



# How Compatible is Alexa with Dual Tasking? – Towards Intelligent Personal Assistants for Dual-Task Situations

Shashank Ahire  
Human-Computer Interaction  
Leibniz University Hannover  
Hannover, Germany  
shashank.ahire@hci.uni-hannover.de

Aaron Priegnitz  
Human-Computer Interaction  
Leibniz University Hannover  
Hannover, Germany  
aaronpriegnitz@gmail.com

Oguz Önbas  
Human-Computer Interaction  
Leibniz University Hannover  
Hannover, Germany  
oguz.oenbas@outlook.de

Michael Rohs  
Human-Computer Interaction  
Leibniz University Hannover  
Hannover, Germany  
michael.rohs@hci.uni-hannover.de

Wolfgang Nejdl  
L3S  
Leibniz University Hannover  
Hannover, Germany  
nejdl@L3S.de

## ABSTRACT

Previous literature has reported that users consider hands-free and eyes-free interaction as one of the prime features of IPAs (Intelligent Personal Assistants). Hands-free and eyes-free interaction enables dual tasking. Although users prefer dual tasking with IPAs, it is unknown to what degree current IPAs are compatible with dual tasking. To determine IPA efficiency while dual tasking, we investigate cognitive load in dual-task scenarios with IPAs. In our experiment, we selected a rhythm game as the primary task and everyday IPA requests as secondary tasks. The secondary tasks belonged to four common categories: information search, multimedia control, smart home control, and turn-taking conversations. The findings show that IPAs need significant improvement to support dual tasking. Out of the four categories, only tasks in the smart home and multimedia categories were appropriate for dual tasking, whereas turn-taking conversation and information search had a high cognitive load. Task completion time was significantly different between tasks, but the penalty on the accuracy of the primary task was small. In interviews we found that, due to information abundance in IPA responses and high time pressure during task completion, users tended to make several mistakes. Based on our findings and observations we derive four design recommendations that facilitate dual-tasking while using IPAs.

## CCS CONCEPTS

• **Human-centered computing** → **Natural language interfaces; Sound-based input / output.**

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

*HAI '21, November 9–11, 2021, Virtual Event, Japan*

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8620-3/21/11...\$15.00

<https://doi.org/10.1145/3472307.3484165>

## KEYWORDS

speech interface; intelligent personal assistant; cognitive load; dual tasking; design recommendations.

### ACM Reference Format:

Shashank Ahire, Aaron Priegnitz, Oguz Önbas, Michael Rohs, and Wolfgang Nejdl. 2021. How Compatible is Alexa with Dual Tasking? – Towards Intelligent Personal Assistants for Dual-Task Situations. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21), November 9–11, 2021, Virtual Event, Japan*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3472307.3484165>

## 1 INTRODUCTION

User behaviour of interacting with IPAs when engaged in another task has been reported in various studies [7, 17, 18]. Users have been interacting with IPAs when engaged in activities like cooking, performing household chores, working, studying, or taking care of the kids. While performing two tasks at once we divide our cognitive resources between a primary and a secondary task. Some primary tasks require low attention and generate a low mental demand. Such tasks are perfectly suited for simultaneously performing a second task. But, if the secondary task demands high cognitive workload and attention, then such a secondary task tends to be detrimental to the performance of both the tasks. Efficient support for dual tasking is a desirable feature for IPA users. Hence, it is important to consider dual tasking situations while designing IPAs.

IPAs are capable of performing various requests of different length and structure. Frequently performed requests like “What is the time?” or “How is the weather?” are typically very brief, on the order of 3-5 words. On the other hand, sometimes more complex requests are performed, like “Set me a reminder for tomorrow at 5pm to pay the bill.”, which is a 12 word request. Complexity of a request is difficult to measure. A simplistic measure would be the number of words or the number of information items. However, these represent only rough indicators of complexity. In an IPA interaction, the user has to perceive, think, react, and make decisions. Thus, it is important to investigate the cognitive workload while interacting with the IPA. Until now, only a few studies have considered cognitive demand while interacting with an IPA in a dual task situation [9, 10, 40].

Dual tasking has been studied largely in the automotive context with various secondary tasks, such as searching for a song, having a call, or interacting with a voice assistant [13, 15, 26]. This research has helped to gauge the effects of various secondary tasks on driving and has suggested techniques to reduce the cognitive load in such interactions. On the other hand, the impact of the secondary task with an IPA on the wide range of possible primary tasks has not yet been researched in detail. Accordingly, there is a need to investigate the interaction with IPAs in dual task scenarios in other contexts, such as the home.

In our research we address three main aspects: (1) Identify and compare the cognitive load for frequent IPA categories of tasks in dual-task scenarios. (2) Determine categories of tasks that are suitable for dual-tasking. Identify measures to gauge if a category is suitable for dual tasking. (3) Envision how to design IPA conversations for having seamless interaction in dual tasking scenarios.

In this paper we investigate cognitive load when dual tasking with IPAs. As a primary task we chose a rhythm game. We selected four categories of common IPA request as secondary tasks. The analysis shows that not all IPA tasks are suitable for dual tasking. Our findings highlights several problems with IPA interactions and responses that the IPA delivered in dual-task situations. During our interviews, we identified a critical user behaviour of device switching when the IPA fails to recognize a request. Based on our findings and observations, we propose four design recommendation that will help IPA designers to develop seamless interactions for dual tasking scenarios with IPAs.

## 2 RELATED WORK

Dual tasking in HCI has been studied with different devices like computers, cars, mobile phones, and voice assistants. Researchers have studied the impact of dual taking on the primary and secondary task [1, 3, 12]. The effects of an interruption has been evaluated with respect to performance, errors, attention, and emotional state [3, 13, 26]. Further, some studies examined the appropriate time to interrupt in a primary task [1, 8, 9].

### 2.1 Effects of Interruption on Dual Tasking

The effects of an interruption is profoundly dependent on timing of an interruption. A wrongly timed interruption can have a strong effect on the primary and secondary task. Czerwinski et al. [8] studied the relationship between time, type of message notification, and the intensity of the interruption on computer. They found that interruption had a harmful effect when the participants were performing a primary task of typing, evaluating search results, or interacting with menus. Further, they noted that interruptions have a detrimental effect on user task performance and emotional state. The magnitude of the disruption depends on the mental load at the point of interruption.

On Comparing the interruptions with postpone option (negotiated) and forced interruption. Users preferred negotiated interruption since they had control over the interruptions. User-controlled interruptions limit the negative impact of an interruption. Forced interruptions required more time for performing a primary task in comparison to the interruptions with negotiate options. Also, negotiated interruption saves users time and memory. Moreover,

interruptions at lower cognitive levels increase the chances of correctly performing the secondary task [16]. Similarly, on studying the effect of interruptions in terms of performance, emotional state, and social attribution, Adamczyk et al. [1] found interruptions to have a minimum effect when occurring towards the end of task completion, but had the largest effect when occurring in the middle of performing a task.

When interruptions allow to maintain an associative link between goal memory and environmental context, the resumption time of the primary task is faster in comparison to conditions where the link cannot be maintained [23]. In a mixed-methods study of how people interrupt others engaged in complex tasks, Edwards et al. [9] revealed that people tend to interrupt sooner when the interruption is urgent. Further they found that people have individual strategies for structuring interruptions in terms of word length, utterance naturalness, clarity, and tone.

### 2.2 Cognitive Load in IPA Interactions

Cognitive load in IPAs has been frequently studied in the domain of automobile user interfaces. Strayer et al. [32] examined different voice assistants and their impact on cognitive load while driving. Their results show a significant effect on cognitive load for all voice assistants, compared to the single task scenario of only driving. They also found that while driving, tasks with lower complexity and time duration tend to have lower cognitive load [31].

In assessing the cognitive load of IPAs in a car through a Wizard-of-Oz approach, comparable ratings were found for cognitive load for voice assistants and hands-free conversations on a phone [15]. On examining how driver's behaviour change while using voice assistants to tune the radio when driving on a highway, it was observed that multitasking with the voice assistant significantly decreases the frequency of mirror checks while driving [27].

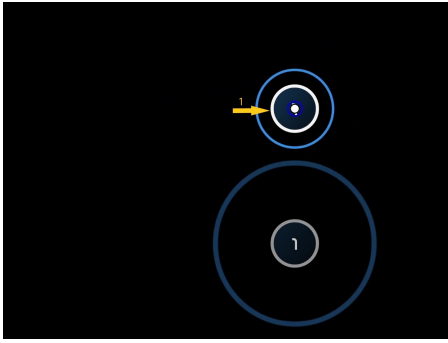
In smart speakers, Edwards et al. [10] investigated the impact caused by voice assistants on the performance of copying and rewording a paragraph. Their results demonstrate a significantly higher workload for rewording a paragraph than for copying. Similarly, on comparing cognitive load for native and non-native English speakers, analyses show higher cognitive load for non-native speakers while interacting with an IPA. Language production and interpretation led to an increase in cognitive load [40]. To determine opportune moments for proactive conversational interactions in domestic settings, it was concluded that opportune moments are dependent on user business, mood, complexity of the primary task, and on the user's social availability [5].

## 3 EXPERIMENT

In our experiment, the participants were asked to perform two tasks simultaneously, the primary task and the secondary task. The primary task was to play a rhythm game and the secondary task was to perform different types of requests with an IPA. Since, Amazon Alexa has the largest market share in Europe<sup>1</sup>, we chose an Amazon Alexa Echo Show 8<sup>2</sup> for processing the requests.

<sup>1</sup><https://voicebot.ai/2018/02/28/amazon-echo-google-home-european-smart-speaker-sales-approach-6-5-million-units-2017/>

<sup>2</sup><https://www.amazon.com/Echo-Show-8/dp/B07PF1Y28C>



**Figure 1: The outer circle encloses the inner circle synchronized with the rhythm of a music beat. On the beat, when outer circle overlaps the inner circle, the user has to click. After each click, the cursor automatically moves to the next target, avoiding the need to move a cursor.**

### 3.1 Primary Task

For our investigation we intended to choose an artificial primary task that requires constant monitoring, requires timely action, and has a predictable and uniform workload pattern similar to our daily activities. Rhythm games have been used before in studies on dual tasking [14, 20]. We chose a simple rhythm game, in which one has to click the mouse button at a precise moment, but without the need for spatial precision. The *osu!* [39] rhythm game generates a consistent and predictable workload with constant visual attention similar to real-world tasks.

While playing *osu!*, players have to click on circles that appear on the screen, matching the rhythm of the music beats (see Figure 1). The more precisely timed the clicks are, higher the score. In order to make the game easier to learn, we modified the game such that the cursor automatically moves to the next target. Hence no spatial precision is required and the users only had to focus on clicking precisely on the beat. We designed a simple rhythm of 3 s, which consisted of 5 clicks with only two intervals of 375 ms and 750 ms (1<sup>st</sup> click, 750 ms pause, 2<sup>nd</sup> click, 375 ms pause, 3<sup>rd</sup> click, 375 ms pause, 4<sup>th</sup> click, 750 ms pause, 5<sup>th</sup> click). This rhythm was repeated over a period of 150 s for all tasks.

For the experiment we created multiple game sessions: A training session, a test session, and a repeat session. The training session had a duration of 150 s and used a pace of 80 bpm. The test session also had a length of 150 s, but was slightly faster at 100 bpm. The repeat session also had a pace of 100 bpm, but only had a duration of 40 s. The repeat session served as a fallback in case a participant was not able to complete the tasks in the test session. In the training session of the game, we had participants practice until they mastered the rhythm of the game and played it reliably.

To cue the participant to perform a secondary task with the smart speaker, an exclamation mark was displayed in the background of the game, as shown in Figure 2. When the exclamation mark appeared, the user was supposed to perform an IPA task. The exclamation mark appeared at an interval of around 30 s.



**Figure 2: On the appearance of exclamation mark, the user is supposed to perform a secondary task with the IPA.**

**Table 1: Categories of IPA tasks and their respective requests.**

Category	Tasks
Multimedia control	<ol style="list-style-type: none"> <li>1. Alexa, play a Justin Bieber song.</li> <li>2. Alexa, set the volume to 5.</li> <li>3. Alexa, skip the song.</li> </ol>
Smart home control	<ol style="list-style-type: none"> <li>1. Alexa, switch on the bedroom lights.</li> <li>2. Alexa, turn on the fan.</li> <li>3. Alexa, turn on the mobile charger.</li> </ol>
Information search	<ol style="list-style-type: none"> <li>1. Alexa, what is the population of Vietnam?</li> <li>2. Alexa, how will be the weather on next Thursday?</li> <li>3. Alexa, who is the fastest swimmer in the world?</li> </ol>
Turn-taking conversation	<ol style="list-style-type: none"> <li>1. Alexa, create a reminder. For – birthday wish. Time – tomorrow 12am.</li> <li>2. Alexa, create an appointment. For – dentist checkup. Time – 25th of January at 5pm.</li> <li>3. Alexa, what’s on my shopping list? Alexa, delete the “fruit” from the shopping list.</li> </ol>

### 3.2 Secondary Task

For secondary tasks, we selected tasks that are frequently performed in our daily activities. Empirical evaluation of smart speakers have found setting up reminders, music control, smart home control, and information search among the most frequently performed tasks [4, 29]. We classified these frequently performed tasks into four categories based on their interaction style, length, and purpose of the task, as shown in Table 1. We categorize them as follows: information search, multimedia control, turn-taking conversation,

and smart home control. Each category consisted of three IPA tasks, so overall each participant had to perform twelve tasks in the experiment. The order of performing tasks was counterbalanced using a Latin Square.

Music streaming is one of the most frequently performed tasks among multimedia tasks. As shown in Table 1, all the three tasks in the multimedia control category were single turn tasks and were between 4 to 6 words long. Likewise, smart home control consisted of tasks that were common single turn tasks and were also between 4-6 words.

Information search tasks were selected based on the type of their response. For instance, “What is the population of Vietnam” has a large number in its response. “Who is the fastest swimmer in the world” has a name in its response, and “How will be the weather on next Thursday?” has a temperature value and a description of weather conditions (sunny, windy, or rainy) in its response. The tasks consisted of 6-9 words each. In the information task, after receiving the answer from Alexa, the participant had to convey the answer to the moderator.

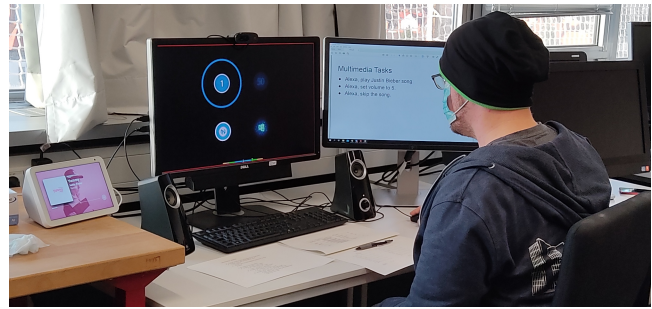
The turn-taking category consisted of a reminder task, an appointment task, and a shopping list task. In the reminder and appointment tasks the participant has to perform these tasks in multiple turns (steps). First, the participant, has to request Alexa to create a particular reminder or appointment. In the following turns, the participant has to provide further information on the task, such as date, time, and name. In the shopping list task, we already added items to the shopping list. In this task, the participant first has to ask Alexa for the items in the shopping list. In the next step, the participant has to identify the fruit (orange) in the shopping list and request its removal. The items in the shopping list were: cake, salad, chicken, spoons, oranges, onion, coffee, lamp, and teapot.

### 3.3 Participants

The participants were recruited by advertising on the university notice board. Overall, 12 participants (11 male, 1 female) were recruited for the study. All of the participants had experience with regularly interacting with smart speakers or voice assistants. The mean age of the participants was 24.7 years with a standard deviation of 2.5 years. Three participants used Siri, three used Alexa, two of them used Google Home/Nest, and only one had experience with using Bixby. Every participant interacted with their IPA between 10-20 times a week. They had experience in performing a dual tasking while studying, exercising, conducting household chores, gaming, or cooking. Each participant was rewarded with a shopping voucher of €10 as a compensation.

### 3.4 Experiment Set-Up

The experiment setup consisted of three components (see Figure 3): the main monitor, the second monitor (reference screen), and the smart speaker (Alexa). The main monitor displayed the *osu!* game. The second monitor displayed the sequence of tasks to be performed in a particular category. The alignment of the second monitor was such that the participant could easily view the screen through their right peripheral vision. Since, users were made familiar with the task before the beginning of the training session, the second monitor was only meant to serve as a reminder. The smart speaker was located in the left peripheral vision of the participant at an



**Figure 3: Experiment setup: The game screen is located in the center, the smart speaker to the left, and the reference screen to the right of the participant. External speakers were used to ensure that the beats are clearly audible.**

audible distance. We used external speakers, so the participants could hear the game beats clearly.

### 3.5 Procedure

After an initial demographic questionnaire, the participants were briefed about the meaning of the different scales of the NASA-TLX questionnaire [38]. To familiarize participants with the tasks we asked them to read the task descriptions and we answered their questions, if any. Next, we trained the participants on the game *osu!*, we asked them to play the training level we designed. The participants were requested to play the training level until they felt comfortable with the game. This practice part took around 2.5 minutes on average. Following this, the participants were asked to fill out the NASA-TLX scales for the played practice trials.

Next, the participant was asked to play the main level of the game. In this period of 150 s all IPA tasks of a given category had to be completed. In case that participants needed a cue, the three IPA tasks of the current category were displayed in textual form on the reference monitor. After performing the tasks of each category, we asked the participant to complete the NASA-TLX questionnaire for the performed tasks. At the end of the experiment an interview was conducted based on the observations made during the experiment and based on the NASA-TLX ratings.

### 3.6 Measures

**3.6.1 NASA-TLX.** For assessing the cognitive workload of the participants in dual-task situation, we used the NASA-TLX<sup>3</sup> questionnaire. The participants were requested to rate each task on six scales: Physical Demand, Mental Demand, Temporal Demand, Effort, Frustration, and Performance. The scales consist of 20 points from “very low” to “very high.”

**3.6.2 Attention demand.** Attention is considered as one of the qualitative parameters for evaluating cognition in voice user interfaces [30]. In particular, we asked the participants to estimate the required level of visual and auditory attention required during the

<sup>3</sup><https://ntrs.nasa.gov/api/citations/20000021487/downloads/20000021487.pdf>

task. Similar to NASA-TLX, we asked our participants to rate attention on a scale of 1-20 (with 1 meaning very low and 20 meaning very high attentional demand).

**3.6.3 Game accuracy.** Accuracy measures were directly computed by osu! game. The score was displayed at the end of each game session<sup>4</sup>. Based on participant click time, they were rewarded with 0, 50, 100, or 300 points. Thus, more timely the click higher the points.

**3.6.4 Task completion time.** We recorded the start and end time for each task. The total time was calculated from the start of the request until the completion of the given task. If the task was not completed in a single attempt, we summed the time taken by all the attempts for completing a particular task.

**3.6.5 Task repetition.** If a participant was not able to accomplish a task on the first try, the task had to be attempted again. Also the participant was asked to repeat the task, if Alexa failed to recognize their request. In the information search category, if the participant failed to perceive Alexa’s response, they were requested to perform the task again.

## 4 RESULTS

We performed quantitative and qualitative analyses of the collected data.

### 4.1 Quantitative Analysis

**4.1.1 Descriptive Statistics.** The median values of NASA-TLX scores (1-20), attention score (1-20), accuracy (percent), and time (seconds) are shown in Table 2. Among the four categories the turn taking category consistently have higher score (except for physical demand and effort), followed by the information search task. The multimedia task is persistently ranked third and the smart home task has the lowest scores on many measures. In Figure 4. For accuracy, turn taking has the lowest value with 38.5% accuracy and the multimedia task has the highest value with 48.3% accuracy.

At 100.9 s the time for the turn-taking task is considerably higher than the other tasks (Table 2). On average, the turn-taking task takes more than twice the time of the information search task (36.4 s), five times as long as the multimedia task (19.3 s), and 6.5 times as long as the smart home task (15.1 s). The repetition count was 15 for information, 14 for turn-taking, 6 for multimedia, and 2 for smart home.

**4.1.2 NASA-TLX scores, attention, and repetition.** We performed a non-parametric Friedman test comparing the four categories on the NASA-TLX scores, the attention score, and repetition counts. As shown in Table 3, the Friedman test showed overall significant differences for all the dependent variables. In the pairwise comparison the smart home and turn-taking category had significant differences for all measures. Likewise, the smart home and information search categories had significant differences for several measures except physical demand, temporal demand, and repetition. In comparison to the information and turn-taking categories, the multimedia category was only significant on five measures out of eight.

<sup>4</sup>[https://osu.ppy.sh/wiki/en/Beatmap\\_Editor/Song\\_Setup#overall-difficulty](https://osu.ppy.sh/wiki/en/Beatmap_Editor/Song_Setup#overall-difficulty)

**Table 2: Median values by category for NASA-TLX scores, attention, accuracy, and time. T: Turn-Taking, M: Multimedia, I: Information, S: Smart Home.**

Measures	M	I	T	S
Physical Demand	4.0	2.0	3.5	2.0
Performance	7.0	12.0	13.5	6.0
Temporal Demand	3.5	8.5	11.0	6.5
Mental Demand	8.5	13.0	14.5	7.5
Effort	8.0	14.0	13.0	7.5
Frustration	4.0	10.0	13.0	3.5
Attention	10.0	15.0	15.5	10.0
Accuracy (%)	48.3	42.2	38.5	45.1
Time (seconds)	19.3	36.4	100.9	15.1

**4.1.3 Time and accuracy.** We performed a repeated-measures analysis of variance for the objective measures time and accuracy. The ANOVA revealed that the difference in accuracy across tasks was not statistically significant ( $F_{11,3} = 1.719, p = 0.18$ ). However, there was a significant effect of task on task completion time ( $F_{9,3} = 7.550, p = 0.0008$ ). A post-hoc Bonferroni-Dunn pairwise comparison exhibited significant differences of completion time between turn-taking and all other categories (multimedia, information search, and smart home).

### 4.2 Qualitative Findings

**4.2.1 Information overload.** While performing tasks in the information category we observed that when Alexa responded with multiple information items, several participants failed to perceive the complete message in a single attempt. Thus, causing them to repeat the request. This particularly happened for the weather task in the information category. In this task, Alexa responded as follows: “On Thursday the 25<sup>th</sup> of July in Bilbao Spain, there will be hazy sunshine with a high of 23°C and low of 20°C.” The response provided five pieces of information in a single sentence: date, location, minimum temperature, maximum temperature, and type of weather (sunny, windy, or rainy).

The high information density in a single response led participants to miss the provided information. One of the participants stated: “It said something like top of 5 and low of 4, it was so close that I couldn’t really understand the numbers.” (P2). Another user stated: “Mental demand was higher for information search and turn-taking since I had to remember more information.” (P3).

**4.2.2 Time pressure.** Some users experienced time pressure while interacting with the IPA. After wake word detection the IPA only listens for a limited amount of time, which leads to time pressure to deliver the request. This pressure increases the temporal demand. In the turn-taking task, the requested information needs to be provided within a limited span of time. The time pressure contributes to errors and mistakes, therefore leading to an increase in the number of attempts. “Waiting for the right moment to say the time and date took my complete attention, so I rated it 20 of 20.” (P6).

**4.2.3 Frequent recognition errors and lack of correction mechanism.** In the interviews several users mentioned that they had stopped

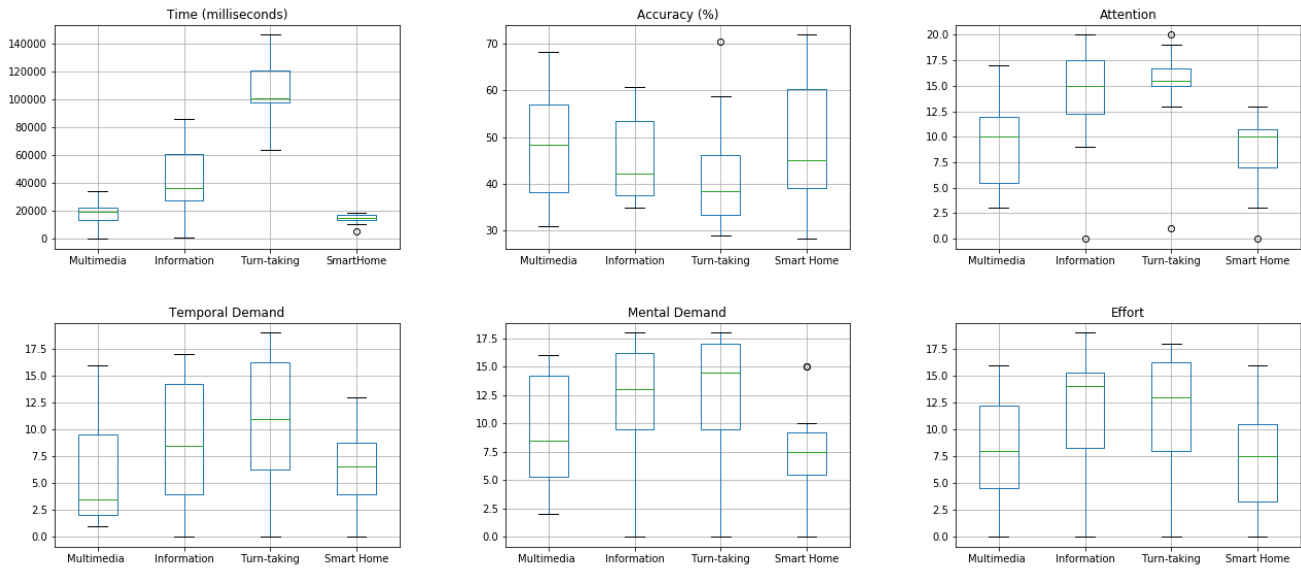


Figure 4: Boxplots of time, accuracy, attention, and the NASA-TLX measures temporal demand, mental demand, and effort.

Table 3: Friedman test of measures. Tick mark(✓) signifies  $p < 0.05$ . T: Turn-Taking, M: Multimedia, I: Information, S: Smart Home.

Measures	Overall		Pairwise					
	$\chi^2$	P-value	M:I	M:T	M:S	I:T	I:S	T:S
Physical Demand	8.12	0.043			✓			✓
Performance	14.13	0.002	✓	✓			✓	✓
Temporal Demand	14.99	0.001	✓	✓				✓
Mental Demand	17.97	0.000	✓	✓			✓	✓
Effort	13.42	0.003	✓				✓	✓
Frustration	16.80	0.000	✓	✓			✓	✓
Attention	16.02	0.001		✓			✓	✓
Repetition	9.5	0.023						✓

setting up appointments and reminders with their smart speaker. Setting up an appointment or reminder consists of providing a name, a date, and a time. Even if a single information item is not recognized correctly, the whole task has to be redone. Additionally, the participants perceived it as difficult to correct mistakes.

*“It is always a frustrating task to set up an appointment and it is much easier to do it with the phone, because getting the time, date, and name of the appointment correct is difficult. When it comes to recognizing certain words, they [smart speakers] always find it difficult.”* (P3).

Due to frequent recognition errors and lack of an error correction technique, participants preferred using the app on their phone for appointments and reminder tasks. Similarly, another participant expressed the same behaviour of device switching while performing information search on smart speakers. *“Even after speaking louder and slowing down, the smart speaker is unable to recognise my request. Then, I will Google it on my mobile phone.”* (P9).

**4.2.4 Pace switching and emphasizing.** We noticed that participants changed their normal utterance pace and slowed down while dual tasking with the smart speaker. Several participant switched their utterance pace during the task and justified this by saying *“I change my pace, because if I speak faster she [Alexa] doesn’t understand me. At the end it takes longer than if I speak more slowly.”* (P8).

Similarly, some of the participants also emphasised on a specific vowel in the word if the device failed to recognised it. For instance the user emphasised on the sound of ‘S’ in the word swimmer, when the device could not recognise it in the first attempt. Asking the participant about it, they responded: *“I just tried to speak clearer and slower and speak ‘S’ louder.”* (P9). The same participant also tended to incline their upper body towards the smart speaker during the interaction. *“In the first task she [Alexa] didn’t react to me, so I wanted to make sure she can understand me properly.”* (P8). The user behaviour of pace switching and emphasizing is also reported in previous literature [7, 22].

## 5 DISCUSSION

### 5.1 Task Comparison

The primary aim of our experiment was to analyze and compare the cognitive load for the different categories of tasks in dual task scenarios. Also, determine the categories of tasks that are compatible for dual tasking. Our findings show that the cognitive load for our selected categories considerably varied on measures like time, temporal, mental demand, and performance (shown in Figure 4). Overall, the turn taking category scored higher on many measures, followed by information search and multimedia category. The smart home category scored consistently low on many measures.

On comparing the accuracy of the primary task, all the tasks had low accuracy (below 50%) while dual tasking. Among all the tasks the multimedia task had the highest accuracy with 48.3%, followed by smart home and information search with 45.1% and 42.2% respectively. Turn taking had lowest accuracy at 38.5%. ANOVA revealed no significant differences regarding the accuracy of tasks. The average time taken to execute tasks varied significantly. The time difference between the lowest (smart home) and highest (turn-taking) task was 85 s. Turn taking required a relatively high task completion time.

### 5.2 Dual Tasking with IPAs

Although users prefer to interact with IPAs while performing a primary task. The findings show that not all tasks are suitable for dual tasking. Among the tasks that we considered, only the multimedia and smart home category had relatively low NASA-TLX scores (Table 2) and were thus suited for dual tasking with an IPA (Alexa). On the other hand, information search and turn taking had significantly higher cognitive load scores. Comparing them with multimedia and smart home shows that they are significantly different on several measures (Table 3).

Overall, pairwise comparison revealed an evident division: multimedia and smart home had significantly lower workload when dual tasking than information search and turn taking. Our findings are consistent with prior work, which states that tasks with lower complexity and time tend to have lower cognitive load [31]. In our experiment, the tasks that had less frustration, completion time, mental demand, and better performance were appropriate for dual tasking with IPAs.

### 5.3 Response Generation in IPAs

For the information search category, the effort for delivering the request was not huge (6-9 words) in comparison to the multimedia and smart home control categories (4-6 words). However, overall the median values in the information search category were significantly higher in comparison to the multimedia and smart home categories. This means that the response generated for the information search tasks had higher cognitive load. In some measures it was even higher than for the turn taking category.

To reduce cognitive load in the information search category, information presentation plays a critical role. After recognizing the request accurately, it is important to present the requested information in an easily accessible manner. It is essential that the user spends a minimum time, effort, and attention to process a

particular piece of information. In order to build a seamless interaction for dual-task scenarios, IPAs need to prioritize the information and present the requested information in a short and uncluttered fashion. Additionally, we also need to reduce the verbiage, which current IPAs are overloaded with.

### 5.4 Abandoning IPAs for Turn-Taking Tasks

Due to persistent errors while inputting text and limited error correction mechanisms, users have to discard the current task and start over again or even switch to another device to perform the task. Error correction with IPAs has been emphasized for a long time [18, 34]. Inconvenience in correcting errors has been one of the major problems with IPAs. In turn-tasking tasks, the user has to input multiple information items (date, time, name, and location). If the IPA commits a mistake in recognizing any information item, the user has to redo the task. Redoing the task increases user frustration and task completion time. Hence, users are compelled to abandon IPAs for turn-tasking tasks.

Previous literature have found non-essential use of IPA, unmet user expectations, non-proactive interface as the prime reasons for abandonment of IPA [6, 35]. But, until now, abandonment of IPA due to recognition errors in longer tasks (turn-taking) was an unreported user behaviour.

## 6 DESIGN RECOMMENDATIONS

Based on the observations and findings in the experiment. We put forward four design recommendations.

### 6.1 Information Prioritization Modeling

Information search and presentation is one of the core functions of IPAs. Although, IPAs have improved in finding information for the user, their presentation style is fairly monotonous. Currently, IPAs are delivering diverse responses, but information presentation pattern is comparatively uniform. As information search is one of the most frequently performed tasks of IPAs [4], it is essential for IPAs to consider information prioritization. For the request “What is the population of Mumbai?” Alexa answers “According to the most recent Indian census in 2011, the population of Mumbai is about 12.5 million.” In long responses, information prioritisation should also be taken into consideration. Prioritisation of information depends on the information the user is asking for, i.e., the primary information and the auxiliary information of the response, i.e., secondary information. The primary information should be presented first in the response, followed by secondary information. In case of above request the response should have been “The population of Mumbai is about 12.5 millions, according to the most recent Indian census in 2011.” This helps the user to focus on primary information first, followed by the secondary information, which eventually minimizes their attention requirement.

### 6.2 Reducing Verbiage

To solve the problem of information overload we suggested in Sect. 4.2.1 that IPA designers should try to reduce the verbiage in an IPA response. Excessive use of words to present a simple piece of information leads to more time, attention demand, and mental demand. This likely affects the performance of the primary task.

Minimal dialogue and meticulously drafting the spoken output has been also echoed in guidelines for voice user interfaces (VUIs) [19, 33]. IPAs should have a model for composing and presenting a response in the simplest possible manner. The goal of the model is to structure the response in a way such that it consumes minimal mental resources of the user. For example, in the case of the weather task in the information search category, the response could have been “The weather will be hazy with temperatures between 20°C and 23°C.”

### 6.3 Command-Based Error Correction

As found in the Sect. 4.2.3, if an IPA misunderstands the input, there should be a provision to allow the user to correct an error. Currently, there is no provision for correcting errors in IPAs. If an error occurs, the user has to start over. The frequent need to repeat a task leads to frustration and abandonment of IPAs, leading users to switch to another modality for performing the task. Error prevention and recovery have been highlighted in existing heuristic and design guidelines for VUIs [19, 33, 37]. To deal with the problem of error correction, one solution we propose is to provide commands for modification. A command-based editing strategy is considered as one of the convenient approaches for editing a dictated text [11]. Upon encountering a particular error correction command the IPA will enter a correction mode for the ongoing task. For instance, while setting up a reminder if the “name” field is incorrectly set, then if the user asserts commands like “replace” or “edit” the IPA can edit or replace the contents of the field.

### 6.4 User-Controlled Interaction

IPAs are interrupting the user proactively for reminding to attend meeting, to take medication, and for home safety and security events [25, 28]. IPAs seem to evolve towards proactive forms of interaction [5, 36]. It is important for users to have control of their interactions. Currently, IPA interruptions are short and can be negotiated using yes or no. But as interruptions are associated with higher information density and require immediate and urgent attention, it will be necessary to have complete user control of the on-going interaction. This situation is unlike in a GUI, where user has control over the pace of a task while inputting data. In IPA interactions, currently the user has negligible control over the progress of on-going interactions. Previous literature has showed users prefer controlled interruption (negotiated) than uncontrolled interruption [16]. User control in VUIs has also been emphasized in guidelines [19, 37].

If IPA interaction is time bound, there should be a provision to pause the interaction. Such a feature would not only be helpful for mainstream users, but also to the older adults and users with disabilities. For instance, while setting up an appointment, if the user requires some time for collecting the information to be submitted, there should be a mechanism to hold the ongoing interaction. The user could issue a command that pauses the IPA interaction until the user utters a command to resume the on-going interaction.

## 7 LIMITATIONS

The experiment was conducted during the COVID-19 pandemic under a strict hygiene protocol. Nonetheless, some participants

might have felt stressed or feared a potential health risk when participating in the study. This might have impacted the subjective ratings. Our participants were relatively young (mean age 24.7 years, standard deviation 2.5 years) and biased towards a male population. Also, the experiment setup was a best-case scenario with no external noise and with an ideal distance from the IPA for voice recognition.

## 8 FUTURE WORK

For future work, we intend to propose an adaptive user interface for IPAs. If a user is dual tasking IPA, could adapt its interaction and response. Further, the IPA could curtail the response and deliver only primary information in the response, using an information prioritization technique as proposed. Also, an IPA could minimize the number of user inputs when it finds users dual tasking.

## 9 CONCLUSION

As the IPA usage is increasing, they are designed for different user groups [21] and contexts [24] and are built for human-like conversations [2]. It is likely that dual tasking behaviour will be also observed in different user groups and contexts. Hence, it is important to design and develop IPAs that are compatible with dual-task situations. From our investigation we conclude that, even though users prefer to dual task in hands-free and eyes-free interactions with IPAs, current IPAs are not compatible with dual tasking. Only the smart home and multimedia control tasks were suitable for dual tasking. In qualitative interviews we found the user behaviour of device switching when performing some tasks. Further, we noticed that current IPA responses lack information prioritization, are densely packed with information, and are unnecessarily verbose.

To change this situation, we appeal to IPA designers to consider dual tasking as one of the prime use case when designing IPA interactions. Furthermore, through our design recommendations we aim to envision IPAs with lower cognitive load and better compatibility with dual tasking. Our design recommendations include: information prioritization, curtailing verbiage, command-based error correction, and user-controlled interaction.

## REFERENCES

- [1] Piotr D. Adamczyk and Brian P. Bailey. 2004. *If Not Now, When? The Effects of Interruption at Different Moments within Task Execution*. Association for Computing Machinery, New York, NY, USA, 271–278. <https://doi.org/10.1145/985692.985727>
- [2] Shashank Ahire and Michael Rohs. 2020. Tired of Wake Words? Moving Towards Seamless Conversations with Intelligent Personal Assistants. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) (CUI '20). Association for Computing Machinery, New York, NY, USA, Article 20, 3 pages. <https://doi.org/10.1145/3405755.3406141>
- [3] Brian P. Bailey, Joseph A. Konstan, and John V. Carlis. 2001. The Effects of Interruptions on Task Performance, Annoyance, and Anxiety in the User Interface. In *Proceedings INTERACT '01*. IOS Press, 593–601.
- [4] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (Sept. 2018), 24 pages. <https://doi.org/10.1145/3264901>
- [5] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 74 (Sept. 2020), 28 pages. <https://doi.org/10.1145/3411810>
- [6] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a Kind Friend is Now a Thing: Understanding How Conversational Agents at Home Are Forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego,

- CA, USA) (*DIS '19*). Association for Computing Machinery, New York, NY, USA, 1557–1569. <https://doi.org/10.1145/3322276.3322332>
- [7] Benjamin R. Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. “What Can i Help You with?”: Infrequent Users’ Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) (*MobileHCI '17*). Association for Computing Machinery, New York, NY, USA, Article 43, 12 pages. <https://doi.org/10.1145/3098279.3098539>
- [8] Mary Czerwinski, Ed Cutrell, and Eric Horvitz. 2000. Instant Messaging and Interruption: Influence of Task Type on Performance. (December 2000), 356–361. OZCHI 2000 Conference Proceedings.
- [9] Justin Edwards, Christian Janssen, Sandy Gould, and Benjamin R. Cowan. 2021. Eliciting Spoken Interruptions to Inform Proactive Speech Agent Design. In *CHI 2021 - 3rd Conference on Conversational User Interfaces* (Bilbao (online), Spain) (*CHI '21*). Association for Computing Machinery, New York, NY, USA, Article 23, 12 pages. <https://doi.org/10.1145/3469595.3469618>
- [10] Justin Edwards, He Liu, Tianyu Zhou, Sandy J. J. Gould, Leigh Clark, Philip Doyle, and Benjamin R. Cowan. 2019. Multitasking with Alexa: How Using Intelligent Personal Assistants Impacts Language-Based Primary Task Performance. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, Article 4, 7 pages. <https://doi.org/10.1145/3342775.3342785>
- [11] Debjyoti Ghosh, Can Liu, Shengdong Zhao, and Kotaro Hara. 2020. Commanding and Re-Dictation: Developing Eyes-Free Voice-Based Interaction for Editing Dictated Text. *ACM Trans. Comput.-Hum. Interact.* 27, 4, Article 28 (Aug. 2020), 31 pages. <https://doi.org/10.1145/3390889>
- [12] Mingjun He, Jianbin Guo, and Shengkui Zeng. 2020. *Cognitive Load Measurement and Impact Analysis on Performance in Dual-Task Situations*. Association for Computing Machinery, New York, NY, USA, 303–307. <https://doi.org/10.1145/3425329.3425388>
- [13] Shamsi T. Iqbal, Yun-Cheng Ju, and Eric Horvitz. 2010. *Cars, Calls, and Cognition: Investigating Driving and Divided Attention*. Association for Computing Machinery, New York, NY, USA, 1281–1290. <https://doi.org/10.1145/1753326.1753518>
- [14] Myoungsoon Jeon, Benjamin K. Davison, Michael A. Nees, Jeff Wilson, and Bruce N. Walker. 2009. Enhanced Auditory Menu Cues Improve Dual Task Performance and Are Preferred with In-Vehicle Technologies (*AutomotiveUI '09*). Association for Computing Machinery, New York, NY, USA, 91–98. <https://doi.org/10.1145/1620509.1620528>
- [15] David R. Large, Gary Burnett, Ben Anyasodo, and Lee Skrypchuk. 2016. Assessing Cognitive Demand during Natural Language Interactions with a Digital Driving Assistant. In *Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Ann Arbor, MI, USA) (*AutomotiveUI '16*). Association for Computing Machinery, New York, NY, USA, 67–74. <https://doi.org/10.1145/3003715.3005408>
- [16] Terri Lenox, Neil Pilarski, and Lance Leathers. 2012. The Effects of Interruptions on Remembering Task Information. 5 (10 2012).
- [17] Yun Liu, Lu Wang, William R. Kearns, Linda Wagner, John Raiti, Yuntao Wang, and Weichao Yuwen. 2021. *Integrating a Voice User Interface into a Virtual Therapy Platform*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411763.3451595>
- [18] 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*). ACM, New York, NY, USA, 5286–5297. <https://doi.org/10.1145/2858036.2858288>
- [19] Christine Murad, Cosmin Munteanu, Benjamin R. Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (2019), 33–45. <https://doi.org/10.1109/MPRV.2019.2906991>
- [20] Babette Park and Roland Brünken. 2015. The Rhythm Method: A New Method for Measuring Cognitive Load—An Experimental Dual-Task Study. *Applied Cognitive Psychology* 29, 2 (2015), 232–243. <https://doi.org/10.1002/acp.3100> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.3100>
- [21] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, Shashank Ahire, Anirudha Joshi, Deepak Sahoo, Nimish Maravi, and Bhakti Bhikne. 2019. StreetWise: Smart Speakers vs Human Help in Public Slum Settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (*CHI '19*). ACM, New York, NY, USA, Article 96, 13 pages. <https://doi.org/10.1145/3290605.3300326>
- [22] 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* (Montreal QC, Canada) (*CHI '18*). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3174214>
- [23] R. Ratwani, Alyssa E. Andrews, Jenny D. Sousk, and J. Trafton. 2008. The Effect of Interruption Modality on Primary Task Resumption. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52 (2008), 393 – 397.
- [24] Simon Robinson, Jennifer Pearson, Shashank Ahire, Rini Ahirwar, Bhakti Bhikne, Nimish Maravi, and Matt Jones. 2018. Revisiting “Hole in the Wall” Computing: Private Smart Speakers and Public Slum Settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (*CHI '18*). ACM, New York, NY, USA, Article 498, 11 pages. <https://doi.org/10.1145/3173574.3174072>
- [25] Ben F. Rubin and Ry Crist. 2019. *What Amazon’s Alexa will tell us in 2019*. Retrieved September 6 2021 from <https://www.cnet.com/home/smart-home/what-amazon-alexa-will-tell-us-in-2019/>
- [26] Dario D. Salvucci, Daniel Markley, Mark Zuber, and Duncan P. Brumby. 2007. iPod Distraction: Effects of Portable Music-Player Use on Driver Performance (*CHI '07*). Association for Computing Machinery, New York, NY, USA, 243–250. <https://doi.org/10.1145/1240624.1240665>
- [27] Ben Sawyer, Joonbum Lee, Jonathan Dobres, Bruce Mehler, Joseph Coughlin, and Bryan Reimer. 2016. Effects of a Voice Interface on Mirror Check Decrements in Older and Younger Multitasking Drivers. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60 (09 2016), 16–20. <https://doi.org/10.1177/1541931213601004>
- [28] ERIC HAL SCHWARTZ. 2019. *LifePod is Taking Reservations for ‘Proactive’ Voice Assistant and Smart Speaker for the Elderly*. voiceBot. Retrieved September 6 2021 from <https://voicebot.ai/2019/06/27/lifepod-is-taking-reservations-for-proactive-voice-assistant-and-smart-speaker-for-the-elderly/>
- [29] Alex Sciuto, Armita 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>
- [30] Katie Seaborn and Jacqueline Urakami. 2021. *Measuring Voice UX Quantitatively: A Rapid Review*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411763.3451712>
- [31] David Strayer, Joel Cooper, Jonna Turrill, James Coleman, and Rachel Hopman. 2016. Talking to your car can drive you to distraction. *Cognitive Research: Principles and Implications* 1 (12 2016). <https://doi.org/10.1186/s41235-016-0018-3>
- [32] David Strayer, Joel Cooper, Jonna Turrill, James Coleman, and Rachel Hopman. 2017. The smartphone and the driver’s cognitive workload: A comparison of Apple, Google, and Microsoft’s intelligent personal assistants. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 71 (06 2017), 93–110. <https://doi.org/10.1037/cep0000104>
- [33] Bernhard Suhm. 2003. Towards best practices for speech user interface design.
- [34] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal Error Correction for Speech User Interfaces. *ACM Trans. Comput.-Hum. Interact.* 8, 1 (March 2001), 60–98. <https://doi.org/10.1145/371127.371166>
- [35] Milka Trajkova and Aqueasha Martin-Hammond. 2020. “Alexa is a Toy”: Exploring Older Adults’ Reasons for Using, Limiting, and Abandoning Echo. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376760>
- [36] Jing Wei, Tilman Dingler, and Vassilis Kostakos. 2021. *Developing the Proactive Speaker Prototype Based on Google Home*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3411763.3451642>
- [37] Zhuxiaona Wei and James A. Landay. 2018. Evaluating Speech-Based Smart Devices Using New Usability Heuristics. *IEEE Pervasive Computing* 17, 2 (2018), 84–96. <https://doi.org/10.1109/MPRV.2018.022511249>
- [38] Wikipedia contributors. 2019. NASA-TLX — Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/w/index.php?title=NASA-TLX&oldid=916786772> [Online; accessed 24-February-2021].
- [39] Wikipedia contributors. 2021. Osu! — Wikipedia, The Free Encyclopedia. <https://en.wikipedia.org/w/index.php?title=Osu!&oldid=1007577903> [Online; accessed 21-February-2021].
- [40] Yunhan Wu, Justin Edwards, Orla Cooney, Anna Bleakley, Philip R. Doyle, Leigh Clark, Daniel Rough, and Benjamin R. Cowan. 2020. Mental Workload and Language Production in Non-Native Speaker IPA Interaction. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (Bilbao, Spain) (*CHI '20*). Association for Computing Machinery, New York, NY, USA, Article 3, 8 pages. <https://doi.org/10.1145/3405755.3406118>