Evaluating User Preferences for Augmented Reality Interactions with the Internet of Things

Shreya Chopra
Department of Computer Science
University of Calgary
Calgary Alberta Canada
shreya.chopra@ucalgary.ca

Frank Maurer
Department of Computer Science
University of Calgary
Calgary Alberta Canada
fmaurer@ucalgary.ca

ABSTRACT

We investigate user preferences for controlling IoT devices with headset-based Augment Reality (AR), comparing gestural control and voice control. An elicitation study is performed with 16 participants to gather their preferred voice commands and gestures for a set of referents. We analyzed 784 inputs (392 gestures and 392 voice) as well as observations and interviews to develop an empirical basis for design recommendations that form a guideline for future designers and implementors of such voice commands and gestures for interacting with the IoT via headset-based AR.

CCS CONCEPTS

• Human-centered computing~User studies • Mixed / augmented reality • HCI theory, concepts and models • Interaction techniques • Empirical studies in interaction design

KEYWORDS

Augmented Reality, Internet of Things, gestural interaction, voice commands, user-centered design, human computer interaction, elicitation study.

INTRODUCTION

Our work investigates user preferences for interacting with the Internet of Things (IoT) using Augmented Reality (AR) headsets as output devices and gesture or voice input. The Internet of Things is a buzzword to describe the connection of everyday objects to the Internet, enabling them to send and receive data to each other [17]. This includes devices being able to generate data and output it to the network or receive control commands from the network. Examples of such devices are televisions, lights, and thermostats. Usually, if these devices have Internet capabilities, they can be controlled by phone apps (such as the Phillips Hue app to control smart lights) or voice-controlled home systems (such as the Google Home). Our work investigates what it would be like to interact with these IoT devices with augmented reality headsets. When we understand how users want to interact using gestures and voice commands, we can develop design guidelines for future systems. Concretely, we investigate what kind of gestures and voice commands people want to use and whether they prefer one control method over the other. The results of our elicitation study will help designers and implementers of such interactions to improve the user experience of the applications. We also develop a way to make sense of voice command elicitation results to form a conclusion.

AVI '20, September 28-October 2, 2020, Salerno, Italy
 ACM ISBN 978-1-4503-7535-1/20/09…$15.00
https://doi.org/10.1145/3399715.3399716

1 Motivation

Currently, headset AR technology often relies on gesture-based interactions. Home-based IoT devices can already be controlled using voice commands (Siri, Alexa, Google). Thus, hand gestures and voice commands were selected as the main input methods. Device-based interaction techniques using controllers & touch display have been discussed by other work [4, 7, 11, 18, 23, 32]. However, we wanted to focus on hands-free interactions. The importance of hands-free interactions is motivated by tasks such as cooking and mechanics where the hands of the user are used for performing work. In addition, hands can get greasy and this could expose any controller used for interacting with the device to the same. Additionally, gaze input is a method of hands-free input often used in headset-based AR in conjunction with voice or gestures. However, we keep our study at 2 variables (gestures and voice) to determine how users want to interact with these two major methods of input. Furthermore, we wanted to see whether users have a preference between voice and gesture input as mainstream headsets such as the Microsoft HoloLens provide both options to the user.

We also assume that interaction with IoT devices via headset-based AR in the near future will be using current interaction paradigms at least for some parts of the UI. Thus, we decided to develop WIMP (windows, icons, menus, pointers) based interfaces for our study. By doing this, we are focusing on how users want to interact with AR and IoT using a WIMP interface: which is what we assume will occur as we make a transition between computers and direct spatial interaction. We focus on home usage scenarios of AR interaction with IoT devices. Our aim is to understand the needs of the general user rather than a specialized user (i.e. factory, industrial). This contributes to bringing this technology to the general public. We also see that a lot of AR IoT scenarios are based on home tasks.

2 Research Questions

Our research aims to answer the following questions:

1. What gestures and voice commands do users want to use to trigger specific IoT control tasks?

2. What can be derived from the elicitation study results to make recommendations to future designers and implementers of headset AR-based IoT controls?
3. Do users have a preference between voice control versus gesture control for headset AR based IoT controls?

RELATED WORK

A major goal of user elicitation studies is enabling designers to create interactions based on what is intuitive and simple for the end user. Current interaction techniques for combining AR and IoT applications are based on a designer's intuition. However, there is a lack of empirical data specifically for this combination. Gestural, voice commands, and multimodal elicitation studies exist for other hardware, and these are useful to us in terms of method and design.

1 Technology and Input

Current AR headsets offer multiple means of input. Hand-held controllers that can be used to interact with holographic components exist [10, 11]. Gaze input is used to control a cursor [20, 30]. Some headsets offer a touchpad on the glasses [13]. Headpose (the rotation of the head) can also be used as input [1]. External forms of input such as a mobile apps and keyboard also exist [1]. Currently, the most common forms of headset AR input are gaze, hand gestures, and voice commands. Current IoT technology also offers various forms of input. IoT devices may fall under the category of devices or control points. Control points are IoT devices or applications through which users can control other IoT devices. Examples of control point IoT devices are the Google Home and Amazon Echo which utilize voice command input [26]. Oftentimes control points are mobile or web applications such as IFTTT or the Philips Hue App which are based on keyboard & mouse or touch screen input [16].

2 User Elicitation Studies

Elicitation studies are conducted to gain insight into how the end-user wants to interact with a system. User-preferred gestural input has been extensively studied as compared to voice command input. Over 38 gestural elicitation studies are discussed by earlier work [5, 6, 9]. An elicitation study usually consists of a list of referents (or tasks) such as "insert" or "cut" for which users propose gestures [31]. Sometimes users are asked for multiple proposals and subsequently their favourite one. This usually avoids legacy bias where previous experience influences proposals [3, 5, 6, 22, 29]. Wizard-of-oz approaches in which the user believes the system to be controlled by them while the researcher controls it [3, 19] are often used in elicitation studies. These techniques are employed to elicit what interactions users want without the limitations of the current hardware/software. Emergent gestural elicitations are performed by researchers and we look to these for inspiration regarding methodology [5, 6, 14, 21, 22, 23, 28, 31, 32].

There is a limited amount of work, especially in the same context as ours, that acknowledges voice commands outside of multimodal elicitation studies. *Hüttenrauch et al.* elicit voice commands for controlling mobile devices and services [15]. Their process includes an elicitation study in which participants provide commands and a subsequent validation in which commands are given to participants who convey their understanding of what the command would produce. "Multimodal" refers to the usage of multiple modes of user input. One of the few works that explicitly study gestural and voice elicitation in combination (separately to other multimodal work) is *Bolt's* elicitation with earlier technology [4]. Previous multimodal studies utilize wizard-of-oz approaches, online surveys, and video interviews [25, 27]. We refer to multimodal and solely voice based elicitations to develop our voice elicitation methodology. Our combo of voice input versus gestural input for controlling IoT devices with AR is, to the best of our knowledge, novel. We also directly compare voice and gestures.

3 Analysis Techniques

Elicitation studies are analyzed using qualitative and quantitative methodologies. Agreement rates are used to determine whether there is enough consensus amongst participants to lead to a conclusive set of interactions that should be provided by a system [7, 21, 31, 32]. Other calculations such as number of occurrences, user preference count, and percentage are sometimes used [12, 21, 25, 27]. Qualitative observations for gestures include finger and hand usage [5, 6, 12]. Observations for voice commands include relevance of words and syntactic structure [12]. *Peshkova et al.* develop mental models regarding how participants choose to formulate control [27].

ELICITATION STUDY

We conducted an elicitation study to determine user preferences for controlling IoT devices through headset AR.

1 Participants

16 volunteers participated in the main study (6 females, 9 males, 1 preferred not to say). We recruited through email lists, word of mouth, and snowball-sampling. The age range of participants was from 19 to 52 years (Mean = 27.125, Median = 23.5, SD = 8.085). 13 participants came from a Computer Science background along with one each from Math, Geomatics, and Mechanical Engineering. 8 out of 16 participants reported having prior experience with some form of AR and 13 reported interacting with IoT devices.

2 Apparatus

The 1st generation Microsoft HoloLens was used to display the holographic components of the experiment to the participants. We developed a wizard-of-oz study in which the system was controlled by the researcher (albeit in front of the participant), ensuring that any interaction the user chose to use would work. The user was being prompted at each step of the task so that they knew exactly what to accomplish (i.e. "set the hour slider to 4"). Each UI component was mapped to our laptop keyboard so that the researcher could simultaneously execute the user's input and show it to them via output. Holographic remoting was implemented via a laptop so the researcher could see what the participant saw and go to the next step when needed. This design provided the user freedom to perform any interaction they wanted without technical limitations such as gestural, command, accent, and gaze-and-commit recognition.

3 Scenarios

The difference between general purpose AR and IoT-specific AR is that IoT devices are expected to respond as a result of being controlled by AR. This is a step further than the sole augmentation of holographic objects on physical objects that are non-responsive.

However, our scenarios were based on input rather than output as we wanted to elicit requirements for user input. Thus, we created wizard-of-oz implementations to avoid interaction limitations resulting from technology. The goal was to determine how users want to control IoT devices in their homes using an AR headset as the output device. We initially investigated and developed a list of small actions that a user would potentially execute when controlling IoT devices with AR in a home. The list consisted of actions such as: copy/paste, save, play/pause, drag and drop, video capture, connect/link, access information, etc. We grouped the actions into 4 scenarios: Interacting with a Menu System, Environmental Control, Media Control, and Following a Workflow. This became our list of typical usage scenarios for IoT controlled by AR. For each of these four scenarios, we chose two different tasks to avoid users repeating a task with different interaction modes. For each task, the participant was told to complete the task using either voice commands or gestures.

3.1 Interacting with a Menu System

Menu systems are quite common in current user interfaces and we expect that they will also be utilized in headset-based AR interfaces. Hence, we consider that interacting with a Menu System will be a typical usage scenario for headset-based AR. The two tasks (contexts) in this scenario are: Connecting a Computer to a Print Queue and Setting up a Lights Schedule as seen in the sequential figures below.

Figure 1: Connecting a Computer to a Print Queue Interface. [Left] Initially, the holographic link button appears next to the physical computer. The user selects the *link* button: either via voice commands or gestures. [Right] The hierarchical menus are expanded. The user selects appropriate intermediate steps to connect to Printer 3. The system indicates successful connection. Lastly, the user is asked to collapse the entire menu.

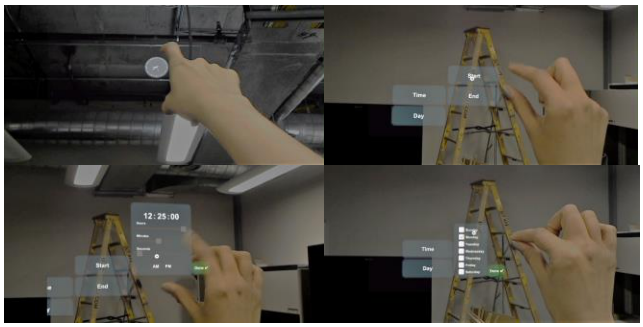

Figure 2: Setting Up a Lights Schedule Interface (sequential). [Top Left] Holographic *clock* button appears next to a physical light when the user looks at it, and the button is selected to expand the menu. [Top Right] The expanded menu appears in front of the user, and the user selects *start* time. [Bottom Left] The user: sets the *hours*, *minutes*, *seconds* sliders and the *AM/PM* toggle for the start time. Then, *done* is selected, and the same menu appears for the *end* time. [Bottom Right] The user toggles on the days that they want the light to turn on.

3.2 Environmental Control

Environmental control is selected as a scenario as to control the physical environment is considered a typical IoT home application. The two tasks in this scenario are: Blinds Control and Thermostat Control as seen in the sequential figures below.

Figure 3: Blinds Control Interface. [Left] A display denotes a window where the paper with vertical lines is used to communicate the orientation of the blinds to the participants. The three buttons on the holographic component denote, respectively: *rotate left*, *open/close window*, *rotate right*. [Right] The user holds down the rotate right button (and it turns black).

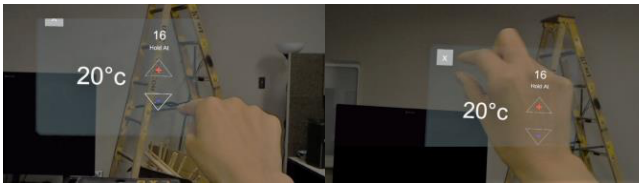

Figure 4: Thermostat Control Interface. [Left] The user selects the *up* button to increase the temperature and *down* button to decrease the temperature. [Right] The user selects the *close* button to close the thermostat control.

3.3 Media Control

Media control is selected as a scenario because a lot of home-based IoT tasks that emerge from initial ideation are based on some sort of media transfer such as capturing/sharing/playing media. The two tasks in this scenario are: Speaker Control and Video Control as seen in the sequential figures below.

Figure 5: Speaker Control Interface. [Left] The speakers can be imagined as embedded on the ceiling. The user selects a holographic *audio thumbnail* to play it. [Right] The *progress bar* replaces the *audio thumbnail* when that is selected. Here, the user selects the *pause* button to pause the song. When the song ends, the thumbnail reappears.

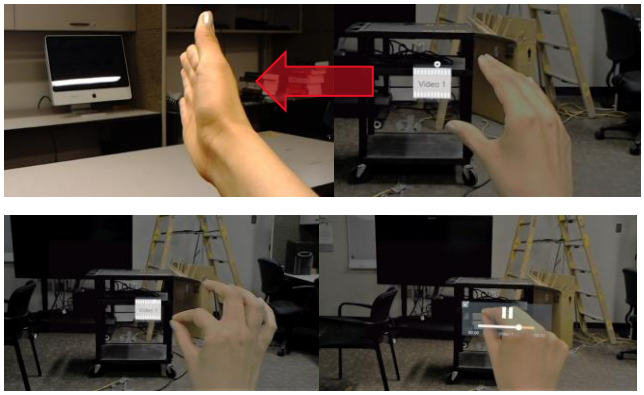

Figure 6. Video Control Interface. [Top] There are two displays identified to the user (one on their left and another on their right). They are asked to play the video on Display 1. [Bottom Left] The user picks the *video thumbnail* to play. [Bottom Right] The *progress bar* replaces the *video thumbnail* when that is selected. The user closes the video player to stop playing.

3.4 Following a Workflow

Following a workflow is selected as a scenario as accessing info in a spatial context is deemed a major advantage of AR. This scenario captures examples where users need their hands for executing the work. The two tasks in this scenario are: Cooking a Recipe and Fixing a Boombox as seen in the sequential figures below.

Figure 7: Cooking a Recipe Interface. [Left] The user reads the instruction, adds cooking oil, and selects the *next* button to go to the next step. In the same way, the user proceeds through the remaining steps until there are no more steps. [Right] One of the steps require the user to go to a different counter to add flour (to simulate a kitchen scenario).

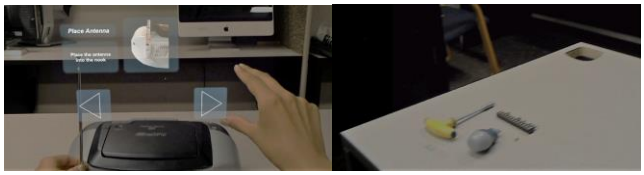

Figure 8: Fixing a Boombox Interface. [Left] The user reads the holographic instruction, places the antenna on the boombox, and selects the *next* button. In the same way, the user progresses through the instructions until the last step. [Right] Image of the physical tools that are used by the user in this task.

4 Task Design Choices

Interaction for each task was designed to fit the usage scenario. For example, connecting a computer to a print queue is designed to test user interaction with a menu system: it could have been designed to fit in the media control scenario, but was specifically designed with a multitiered menu to test the scenario. The same applied to all tasks. Some tasks could overlap in two or more scenarios if explored in further breadth. However, the point was to explore the four scenarios, so tasks were designed as such with focus on specific aspects of interaction. Moreover, design choices were influenced by current HoloLens interfaces as it supports both voice and gesture input. Thus, by explicitly asking the user for both a voice and gesture input round and taking inspiration from current designs of HoloLens and WIMP interfaces, we aimed to design in comparative terms for both voice and gesture input.

5 Rounds

There were 16 task/interaction combinations since all 8 tasks could be performed with either voice or gestural control.

Combo #	Scenario	Task (Context)	Control Method
1	Interact with a Menu System	Computer to Print Queue	gesture
2	Interact with a Menu System	Computer to Print Queue	voice
3	Interact with a Menu System	Lights Schedule	gesture
4	Interact with a Menu System	Lights Schedule	voice
5	Environmental Control	Blinds	gesture
6	Environmental Control	Blinds	voice
7	Environmental Control	Thermostat	gesture
8	Environmental Control	Thermostat	voice
9	Media Control	Speaker Control	gesture
10	Media Control	Speaker Control	voice
11	Media Control	Video Display	gesture
12	Media Control	Video Display	voice
13	Follow a Workflow	Cooking	gesture
14	Follow a Workflow	Cooking	voice
15	Follow a Workflow	Fixing a Boombox	gesture
16	Follow a Workflow	Fixing a Boombox	voice

Table 1: All possible 16 combinations of tasks. Half of the participants performed combo 1-5-9-13 for round one and 4-8-12-16 for round two. The other half performed combo 2-6-10-14 for round one and 3-7-11-15 for round two. Thus, half of the participants performed all scenarios with gestures first and with voice second. The other half did it vice versa.

Each participant was asked to take part in 8 combinations: where they performed each task only once with either voice or gestures. The study was broken down into 2 rounds. Round 1 sequentially consisted of: connecting computer to print queue, blinds control, speaker control, and cooking a recipe. Round 2 sequentially consisted of: setting a lights schedule, thermostat control, video control, and fixing a boombox. Half of the participants performed round 1 with voice and round 2 with gestures while the other half performed round 1 with gestures and round 2 with voice. This method reduced the impact of a learning effect, avoided biases, and ensured the testing of both control methods for all tasks. The breakdown of combinations and rounds can be observed in table 1.

6 Procedure

Participants were asked to fill out a pre-study questionnaire regarding their previous experience with controlling devices, AR, VR, and IoT. Next, they were fitted with the headset and provided instructions for round 1. They were given a prompt for each task and were given the freedom to interact in any way they wanted (within the specified method of voice or gesture) to accomplish the task at hand. After round 1 was completed, there was a semi-structured interview regarding their experience and the method of

control that they were asked to use for round 1. Next, round 2 was completed in the same way, and the same semi-structured interview took place for round 2. Lastly, a comparative semi-structured interview took place in which the focus was to reflect upon the pros and cons of each control method (voice vs. gesture) and to provide insight into their preference. The participants' interaction and interviews were videotaped using a phone camera.

ANALYSIS AND RESULTS

1 Analysis

The 16 participants were prompted for input 49 times (as each task included multiple sub-tasks) resulting in a total of 784 inputs. Of these, 392 were gestures and 392 were voice commands. We used open coding from grounded theory to analyze the data [24]. Initially, each of the 784 inputs were recorded on video and a short description was written for each of them in a table. Data was used for agreement rate calculations, defining a consensus set for voice commands and gestures, and forming design recommendations.

Tasks were designed organically, and this led to some referents having multiple proposals from the same user. For example, rotating blinds to the left and then to the right were considered as two proposals for the same referent. However, users were not explicitly asked for two diverging proposals, but some users did this i.e.) rotating a hand versus tapping a button. They were not asked to pick a favourite. For this reason, *Chen et al.'s* agreement rate was used [7]. This agreement rate accounts for multiple proposals by some or all participants (also for cases where a user does not pick their favourite proposal). It calculates the agreement rate for the proposal that is proposed by the highest number of people. It also accounts for the possibility of 0 in the case where no one agrees with each other. The formula:

$$AR = \frac{\max_{P_i \subseteq P} |P_i| - 1}{|P| - 1} \quad (1)$$

where $|P|$ is the number of participants, Pi is a set of participants who made proposal i, so $\max_{Pi \subseteq P} |Pi|$ is the number of participants who made the most popular proposal.

1.1 Gestures

12 unique gestures from 392 gesture inputs were initially identified. The most commonly used gesture for each referent was determined and the agreement rate was derived to see if participants agree with each other. Depending on whether agreement rates were high enough, a gesture set would be derived out of the initial 12 gestures.

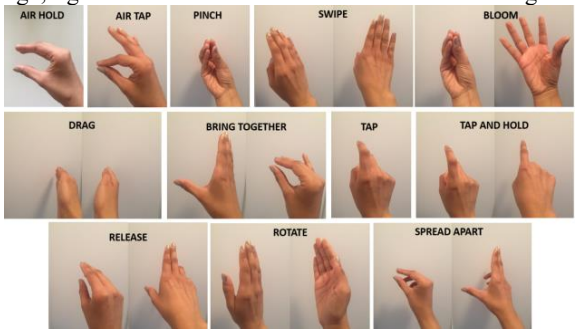

Figure 9: 12 unique gestures identified out of 392 gesture inputs. A combination of gestures was also used sometimes. A gesture set was narrowed down from these gestures.

1.2 Voice Commands

Voice commands are very flexible. We want to determine if there is a pattern that people use when they try to use a voice command for IoT interactions. We are inspired by grammars and attempt to develop a grammar for voice commands. Grammars that we want are simple and IGNORE sequence. Some components can be skipped. We heuristically state that a user is following the grammar when they use more than 50% of the components. Our method utilizes looking for commonalities in the voice commands that each participant used to accomplish the same result. We introduce a **voice command pattern template** for analyzing proposed commands. This should result in a pattern for the command. This was done for each time a voice command was required. An example of this is the "pick directional button/ specify direction" referent for the environmental control scenario.

P1	"move to left"
P2	"rotate to the left using the button"
P3	"rotate blinds to the left"
P4	"rotate the blinds to the left"
P5	"click"
P6	"spin left"
P7	"blinds rotate to the left"
P8	"rotate blinds to left"

Table A

P1	"move to left"
P2	"rotate to the left using the button"
P3	"rotate blinds to the left"
P4	"rotate the blinds to the left"
P5	"click"
P6	"spin left"
P7	"blinds rotate to the left"
P8	"rotate blinds to left"

Table B

Table 2: [Table A] Raw data from 8 participants that were tasked with changing the orientation of the blinds in the left direction. [Table B] Initially, the most common factor that was seen is the usage of "[to the] left".

P1	"move to left"
P2	"rotate to the left using the button"
P3	"rotate blinds to the left"
P4	"rotate the blinds to the left"
P5	"click"
P6	"spin left"
P7	"blinds rotate to the left"
P8	"rotate blinds to left"

Table A

P1	"move to left"
P2	"rotate to the left using the button"
P3	"rotate blinds to the left"
P4	"rotate the blinds to the left"
P5	"click"
P6	"spin left"
P7	"blinds rotate to the left"
P8	"rotate blinds to left"

Table B

Table 3: [Table A] Next, it was noted that the word "blinds" is used quite often. [Table B] Lastly, it was noted that verbs like "rotate" and "spin" or "move" were used various times.

Since 7 people used at least 2 of 3 criteria (≥ 1⁄2 of the criteria) as highlighted in the 3 colours, they were counted as using the same general formula of:

"rotate"/other verb ("spin"/ "move") + "blinds" + "[to the] left" (7 of 8 users)

where "rotate" was the most commonly used verb. It should be noted that sequence of components does not matter i.e.) P7 said "blinds" first.

Next, the voice commands from proposal 2 (orienting blinds to the right direction) were also analyzed in the same way. The formula derived from that was the following (where the same 7 people used the formula):

"rotate"/ other verb ("spin"/ "move") + "blinds" + "[to the] right" (7 of 8 users)

It was noted that "to the" are filler words only used by some users, and thus were dropped. From here, looking at both proposals (and removing any uncommon factors), a formula for the user-preferred command was derived. It was inferred that if a system accepts a command that takes in:

"rotate"/other verb ("spin"/ "move") + "blinds" + "left"/ "right" (in non-specific order) there is a high chance that it will coincide with what words the users would intuitively use. The total number of unique individuals who used this formula was **7 out of 8**.

If there was another task with the same referent, that was used for analysis as well. In this case, thermostat control also required a user to pick a directional button/specify direction when changing temperature. This was done with the other 8 people. The same analysis procedure was repeated, and **7 out of 8** people proposed:

"change"/ other verb ("set"/ "select") + words on UI for temperature value ("hold at") + numeric value

Next, these formulas were examined for the removal of uncommon factors (none exist here), and a super-formula was developed:

"rotate"/"change"/ other verb ("spin"/"move"/"set"/"select") + object ("blinds"/ UI words for temperature) + value ("left"/"right"/numeric value)

Next, *Chen et al.'s* agreement rate was used to determine consensus amongst participants [7]. The calculation was:

$$\frac{(Task\ 1 + Task\ 2\ users\ who\ align\ with\ pattern) - 1}{Total\ number\ of\ users - 1} = \frac{(7+7)-1}{16-1} = \frac{13}{15} = 0.87$$

2 Results

The analysis resulted in agreement rates that were high enough for us to specify a gesture and voice commands set. We created a table of appropriate commands and gestures for each referent as seen in table 4.

There was also a count of control method preferences (Table 5). At the end of the study, users were asked which method of control they prefer. This was done to determine whether there was one method that was overwhelmingly preferred. Users were also interviewed on their motivation of picking the preferred method. There was not enough difference between the 2 methods to be able to declare one as preferable. After users cited their reasons for their preferences, they were asked whether they would always use their preferred method or if it depends on what is to be done. Everyone said technology that offers a combination of both voice control and gesture control would be best.

Method of Preference	Reasons Given
Gestures 7 Users	• more subtle with others around/silent • more fun/entertaining • physicality • more immersive • predictability (knew what to do) • simplicity • consistency • seems cooler/can do fancy things with hands • exercise • feels weird to talk like talking to a person • voice commands feel variable • muscle memory/experience • more control/more flexibility in terms of controlling • avoiding language barriers
Voice 6 Users	• more comfortable because not HoloLens user • gestures felt more awkward as maybe not used to it • just more comfortable • less tedious on a big menu /don't want to click continuously on menus • works well for most situations • more fun • better experience • no need of moving hand • just say what you are thinking • more natural- does not feel like you are waiting for it to respond • like to talk • talking is easier • good for laziness • arm tiring out from gesturing on long/detailed menu
Mixture: not picking one 3 Users	• Can do complex, personalized things with gestures while voice is hands free. Also, don't want to set the gaze. • Voice may have problems with accents/dialects. While gestures may be a problem if someone doesn't have five full digits. Prefer voice for cognitive but hand motions for physical analogs (still prefer voice for recipe even though it is physical analog). • Prefer gestures for longer interactions (set up time/temperature) while voice for short interactions/ one word (to fix something)

Table 5: Users' method of preference and reasons given. Table

Referent	Most Popular Gesture agreement rate	Voice Command Template (Order of Components Does Not Matter) agreement rate
Interacting with Menu System		
Expand Menu	tap 0.53	"click"/"set"/ other verb ("hit"/"open"/"select"/"program") + object ("button"/"timer") 0.67
Pick Button	tap 0.67	exact words on button 1
Set Slider	drag 1	slider name + value 1
Set Toggle	tap 0.71	toggle value 1
Collapse Menu	tap 0.29 **low rate*	"collapse"/ other verb ("erase"/"click") + "menu" 0.86
Environmental Control		
Pick Directional Button/ Specify Direction	tap 0.53	"rotate"/"change"/ other verb ("spin"/"move"/"set"/"select") + object ("blinds"/ UI words for temperature) + value ("left"/"right"/numeric value) 0.87
Open Blinds Entirely	swipe 0.71	"open" + "blinds" 1
Close the Control Panel	tap 0.86	"close" + "menu"/"it" 0.57
Media Control		
Select Media to Play	tap 0.6	"play" + media file name 0.93
Select Button/Modify Playing Status	tap 0.67	"play"/"pause" / "stop" 0.80
Select Physical Display for Media	drag 0.57	"play" + media file name + "on" + display name 0.86
Following a Workflow		
Go to Next/ Previous Step (not enough previous users)	tap 0.6	"next" 1

Table 4: Referent table with the resulting voice command set and gesture set. Chen's agreement rates are included.

INTERPRETATION

The proposed gesture set and voice commands set is a partially subjective derivation from the data. For example, gestures that do not have high agreement scores are ignored. This is a partially subjective perspective via which we try to make sense of the data. In terms of voice commands, users generally opted to use the minimal amount of words needed to get the point across. Moreover, when words were available on the interface, users utilized them as a part of the command. There was also extensive use of verbs and names of physical and holographic objects.

The final gesture set consisted of 3 gestures: tap, drag, and swipe. Tap was the most common and was seen across all 4 scenarios. Users opted for this when there was some graphical component on the field of view that they could "select" such as holographic buttons, etc. The Drag was used when there was some physical movement implied: one being "moving" a graphical media thumbnail to a physical TV/Computer to play it. Swipe was used when there was some physical movement that was like the movement of an actual physical object: opening "blinds" like one would "draw curtains" with both arms. The user interface was designed via the WIMP paradigm standard (windows, icons, menus, pointer) [18]. This takes inspiration from current technology, and it is something that a future AR environment could look like. This design standard could have steered interaction such that users mostly tapped. However, this design provides an easy and consistent way for the users to interact. A relatively consistent user environment is provided where both implementation and interaction can be simple.

DESIGN RECOMMENDATIONS

Apart from the gesture and voice commands set, there were things like recurring or interesting behaviour that offer insights to implementers of AR-controlled IoT technology. Users also provided comments on the entire study experience. Some of the insights are intuitive or serve as confirmation for what we already know from other UI design contexts. As stated in the related works, there are other gestural elicitations that provide similar results for other technologies [5, 6, 14, 21, 22, 23, 28, 31, 32]. Similarly, there are also voice command and multimodal ones [4, 15, 25, 27]. We present insights for the context of both voice and gestures as related to using headset AR to control IoT devices.

Most recommendations were based on repetitive behaviour of at least half of the participants. Some observational recommendations were included even when a few individuals did them purely because they seemed to be insightful (as indicated).

Gesture round insights include the following. Technology should accommodate all 6 hand orientations with palm facing: *up, down, left, right, in,* and *out*. Furthermore, hand-switching and equal capabilities for both hands should be incorporated. The use of the pointer and thumb fingers should be prioritized, but the occasional use of the other three fingers can also be considered. Various degrees of freedom should be accommodated for considering gestural motion: *left-to-right, right-to-left, forward, down,* and *up*. It should also be considered that an entire arm may be used for input. Gestures should be designed as a means to bridging physical gaps if they exist i.e.) "sending" a holographic video thumbnail to a physical TV. Lastly, gestures may be influenced by physical world influences. i.e.) Swiping both arms apart to open blinds as if drawing physical curtains.

Voice round insights include the following. Technology should consider minimalistic commands. i.e.) A single *word* was used most often followed by *phrases* and then *sentences*. Furthermore, commands are influenced by words on the interface, by user interface components (i.e. "button") and by the physical environment/objects i.e. "blinds". Voice technology should also consider both natural conversation structure and previously learned command styles. i.e. "Hey Google", "Hey Tim" were both used. Moreover, commands should be designed with the consideration of natural behaviour and intuition. This could be even more important than voice commands that users propose. For example, users often said "ok next" or "done next" to go to the next step while following a workflow. Users perhaps expected the system to react just to the words "ok" or "done". Thus, natural language emerged in this example. Lastly, it could be inferred that "back" is a command that could align well with user intuition to go to the previous step. Only 3 users used the previous step and they all used "back".

Overall observations provided the following insight. Technology should be receptive to multimodal input. Users were often supplementing their gestures with commands and commands with gestures. It should also be noted that interaction with technology may change based on affordances. i.e. while kneading dough, users sometimes cleaned their hand to proceed using the system with the same hand or started using the other hand. Some voice command users cleaned hands before proceeding as well.

User comments provided the following insight regarding interaction and strategy. Users commented that current standard technology influences how they choose to interact with new technology. Additionally, when consecutive components are at a distance from one another (i.e. outside the same field of view), users wanted some visual cues like arrows, animations, or a glow should indicate where to look next. Lastly, most users see themselves as interacting consistently; however, some want to change their interactions (i.e. saying "pause the video" one day while saying "shut it off" the next day).

User comments provided the following insight regarding control method preferences. Everyone agreed that there are some cases where they would use their non-preferred method. Users who preferred gestures would use commands for: being more natural, efficiency (when hands are busy), being lazy, driving, shorter tasks, and multitasking. Users who preferred voice commands would use gestures for: gaming, tasks with multiple choices (like calling someone from a list of people), following a workflow with visual cues (like at work), when one must be quiet, not having to memorize object names (but rather just pointing at the object). Multi-modal interaction (being able to use both commands and gestures at the same time) is an option proposed by users. Favouring gestures or voice based on the scenario or situation (i.e. preferring voice when holding groceries) was also proposed by users. Based on user comments, no one control method is all-encompassing, and thus, a combination/choices are the best option.

Based on the study scope, design recommendations and result conclusions can be applied to relevant home-based IoT scenarios in which control would be administered with headset AR. Given the

participant demographic, our conclusions would likely be most applicable to users who engage with technology on a regular basis.

LIMITATIONS

There are some factors that may be limitations to the study.

1 Priming

As noted by *Chan et al.* [5, 6], priming may influence users’ proposals in elicitation studies. Thus, traditional priming was avoided. We did not provide example interactions for the user in order to avoid biased elicitation. However, there may be some way by which users can be better prepared without providing proposal examples. For example, performing a different task on AR headset technology may better prepare them for the study.

Additionally, the interface design is likely resulting in priming effects. We make a few assumptions. We assume that head-mounted displays will be more widely used in the future. We assume that user interfaces in the foreseeable future will be similar to what we currently use and have been using (WIMP (windows, icons, menus, pointers)) due to familiarity and (presumed) learnability. We design our interface based on WIMP and aim to understand how users would want to interact in AR using WIMP. We realize that for certain tasks, WIMP interfaces are suboptimal in AR applications but also realize that they are a good UI design choice for other system features. We stated that using WIMP could have influenced users to tap. This also applies to other gestures and voice commands. However, using WIMP or some variation of it is beneficial in that it provides guidance into system functionality. The user does not have to know what the system can provide before starting to use it. WIMP provides cues and increases detectability/findability of functionality. Although such menus may be “old-style”, they come with benefits. The choice of words written on the buttons or other user interface components seem to have influenced participants in the words that they propose in their voice commands. This may have been reduced by using more images instead of words on the UI.

2 Time & Rounds

Due to a time limitation, each user performed each task with either gesture or voice only once. Ideally, if each task was performed twice by each participant, they may be able to make an even better comparison between the two methods/ rounds.

3 Inter-Rater Reliability

Some other elicitations determine the inter-rater reliability. However, since we had a WIMP design, users were mostly pointing at things, and there was not much room for misinterpretation. Hence, only one researcher coded the results.

4 Recommendations/ Results Generalizability

A main limitation of the recommendations is that they can only be generalized to a certain extent as we do not have a random sample of the population of interest. Other limitations are situational/environmental in that they may apply more explicitly to a certain context based on a task. Also, participants were recruited via snowball sampling (non-random) and were predominantly from a computer science background. They were mostly in their early 20’s. Consequently, generalizations at best would be limited to users with similar backgrounds and age group. The number of participants was also limited to 16. The non- random sampling and limited number of participants do not allow for generalizations to all possible users.

5 Presence of Others

Participants may or may not interact with technology differently if someone is watching them. Some participants discussed using gestures or voice commands based on if they were by themselves. There was also discussion on how it feels to talk to a machine. Thus, factors such as being video-captured or performing in a research environment may influence user interaction.

FUTURE WORK

The summary and limitations highlight gaps that can be filled through future work. A study with a group of people from a different professional or cultural background may show different results in terms of interaction and preferences. Additionally, situational and environmental dependencies could be explored. For example, how do people interact when it is dark or when they are in a room full of people versus by themselves. Perhaps there would be one method that would stand out from the other if certain circumstances are tested. Furthermore, alternative elicitation methods could be explored to see if alternative insights arise. Examples being asking users to select gestures or voice commands based on a predetermined set of possible proposals or letting users make three proposals and asking them for their favourite. Moreover, the backend could also be fully implemented to see whether that makes a difference in interaction preferences (based on system delays, etc.) Additionally, the experiment could be carried out in a real home environment along with devices that are actually IoT-embedded rather than wizard-of-oz. Perhaps a more realistic environment and devices would allow for environmental affordances to also become highlighted. Other variables and capabilities could also be tested. For example, gaze input or controlling devices from another room could be probed on.

CONCLUSION

We identified a need for user involvement in the design of gestures and voice commands for AR-based IoT controls. We further established that a compare and contrast approach between voice commands and gestures would provide clarity on whether users prefer one method over another. We conducted an elicitation study based on home scenarios with the end users where 784 voice commands and gestural inputs were recorded. We contributed a novel way of analyzing voice commands called voice command pattern template that can further understanding regarding the types of voice commands users want to use. We presented the core findings including the agreement rates, gesture set, voice commands set, and a count of user preferences. Furthermore, we presented design recommendations based on observations and user comments. An extensive version of our work can be studied in the corresponding thesis [8]. Overall, our research can inform researchers and designers of AR and IoT controls regarding end user preferences and elicitation methodology.

REFERENCES

[1] AR Critic. What are the Magic Leap One Input Methods?. 2018. Retrieved: September 13, 2019 from https://arcritic.com/2293/what-are-the-magic-leap-one-input-methods/.

[2] Augmented Reality | Definition of Augmented Reality by Lexico. 2019. Retrieved: September 13, 2019 from https://www.lexico.com/en/definition/augmented_reality.

[3] Ceylan Beşevli, Oğuz Turan Buruk, Merve Erkaya, and Oğuzhan Özcan. 2018. Investigating the Effects of Legacy Bias.In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility (DIS '18).* ACM Press, New York, NY, 277-281. http://doi.acm.org/10.1145/3197391.3205449.

[4] Richard A. Bolt. 1980. “Put-that-there”: Voice and gesture at the graphics interface. In *Proceedings of the 7th annual conference on Computer graphics and interactive techniques (SIGGRAPH '80)*. ACM, New York, NY, USA, 262- 270. DOI=http://dx.doi.org/10.1145/800250.807503.

[5] Edwin Chan. 2017. *User-defined Single-hand Microgestures*. Master’s thesis. University of Calgary, Calgary, Canada. Retrieved: March 28, 2020 from: https://prism.ucalgary.ca/bitstream/handle/11023/3827/ucalgary_2017_chan_edwin.pdf?sequence=3

[6] Edwin Chan, Frank Maurer. 2016. User Elicitation on Single-hand Microgestures. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16).* ACM Press, New York, NY, 3403-3414. http://doi.acm.org/10.1145/2858036.2858589.

[7] Qing Chen. 2018.*Immersive Analytics Interaction: User Preferences and Agreements by Task Type*. Master’s thesis. University of Calgary, Calgary, Canada. Retrieved: March 28, 2020 from: https://prism.ucalgary.ca/handle/1880/106633

[8] Shreya Chopra. 2019. *Evaluating User Preferences for Augmented Reality Interactions for the Internet of Things.* Master’s thesis. University of Calgary, Calgary, Canada. Retrieved: March 28, 2020 from https://prism.ucalgary.ca/handle/1880/111455e

, Lester F. Ludwig. 1991. Multidimensional audio window management, International Journal of Man-Machine Studies, Volume 34, Issue 3, 1991, Pages 319-336, ISSN 0020-7373, https://doi.org/10.1016/0020- 7373(91)90023-Z.

[10] Janet Groeber. 2015. Mixed Reality. *Design : Retail*. 27, 4 (Apr/May 2015) 22.

[11] Alaric Hamacher, Jahanzeb Hafeez, Roland Csizmazia. 2019. Augmented Reality User Interface Evaluation – Performance Measurement of Hololens, Moverio and Mouse Input.*International Journal of Interactive Mobile Technologies (iJIM)*. 13, 03 (2019), 95-107. https://doi.org/10.3991/ijim.v13i03.10226

[12] Alexander G.Hauptmann, Paul McAvinney. 1993. Gestures with speech for graphic manipulation. *International Journal of Man-Machine Studies*. 38, 2 (Feb. 1993), 231-249. https://doi.org/10.1006/imms.1993.1011

[13] Help – Google Glass. 2019. Retrieved: September 13, 2019 from https://www.google.com/glass/help/.

[14] O. Hilliges D. Kim S. Izadi et al. "HoloDesk: direct 3d interactions with a situated see-through display" In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems.* pp. 2421-2430 2012.

[15] Helge Hüttenrauch, Mike Tate, Martin Böcker, Rosemary Orr, and Françoise Petersen. 2008. Standardising SPOKEN commands for mobile devices and services. In *Proceedings of the International Conference on Mobile Technology, Applications, and Systems* (Mobility '08). https://doi.org/10.1145/1506270.1506345

[16]IFTTT. 2019. Retrieved: September 13, 2019 from https://play.google.com/store/apps/details?id=com.ifttt.ifttt&hl=en.

[17] Internet of things | Definition of Internet of things by Lexico. 2019. Retrieved: September 13, 2019 from https://www.lexico.com/en/definition/internet_of_things.

[18] Stephen Kimani. 2009. WIMP Interfaces. *In: Ling Liu, M. Tamer Özsu (Eds.) Encyclopedia of Database Systems.* (2009), 3529-3533. https://doi.org/10.1007/978-0-387-39940-9_467

[19] Sang-Su Lee, Jeonghun Chae, Hyunjeong Kim, Youn-kyung Lim, and Kun-pyo Lee. 2013. Towards more natural digital content manipulation via user freehand gestural interaction in a living room. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing* (UbiComp '13), 617-626. https://doi.org/10.1145/2493432.2493480

[20] Cherlynn Low. Lenovo launches ThinkReality AR and VR headset for enterprises. 2019. Retrieved: September 13, 2019 from https://www.engadget.com/2019/05/13/lenovo-thinkreality-ar-vr-headset-hololens-2/#/.

[21] Meredith Ringel Morris. 2012. Web on the wall: Insights from a Multimodal Interaction Elicitation Study. In *Proceedings of the 2012 ACM international conference on Interactive tabletops and surfaces* (ITS '12), 95-104. https://doi.org/10.1145/2396636.2396651

[22] Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, M.C. Schraefel, and Jacob O. Wobbrock. 2014. Reducing legacy bias in gesture elicitation studies. *interactions*. 21, 3 (May/June 2014), 40-45. https://doi.org/10.1145/2591689

[23] Meredith Ringel Morris, Jacob O. Wobbrock, & Andrew D. Wilson. (2010). Understanding Users’ Preferences for Surface Gestures. Graphics Interface 2010, 261–268. Canadian Information Processing Society.

[24] Michael Muller. 2014. Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method. *In: Olson J., Kellogg W. (eds) Ways of Knowing in HCI*. (2014), 25-48. https://doi.org/10.1007/978-1-4939-0378-8_2.

[25] Michael Nebeling, Alexander Huber, David Ott and Moira C. Norrie. 2014. Web on the Wall Reloaded.In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces* (ITS '14), 15-24. https://doi.org/10.1145/2669485.2669497.

[26] Scott Orgera. 2019. The Top 100 Google Assistant and Google Home Commands. Retrieved: September 13, 2019 from https://www.lifewire.com/top-google-assistant-and-google-home-commands- 4158256.

[27] Ekaterina Peshkova, Martin Hitz, David Ahlström. 2016. Exploring User-Defined Gestures and Voice Commands to Control an Unmanned Aerial Vehicle.*Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*. (2016), 47-62. https://doi.org/10.1007/978-3-319-49616-0_5

[28] T. Piumsomboon A. Clark M. Billinghurst P. Kotzé G. Marsden G. Lindgaard et al. User-Defined Gestures for Augmented Reality. In *Proceedings of the International Conference on Human-Computer Interaction (INTERACT 2013)*. Heidelberg:Springer pp. 282-299 2013.

[29] Jaime Ruiz, Daniel Vogel. 2015. Soft-Constraints to Reduce Legacy and Performance Bias to Elicit Whole-body Gestures with Low Arm Fatigue. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 3347-3350. https://doi.org/10.1145/2702123.2702583

[30] Jim Tanous. 2019. The Lenovo ThinkReality A6 Is a Lightweight AR Headset for Businesses. Retrieved: September 13, 2019 from https://pcper.com/2019/05/lenovo-thinkreality-a6/.

[31] Radu-Daniel Vatavu, Jacob O. Wobbrock. 2016. Between-Subjects Elicitation Studies: Formalization and Tool Support. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), 3390-3402. https://doi.org/10.1145/2858036.2858228.

[32] Jacob O. Wobbrock, Meredith Ringel Morris, Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '09), 1083-1092. https://doi.org/10.1145/1518701.1518866.