to be more frequent (12.3%). Theoretically, STs should never occur as the Fallback intent would be triggered if a non-number is recognized or given by the user for Q1 to Q3. Since ATs were caused by accumulative incorrect speech recognition across attempts, more frequent STs could stop ATs from happening. In other words, people who are more likely to trigger recognition errors should, in theory, encounter more ATs; however, the accumulative recognition errors are prevented due to STs. Consistent with previous studies [28, 43], our participants reported that such unexpected app triggering and termination were very confusing.

Both termination errors and recognition errors can impact participants' answer quality. For example, a few participants reported that they noticed the speaker had trouble recognizing 2 and, therefore, they tried to avoid answering 2 afterwards during the exit interview. Also, when facing in-situ recognition errors, many participants used the changing answers strategy and the help from others strategy to resolve errors. However, those strategies can imperceptibly lower data quality. For the former one, users are trying to give a recognizable" rather than a valid" answer during the interaction [21]; for the latter one, the recorded answers do not originate from targeted users. Our ndings are in line with previous studies that suggest frequent interaction failures can alter users' ways of interacting with VUIs and discourage them from using VUIs [18, 42, 43].

In Table 2, we also present the ESM response counts and changed answer rate for each participant. While we do believe that recognition errors can drive users to change answers when facing errors [49] or give up interactions [42], we fail to observe any obvious correlations between termination errors and user response counts and changed answer rates. We requested participants to answer ESMs whenever they could. So, it is possible that even if participants encountered many interaction errors, they still tried to answer the ESMs whenever they could as they were participating in a study. Another possibility is that after experiencing some recognition errors, some participants only gave answers that they considered were more easily recognized [9, 63] to avoid errors beforehand, but such behaviors may not be re ected in our short-term study. Additionally, the high percentage of STs stopped more ATs (consecutive and time-consuming recognition errors). Participants with higher ST rates may have not yet understood why such error occurred [43]. Therefore, they did not change answers to recognizable ones during the interaction.

Lastly, we would like to point out the time cost of interaction errors. The answering time presented in Table 1 suggests that an ESM session can be signi cantly extended if participants need to answer a question with two or three attempts. The interaction time can be more than tripled if users encounter two or more errors in one session. Extended interaction time caused by system errors negatively impacts the user experience [20, 54] and increases the user burden [43].

5.3 Implications for VUI Applications

ESM applications can be promising to run on proactive speakers. To capture higher quality data, we suggest that ESM surveys should include more close-ended questions. We use the 5-point numerical scale in our ESM, and we nd numbers, particularly 2 and 4 , cannot be correctly transcribed at times. Numbers under ve are mostly short, single-syllable words, which are likely to have lower recognition accuracy as suggested [27, 36]. The use of pre-de ned labels may be more suitable than numbers. For example, 1 , 2 , 3 , 4 , 5 can be replaced by worst , upset , neutral, good , and excellent . But while the word good has an accuracy rate of 99.9% [36] the usability of other words still need to be evaluated. Furthermore, we suggest that ESM surveys should include fewer open-ended questions. While we nd fewer in-situ systems errors

are triggered at Question 4 (What are you currently doing?), 26.4% of the reported activity data requires manual correction with the use of auxiliary audio recordings. The extensive occurrence of transcription errors for free-form text requires considerable amounts of human e ort for post-analysis. The high cost of human transcription can limit the scale of data collection. Lastly, we recommend that proactive VUI applications should consider enabling the use of the start/end sound of requests[5], i.e., users will hear a sound when the speaker starts listening to commands and after the speaker is done listening to requests. As the interaction with proactive speakers can happen when the speaker is out of sight, the start/end sound can give users audio cues.

On a more general level, we recommend developers use more error-robust words in the dialogue design of their VUI applications. While speech recognition technology is advancing, interaction errors are unlikely to entirely disappear [60]. Once proactive speakers are used in people's homes, interactions are likely to be initiated at inopportune moments, such as when the environment is not quiet. Also, users can have a wider range of accents [34]. Essentially, both ATs and STs are caused by incorrect speech transcription. Therefore, the usability of VUI applications can be enhanced if voice commands are well constructed with words that have higher accuracy rates. For example, the voice commands should incorporate words with two or three syllables (e.g., forecast) and avoid using too many short, single-syllable words and words with homophones (e.g., bean) [12, 27, 36, 51].

5.4 Implications for VUI Research

During the investigation of interaction errors, we nd that the auxiliary audio recordings provide rich data. First, we used the audio recordings as the ground truth of answers to correct transcription errors. When facing recorded errors such as 0 and 85 , we initially assumed that participants gave 1 and 5 . Given the audio recording, we discovered that those recorded answers were actually from the noise in the background. Similarly, we did not expect participants' family members to answer the ESM for them. This replacement would go otherwise undetected if we only relied on recorded answers or logs. Also, we presumed that most incomplete ESMs were caused by consecutive system errors (AT). However, we nd that STs occur four times more frequently than ATs with the real-time recording. Furthermore, we realize that measuring voice interaction should not only focus on the conversation but also include user behaviors such as their speech loudness and the rate of speech or movements and interactions with people within the speaker's vicinity [3, 57]. Microphones are used in many previous studies to understand how users speak to the device and learn how people collaborate together to resolve interaction errors. In our study, even though the multi-turn interaction remains relatively simple, the audio recording still helps us learn many trivial yet important aspects of voice interactions in the wild. Luo et al. [43] used the smart speaker to collect self-reported data and mentioned that their participants needed to manually correct their records. In this case, if there were audio recordings, the user struggle could be captured. On the other hand, audio recordings also introduce privacy concerns. Our system accidentally recorded one participant's kid's voice, which suggests that future VUI researchers should weigh the bene ts of rich audio information against privacy implications. Maybe researchers can consider studying a family as a unit, or the targeted participants can be provided with audio recordings so that they can choose whether they want to delete certain recordings.

Cho and Rader [18] collected two sources of data: what the user said and what the speaker heard. For studying errors with custom VUI applications, we suggest adding another important piece of information: the logs generated by VUI applications. In this study, we extract timestamps from logs of the Proactive and study both the answering time and the system processing and network latency. Learning about the interaction time helps us identify user e ort in answering a question. Learning about the system processing and network latency further helps us understand the user burden. Our empirical data suggests that the time gaps (see Figure 4) mostly remained at 100-200 ms level, but some could be longer than 1 second, especially for the delay between activity response record time and the ending prompt Response log time. When a 1-sec delay is introduced during an interaction, users may consider it is them, rather than the device, that has caused an interaction failure [57]. Understanding at which part delays are generated may reveal shortages of today's framework of smart speakers and o er opportunities for the development of future smart speakers. Therefore, future research should also try to collect time information, especially the back-end processing time.

6 LIMITATIONS AND FUTURE WORK

In this study, we investigate interaction errors that occurred in a 3-week eld voice ESM study of 13 participants. Through manual inspection, we have identi ed di erent interaction and transcription errors as well as user strategies to overcome errors by analyzing logs, recorded ESMs, and audio recordings. While our analysis is mainly based on manual labeling and audio recordings, it should be noted that this method is limited. One missed opportunity is that we did not capture logs stored in Google's My Activity. In other words, we did not have rst-hand evidence of what the Google Home actually heard . This limits us to developing further understanding of the Google Home's speech recognition capability. By referencing logs and recorded data, we may miss some interaction trouble as there is a margin of human errors in manual labeling. Also, microphones are unable to fully depict people's interactions with smart speakers [45]. Our interpretations of user strategies are based on limited information and may not re ect the intention of users [41]. Further, we later identi ed that the USB microphone we used was unstable in di erent environments and captured some hardware noises produced by the Raspberry Pi. We could not accurately count user strategies (e.g., raising voice) and quantify the correlation between environmental noises and interaction errors. Therefore, we adopted a more qualitative approach in summarizing user strategies. Henceforth, future research should try to collect Google Activity logs if possible, use higher-grade microphone arrays to capture richer information (e.g., Angle of Arrival, volumes) of interactions, or even involve participants in the process of mapping and interpreting errors [42].

The unexpected high occurrence of STs and frequent recognition errors of simple numbers unveil that existing smart speakers

[5] https://support.google.com/assistant/answer/9071787

may still face usability issues. But since our investigation of interaction errors is based on a voice ESM application on Google Home, we acknowledge that our findings cannot be generalized to all VUI applications and speaker platforms (e.g., Amazon, Apple). We believe it is necessary to conduct more empirical user studies [7, 71] to quantify and understand interaction errors both in the lab and in the field. For example, a future study with a wider range of participants can consist of two phases. In a controlled environment (i.e., no background noise, identical speaker setting, and fixed distance to the speaker), researchers can first measure the baseline of users' error rates with speakers and investigate the impact of accents and speaking styles [27, 36, 53]. Then, researchers can deploy speakers in people's homes to investigate whether users' error rates and behaviors will be impacted by the environment (e.g., room layouts, ambient noises, the presence of other people). To gain more general insights into interaction errors with speakers, different types of VUI applications can be designed and evaluated, such as ESM applications with various question types (e.g., 7-point scales and word scales) and applications that proactively prompt voice intervention messages to engage users [19]. Finally, another direction to pursue is to study how to reduce the error rate by augmenting speakers with sensors, such as motion sensors or cameras. Sensors can make proactive speakers more context-aware and initiate interactions at opportune moments to reduce the impact from the environment [15, 43]. Camera and other recording devices can capture more detailed user behaviors (e.g., moving and head orientation) around speakers and help us further understand the interaction [54].

7 CONCLUSION

We present an in the wild case study (i.e., in participants' homes) with proactive speakers to investigate interaction errors and user recovery strategies. We examine how users overcome encountered errors while answering voice ESMs. We deployed our ESM application for three weeks with 13 participants and analyzed collected audio recordings of all the interactions between participants and the proactive speaker. With 1,213 ESMs collected, we demonstrate that proactive speakers can capture data via multi-turn conversations. However, our analyses of interaction errors suggest that, even though the implemented ESM application required the speaker only to recognize numbers, interaction errors occurred more frequently than expected. While ultimately, almost all interaction errors were caused by incorrect transcriptions, some errors led to increased user burden and caused abrupt termination of ESMs. Therefore, further research needs to carefully design the VUI applications for data collections. We give specific design recommendations for future voice ESM applications and then discuss implications for more general VUI applications and VUI research methodology. We hope this work can spark future reactive and proactive speakers studies to explore errors during voice interactions. Ultimately, lessons learned from errors can contribute to making smart speakers more usable.

ACKNOWLEDGMENTS

This work is partially funded by ARC Discovery Project DP190102627, NHMRC grants 1170937 and 2004316.

REFERENCES

[1] [n.d.]. https://www.amazon.com/cubic-ai-5-Minute-Plank-Workout/dp/B06XHTCB3Z
[2] [n.d.]. https://assistant.google.com/services/a/uid/000000addca8c8f3
[3] 2021. Smart Speaker Market Global Industry Trends, Share, Size and Forecast Report. https://www.marketwatch.com/press-release/smart-speaker-market-global-industry-trends-share-size-and-forecast-report-2021-02-17?tesla=y
[4] Mohammad Aliannejadi, Manajit Chakraborty, Esteban Andrés Ríssola, and Fabio Crestani. 2020. Harnessing evolution of multi-turn conversations for effective answer retrieval. In Proceedings of the 2020 Conference on Human Information Interaction and Retrieval. Association for Computing Machinery, New York, NY, USA, 33–42.
[5] 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.
[6] Bruce Balentine and David P. Morgan. 2001. How to build a speech recognition application: a style guide for telephony dialogues. EIG Press, San Ramon, CA.
[7] Curtis A Becker. 1979. Semantic context and word frequency effects in visual word recognition. Journal of Experimental Psychology: Human Perception and Performance 5, 2 (1979), 252.
[8] 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. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300473
[9] 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.
[10] Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of Siri, Alexa, and Google Assistant. Journal of medical Internet research 20, 9 (2018), e11510.
[11] Niall Bolger and Jean-Philippe Laurenceau. 2013. Intensive longitudinal methods: An introduction to diary and experience sampling research. Guilford Press, New York, NY, US.
[12] Julia Cambre, Alex C Williams, Afsaneh Razi, Ian Bicking, Abraham Wallin, Janice Tsai, Chinmay Kulkarni, and Jofish Kaye. 2021. Firefox Voice: An Open and Extensible Voice Assistant Built Upon the Web. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445409
[13] Justine Cauell, Tim Bickmore, Lee Campbell, and Hannes Vilhjalmsson. 2000. Designing embodied conversational agents. Embodied conversational agents 29 (2000), 29–63.
[14] Irene Celino and Gloria Re Calegari. 2020. Submitting surveys via a conversational interface: an evaluation of user acceptance and approach effectiveness. International Journal of Human-Computer Studies 139 (2020), 102410.
[15] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello There! Is Now a Good Time to Talk? Opportune Moments for Proactive Interactions with Smart Speakers. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 4, 3 (2020), 1–28.
[16] Ruth Chambers and Paul Beaney. 2020. The potential of placing a digital assistant in patients' homes.
[17] Amy Cheng, Vaishnavi Raghavaraju, Jayanth Kanugo, Yohanes P Handrianto, and Yi Shang. 2018. Development and evaluation of a healthy coping voice interface application using the Google home for elderly patients with type 2 diabetes. In 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC). IEEE, New York, US, 1–5. https://doi.org/10.1109/CCNC.2018.8319283
[18] Janghee Cho and Emilee Rader. 2020. The Role of Conversational Grounding in Supporting Symbiosis Between People and Digital Assistants. Proceedings of the ACM on Human-Computer Interaction 4, CSCW1 (2020), 1–28.
[19] Woohyeok Choi, Sangkeun Park, Duyeon Kim, Youn-kyung Lim, and Uichin Lee. 2019. Multi-stage receptivity model for mobile just-in-time health intervention. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 3, 2 (2019), 1–26.
[20] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705
[21] Richard L Clayton and Debbie LS Winter. 1992. Speech data entry: results of a test of voice recognition for survey data collection. JOURNAL OF OFFICIAL STATISTICS-STOCKHOLM 8 (1992), 377–377.

[22] Michael H Cohen, Michael Harris Cohen, James P Giangola, and Jennifer Balogh. 2004. *Voice user interface design*. Addison-Wesley Professional, Boston, MA, USA.

[23] Hasan Shahid Ferdous, Bernd Ploderer, Hilary Davis, Frank Vetere, and Kenton O'hara. 2016. Commensality and the social use of technology during family mealtime. *ACM Transactions on Computer-Human Interaction (TOCHI)* 23, 6 (2016), 1–26.

[24] Anna K Fletcher and Greg Shaw. 2011. How voice-recognition software presents a useful transcription tool for qualitative and mixed methods researchers. *International Journal of Multiple Research Approaches* 5, 2 (2011), 200–206.

[25] Markus Funk, Carie Cunningham, Duygu Kanver, Christopher Saikalis, and Rohan Pansare. 2020. Usable and Acceptable Response Delays of Conversational Agents in Automotive User Interfaces. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. Association for Computing Machinery, New York, NY, USA, 262–269. https://doi.org/10.1145/3409120.3410651

[26] Shiyoh Goetsu and Tetsuya Sakai. 2020. Different types of voice user interface failures may cause different degrees of frustration. *arXiv preprint arXiv:2002.03582* (2020).

[27] Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication* 52, 3 (2010), 181–200.

[28] Danula Hettiachchi, Zhanna Sarsenbayeva, Fraser Allison, Niels van Berkel, Tilman Dingler, Gabriele Marini, Vassilis Kostakos, and Jorge Goncalves. 2020. "Hi! I am the Crowd Tasker" Crowdsourcing through Digital Voice Assistants. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376320

[29] Jeff Huang and Efthimis N Efthimiadis. 2009. Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management*. Association for Computing Machinery, New York, NY, USA, 77–86. https://doi.org/10.1145/1645953.1645966

[30] Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. Association for Computing Machinery, New York, NY, USA, 143–152. https://doi.org/10.1145/2484028.2484092

[31] Alan Kennedy, Alan Wilkes, Leona Elder, and Wayne S Murray. 1988. Dialogue with machines. *Cognition* 30, 1 (1988), 37–72.

[32] Auk Kim, Woohyeok Choi, Jungmi Park, Kyeyoon Kim, and Uichin Lee. 2018. Interrupting Drivers for Interactions: Predicting Opportune Moments for In-vehicle Proactive Auditory-verbal Tasks. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–28.

[33] Auk Kim, Jung-Mi Park, and Uichin Lee. 2020. Interruptibility for in-vehicle multitasking: influence of voice task demands and adaptive behaviors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–22.

[34] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences* 117, 14 (2020), 7684–7689.

[35] Mitsuki Komori, Yuichiro Fujimoto, Jianfeng Xu, Kazuyuki Tasaka, Hiromasa Yanagihara, and Kinya Fujita. 2019. Experimental Study on Estimation of Opportune Moments for Proactive Voice Information Service Based on Activity Transition for People Living Alone. In *International Conference on Human-Computer Interaction*. Springer, Springer International Publishing, Cham, 527–539.

[36] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. 2018. Skill squatting attacks on Amazon Alexa. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*. USENIX Association, Baltimore, MD, 33–47. https://www.usenix.org/conference/usenixsecurity18/presentation/kumar

[37] Dounia Lahoual and Myriam Frejus. 2019. When users assist the voice assistants: From supervision to failure resolution. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3343413.3377968

[38] Reed Larson and Mihaly Csikszentmihalyi. 2014. The experience sampling method. In *Flow and the foundations of positive psychology*. Springer, Dordrecht, 21–34.

[39] Sunok Lee, Minji Cho, and Sangsu Lee. 2020. What If Conversational Agents Became Invisible? Comparing Users' Mental Models According to Physical Entity of AI Speaker. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–24.

[40] Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M Mitchell, and Brad A Myers. 2020. Multi-Modal Repairs of Conversational Breakdowns in Task-Oriented Dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 1094–1107. https://doi.org/10.1145/3379337.3415820

[41] Anthony J Liddicoat. 2021. *An introduction to conversation analysis*. Bloomsbury Publishing, London, England.

[42] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[43] Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. TandemTrack: Shaping consistent exercise experience by complementing a mobile app with a smart speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376616

[44] Raju Maharjan, Darius Adam Rohani, Per Bækgaard, Jakob Bardram, and Kevin Doherty. 2021. Can we talk? Design Implications for the Questionnaire-Driven Self-Report of Health and Wellbeing via Conversational Agent. In *CUI 2021-3rd Conference on Conversational User Interfaces*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3469595.3469600

[45] Donald McMillan, Moira McGregor, and Barry Brown. 2015. From in the Wild to in Vivo: Video Analysis of Mobile Device Use. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Copenhagen, Denmark) *(MobileHCI '15)*. Association for Computing Machinery, New York, NY, USA, 494–503. https://doi.org/10.1145/2785830.2785883

[46] Michael Frederick McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface*. Springer, Cham.

[47] Hyunsu Mun, Hyungjin Lee, Soohyun Kim, and Youngseok Lee. 2020. A Smart Speaker Performance Measurement Tool. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*. Association for Computing Machinery, New York, NY, USA, 755–762. https://doi.org/10.1145/3341105.3373990

[48] Hyunsu Mun and Youngseok Lee. 2020. Accelerating Smart Speaker Service with Content Prefetching and Local Control. In *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, New York, US, 1–6.

[49] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3173574.3173580

[50] Chelsea M Myers, Anushay Furqan, and Jichen Zhu. 2019. The impact of user characteristics and preferences on performance with an unfamiliar voice user interface. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3290605.3300277

[51] Nick Nikiforakis, Marco Balduzzi, Lieven Desmet, Frank Piessens, and Wouter Joosen. 2014. Soundsquatting: Uncovering the use of homophones in domain squatting. In *International Conference on Information Security*. Springer International Publishing, Cham, 291–308.

[52] Chaewon Park, Yoonseob Lim, Jongsuk Choi, and Jee Eun Sung. 2021. Changes in linguistic behaviors based on smart speaker task performance and pragmatic skills in multiple turn-taking interactions. *Intelligent Service Robotics* 14, 3 (2021), 1–16.

[53] Sonia Paul. 2017. Voice Is the Next Big Platform, Unless You Have an Accent | Backchannel. https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent/

[54] Hannah RM Pelikan and Mathias Broth. 2016. Why that nao? how humans adapt to a conventional humanoid robot in taking turns-at-talk. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 4921–4932. https://doi.org/10.1145/2858036.2858478

[55] Martin Pielot, Bruno Cardoso, Kleomenis Katevas, Joan Ser à, Aleksandar Matic, and Nuria Oliver. 2017. Beyond interruptibility: Predicting opportune moments to engage mobile phone users. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (sep 2017), 1–25. https://doi.org/10.1145/3130956

[56] Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Osaka, Japan) *(UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 825–836. https://doi.org/10.1145/2750858.2804252

[57] 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*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174214

[58] 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*. Association for Computing Machinery, New York, NY, USA, 207–219. https://doi.org/10.1145/2998181.2998298

[59] Aung Pyae and Paul Scifleet. 2019. Investigating the role of user's English language proficiency in using a voice user interface: A case of Google Home smart speaker. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY,

USA, 1–6. https://doi.org/10.1145/3290607.3313038

[60] Stuart Reeves, Martin Porcheron, and Joel Fischer. 2018. 'This is not what we wanted' designing for conversation with voice interfaces. *Interactions* 26, 1 (2018), 46–51.

[61] Melanie Revilla, Mick P Couper, Oriol J Bosch, and Marc Asensio. 2020. Testing the use of voice input in a smartphone web survey. *Social Science Computer Review* 38, 2 (2020), 207–224.

[62] Felicia Roberts, Alexander L Francis, and Melanie Morgan. 2006. The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech communication* 48, 9 (2006), 1079–1093.

[63] 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* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[64] Hyewon Suh, Nina Shahriaree, Eric B Hekler, and Julie A Kientz. 2016. Developing and validating the user burden scale: A tool for assessing user burden in computing systems. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 3988–3999. https://doi.org/10.1145/2858036.2858448

[65] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts. 2007. Information Re-Retrieval: Repeat Queries in Yahoo's Logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR '07)*. Association for Computing Machinery, New York, NY, USA, 151–158. https://doi.org/10.1145/1277741.1277770

[66] Daphne Townsend, Frank Knoefel, and Rafik Goubran. 2011. Privacy versus autonomy: a tradeoff model for smart home monitoring technologies. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, New York, US, 4749–4752.

[67] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search: Perspective Paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (CHIIR '18)*. Association for Computing Machinery, New York, NY, USA, 32–41.

[68] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The experience sampling method on mobile devices. *ACM Computing Surveys (CSUR)* 50, 6 (2017), 1–40. https://doi.org/10.1145/3123988

[69] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. Developing the Proactive Speaker Prototype Based on Google Home. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3411763.3451642

[70] Yukang Yan, Chun Yu, Wengrui Zheng, Ruining Tang, Xuhai Xu, and Yuanchun Shi. 2020. FrownOnError: Interrupting Responses from Smart Speakers by Facial Expressions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376810

[71] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. 2019. Dangerous Skills: Understanding and Mitigating Security Risks of Voice-Controlled Third-Party Functions on Virtual Personal Assistant Systems. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, New York, US, 1381–1396. https://doi.org/10.1109/SP.2019.00016