User-Defined Interaction for Smart Homes: Voice, Touch, or Mid-Air Gestures?

Fabian Hoffmann[1], Miriam-Ida Tyroller[1], Felix Wende[1], Niels Henze[2]
University of Regensburg
Regensburg, Germany
[1]firstname.lastname@student.ur.de, [2]niels.henze@ur.de

ABSTRACT

Smart home appliances and smart homes, in general, are on the verge of ubiquity. Research and industry proposed a range of modalities, including speech, mid-air gestures, and touch displays, to control smart homes. While previous work designed for the individual modalities, it is unclear how they compare from a user-centered perspective. Therefore, we conducted an elicitation study that asked participants to propose commands using speech, mid-air gestures, and a touch display. Also, we asked participants to rate their suggestions and the modalities. The results show that using voice commands or a touch display is clearly preferred compared to the use of mid-air gestures. As we found high agreement scores for voice commands, our results also highlight the potential of elicitation studies for voice interfaces.

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**; **Ubiquitous and mobile computing**; **Graphical user interfaces**.

KEYWORDS

Smart home, voice control, display control, mid-air gestures

ACM Reference Format:
Fabian Hoffmann, Miriam-Ida Tyroller, Felix Wende, Niels Henze. 2019. User-Defined Interaction for Smart Homes: Voice, Touch, or Mid-Air Gestures?. In *MUM 2019: 18th International Conference on Mobile and Ubiquitous Multimedia (MUM 2019), November 26–29, 2019, Pisa, Italy*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3365610.3365624

1 INTRODUCTION & BACKGROUND

Smart home technologies are becoming increasingly popular. Besides enabling automating a variety of functions, they also enable to control entertainment systems, home appliances, room temperature, lighting, and access control. According to a recent market analysis[1], it is expected that nearly 30 million households in the US will adopt smart home technologies soon. In parts, this is enabled through a range of devices, such as Amazon's Echo and Google Home, that can be connected to the user's smart home devices.

Current commercial devices can be controlled through different interaction modalities. Inheriting the legacy of analog buttons, wireless physical buttons can be used to control light and temperature. Also, dedicated devices enable additional interaction modalities. A prime example is Amazon's smart virtual assistant Alexa which enables to control smart devices not only through speech but also through the user's phone and through interactive displays. At the beginning of 2019, over 100 million devices with Amazon's Alexa have been sold[2]. Media reports[3] suggests that there will be over 200 million installed smart speakers at the end of 2019.

Besides speech and touch UIs, previous research proposed further interaction modalities. Especially the use of mid-air gestures has been widely explored. Already Bolt's seminal work combined mid-air gestures with other modalities [5]. In recent years, the use of mid-air gestures has been popularized by entertainment technologies. Previous work used, for example, Nintendo's WiiMote [29] or the Leap Motion [37]. Mid-air gestures have been proposed for a large number of applications, including controlling music playback [17], interact with large displays [1], and smart TVs [31]. When designing gestures, Nielsen et al. concluded that technology-based approaches lead to awkward gestures without intuitive mapping towards functionality and systems which only work under strictly pre-defined conditions [24]. As an alternative, previous work presented human-centered approaches to design gestures [11, 24, 35], which have been coined guessability studies by Wobbrock et al. [35]. As demonstrated by a recent survey [34], guessability studies are the de-facto standard for designing gestures. The produced gesture sets are preferred by users [22] and also easier to remember [23].

To provide a consistent experience across the wide variety of future smart home devices, it is necessary to know which interaction modality provides the higher user experience and which one is preferred by users. The three most promising interaction modalities are speech, touch UIs, and mid-air gestures. All three modalities can be used to control smart homes. While previous work broadly explored the interaction with smart homes (e.g. [3, 14]) and also compared different modalities [32], it is still unclear which interaction modality is the most usable one and which is the one preferred by users.

In this paper, we compare the use of voice commands, touch UI, and mid-air gestures to interact with smart homes. Therefore,

[1]http://mordorintelligence.com/industry-reports/global-smart-homes-market-industry

MUM 2019, November 26–29, 2019, Pisa, Italy

ACM ISBN 978-1-4503-7624-2/19/11…$15.00
https://doi.org/10.1145/3365610.3365624

[2]https://www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp
[3]https://techcrunch.com/2019/04/15/smart-speakers-installed-base-to-top-200-million-by-year-end

we conducted an elicitation study to determine the preferred command for eleven representative smart home tasks. We also asked participants to rate the goodness, ease, enjoyment, and social acceptance, as well as their general fit for the task. The contribution of our work is threefold. 1) Quantitative and qualitative results consistently show that touch UIs and voice commands are preferred compared to mid-air gestures to control smart home tasks. 2) Receiving high agreement scores for voice commands, we show that elicitation studies are not only suitable to design gesture-based but also voice-based interfaces. 3) Analyzing the elicited commands for three modalities, we provide novel taxonomies that can help to analyze future elicitation studies.

2 METHOD

To compare the three interaction modalities touch UI, speech and mid-air gestures, we conducted an elicitation study to determine appropriate commands for each modality. We followed the process introduced by Wobbrock et al. [36] and refined by Vatavu and Wobbrock [33]. Furthermore, we asked participants to rate the commands for each modality.

We used recent market analysis to determine representative tasks[4]. The smart home market can be divided into the six categories *home entertainment, smart household appliances, energy management, networking and control, comfort and light* and *building security*. We excluded *networking and control* as it typically does not require explicit actions. Except for building security, we selected two common tasks for the remaining categories. We selected three building security tasks because of its much larger market share. The five categories with their assigned tasks are listed in Table 1.

After explaining the purpose and the procedure of the study, participants were asked to fill a consent form and a demographic questionnaire. Participants were asked to propose commands for each modality. We randomized the order of the modality and the order of the tasks within each modality to reduce sequence effects. As proposed by Wobbrock *et al.* [36] the eleven tasks were illustrated through pictures (see Figure 1 for an example), which showed the state before and after issuing the command. We also explained the tasks verbally to ensure a consistent understanding. We encouraged participants to explain their choices using thinking-aloud. After each task, participants rated their suggestion's goodness, ease, enjoyment and social acceptance on 7-point Likert items. Participants' commands were captured through video and audio recordings. For the display interaction, participants were additionally offered to sketch the desired interface using pen and paper.

[4]https://www.statista.com/outlook/279/100/smart-home/worldwide

Table 1: Categories with their assigned tasks

Category	Task
Home entertainment	1. Increase the volume of the music 2. Turn on the next TV channel
Household appliances	3. Start multi-colored wash at 60 degree 4. Turn off the oven
Energy management	5. Increase the room temperature 6. Open the shutters
Comfort and light	7. Turn on the light 8. Dim the light
Building security	9. Close the window 10. Lock the front door 11. Turn on the security camera

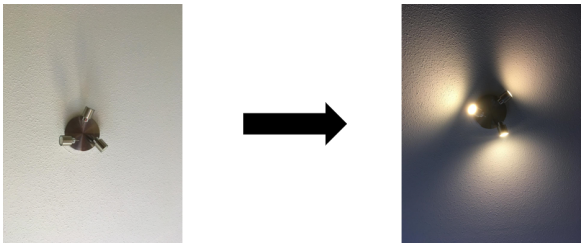

Figure 1: Example for how the tasks where illustrated to the participant (Task 7: *turn on the light*)

After completing all tasks with all modalities, we asked participants to rate the suitability of each modality for the eleven tasks on a 7-point Likert item. They were asked to do this independently of their suggestions. Finally, we conducted a semi-structured interview to explore the motivation of the participants for each choice and allow them to discuss the efficiency, simplicity, naturality, desirability, and enjoyment of the interaction modalities similar to the elicitation study on foot gestures by Felberbaum *et al.* [7]. In total, the study took about an hour per participant.

13 participants (7 female) took part in the study. We recruited them through social networks and personal contacts. Their average age was 33.5 (SD = 15.1). Most participants were students with different subjects. All participants at least heard of smart homes and were familiar with touchscreen-interaction. Ten participants were familiar with both voice control and touchscreens to control other devices, but only one performed mid-air gestures for interaction yet. Seven participants own smart home devices such as Google Home, Amazon Alexa, smart TVs or lamps and use them frequently. No participant owned a fully integrated smart home system.

3 RESULTS

With 13 participants and eleven tasks, we collected 143 suggestions for each of the three modalities, resulting in a total of $13*11*3 = 429$ suggestions. We collected video and audio recordings, subjective ratings of the suggestions, qualitative observations and an assessment on the modalities for each task. We derived taxonomies for each modality, user-defined sets of voice commands, display interactions, and mid-air gestures. Furthermore, we compared the modalities using quantitative and qualitative analysis.

3.1 Command Taxonomies

As the first step, we constructed a taxonomy for each modality. We adopted and adapted the dimensions by Wobbrock *et al.* [36], Ruiz *et al.* [27] and Dingler *et al.* [6]. Additional dimensions were developed when appropriate.

3.1.1 Voice Commands. The participants suggested 43 unique voice commands. We used the dimensions by Wobbrock *et al.* [36] and Ruiz *et al.* [27] and adapted them for voice commands. We accordingly classified the voice commands along five dimensions: *nature*,

form, *flow*, *context* and *complexity*. Within each dimension are multiple categories, shown in Table 2.

The *nature* dimension comprises *action* voice commands which state the action to perform. An exemplary voice command is "increase temperature". *State* voice commands describe the desired condition of a device. For example, a *state* voice command is "cameras on" to start camera surveillance. The *form* dimension describes how many words are used in the voice command and if they have the structure of a full sentence. A *single word* command can be "next" to get to the next TV channel. *Two words* voice commands mostly consist of the device to be controlled and an action or state. *More words* commands are similar to *two words* but use additional filler words. Finally, voice commands that are complete sentences were classified as *sentence*. The *flow* dimension categorizes if a device responds after or while the user acts. A voice command is *discrete*, when a device performs the command after the participant stopped talking. A *continuous* starts an action with a command and stop the ongoing action with another command. The *context* dimension describes, if the voice command requires a specific context or can be performed independently. For example saying "turn off" to turn off the oven is *in-context*, whereas "oven off" is considered *no-context*. The *complexity* dimension describes if the voice command consists of a single command or is a composition of multiple commands. A *compound* voice command can be decomposed into *simple* voice commands.

3.1.2 Display Interaction. Participants suggested 61 unique display interactions. The suggestions generally fit into one of two categories. They either described GUI elements or touch gestures. Therefore, we derived two separate taxonomies. The dimensions and categories for GUI elements were inspired by Wobbrock *et al.* [36] and Ruiz *et al.* [27]. We classified suggestions containing GUI elements along the three dimensions *form*, *elements*, *flow* (see Table 3).

The *form* dimension describes if the command consists of a single action that results in a response or consists of a selection of the desired action that is started through another element. A *direct action* would be a single click on a button to turn the light on. Selecting a washing program and starting it with an additional "start"-button would be considered *selection & confirmation*. The *elements* dimension describes, the type of GUI elements. *Single*

Table 2: Taxonomy for voice commands

Nature	Action	Voice command states the action to perform
	State	Voice command describes the desired condition
Form	Single word	Voice command consists out of a single word
	Two words	Voice command consists out of two words
	More words	Voice command consists out of more words without sentence structure
	Sentence	Voice command uses sentence structure
Flow	Discrete	Response occurs *after* the user acts
	Continuous	Response occurs *while* the user acts
Context	In-context	Voice command requires specific context
	No-context	Voice command does not require specific context
Complexity	Simple	Voice command consists of a single voice command
	Compound	Voice command can be decomposed into simple voice commands

Table 3: Taxonomy for touch UIs using Graphical User Interface (GUI) elements

Form	Direct Action	Interaction directly leads to the action
	Selection & Confirmation	Selection and activation of the action through different element
Elements	Single clickables	Includes one or more one-click elements (button, checkbox, etc.)
	Slider	Includes one or more sliders
	Rotation	Includes one or more rotational elements
	Text & number entry	Includes one or more fields for text or numbers
	Symbolic	Includes one or more special symbolic elements
Flow	Discrete	Response occurs *after* the user acts
	Continuous	Response occurs *while* the user acts

Table 4: Taxonomy for touch UIs using touch gestures

Nature	Symbolic	Gesture visually depicts a symbol
	Physical	Gesture acts physically on objects
	Metaphorical	Gesture indicates a metaphor
	Abstract	Gesture-referent mapping is arbitrary
Form	Static pose	Hand pose is held in one location
	Dynamic pose	Hand pose changes in one location
	Static pose and path	Hand pose is held as hand moves
	Dynamic pose and path	Hand pose changes as hand moves
	One-point touch	Static pose with one finger
	One-point path	Static pose and path with one Finger
Binding	Object-centric	Location defined with respect to object features
	World-dependent	Location defined with respect to world features
	World-independent	Location can ignore world features
	Mixed dependencies	World-independent plus another
Flow	Discrete	Response occurs *after* the user acts
	Continuous	Response occurs *while* the user acts

Table 5: Taxonomy for mid-air gestures

Nature	Symbolic	Gesture visually depicts a symbol
	Physical	Gesture imitates a physical action
	Metaphorical	Gesture indicates a metaphor
	Abstract	Gesture-referent mapping is arbitrary
Flow	Discrete	Response occurs *after* the user acts
	Continuous	Response occurs *while* the user acts
Context	In-context	Gesture requires specific context
	No-context	Gesture does not require specific context
Interaction	One hand	Gesture was performed with one hand
	Two hands	Gesture was performed with two hands
Dimension	Single-Axis	Motion around a single axis
	Tri-Axis	Translational hand motion or wrist rotation
	Six-Axis	Translational hand motion and wrist rotation
Position	Flat hand	Gesture started with flat hand
	Open hand	Gesture started with open hand
	Closed hand	Gesture started with closed hand (fist)
	Single finger	Gesture started with one stretched finger
	Two fingers	Gesture started with two stretched fingers
	More fingers	Gesture started with three or four stretched fingers
Movement	No movement	No change in finger position
	Movement	Change in finger position
Complexity	Simple	Gesture consists of a single gesture
	Compound	Gesture can be decomposed into simple gestures

clickables are elements which cause a response after a single click, such as buttons or checkboxes. The category *slider* is chosen, when one or more sliders are included in the GUI, for example, to change the volume of music. Rotational elements such as the selection of washing programs through a visualized rotary knob is categorised as *rotation*. *Text & number entry* includes options to enter content with a keyboard. Special symbolic elements, such as dragging wood into a fire to increase the room temperature are considered *symbolic*. The *flow* dimension again categorizes if a device responds after or while the user acts. A *discrete* command would be, pressing a "close window" button and after which the command is executed. A *continuous* one would be, dragging the regulator on a slider to adjust the volume of music simultaneously.

For touch gestures, we used Wobbrock *et al.* [36] taxonomy of surface gestures. Thus, Touch gestures were classified along the four dimensions *form*, *nature*, *binding*, *flow* (see Table 4).

3.1.3 Mid-Air Gestures. Participants suggested 55 unique mid-air gestures. We adopted the dimensions *nature* (small adjustment at category *physical*), *flow*, *context* and *complexity* and their corresponding categories from Wobbrock *et al.* [36] and Ruiz *et al.* [27]. *Dimension* from Ruiz *et al.* [27] was adapted to our needs. *Interaction* is inspired by Dingler *et al.* [6]. We further extended the taxonomy by the two dimensions *position* and *movement*.

The *nature* dimension includes *symbolic* mid-air gestures, which visually depict symbols. An example is drawing an "X" into the air to turn off the oven. *Physical* mid-air gestures imitate actions such as locking a door with a key, by rotating the wrist with a fist. An example of a *metaphorical* gesture is wrapping the arms around and rubbing the body, to indicate freezing to raise the temperature in the room. Mid-air gestures that did not fit into any of these three categories are classified as *abstract*. The *flow* dimension is the same as in the other taxonomies. A gesture is*discrete* if the response occurs after the movement and *continuous* when the response occurs during the movement. The *context* dimension describes if the mid-air gesture requires a specific context or can be performed independently. For example, making a horizontal hand movement to change the TV channel is *in-context*, whereas pointing at the TV and then performing the hand movement is considered *no-context*. The *interaction* dimension describes if one hand or both hands are used. The *dimension* is used to describe the number of axes involved in the movement of the hand. Some gestures, such as just rotating the wrist happen along a single axis. The translation of a hand or a rotational motion from the wrist is considered *tri-axis*. The combination of those two movements is classified as *six-axis*. Finger movement is not considered in this dimension but in the *movement* dimension. The *position* dimension describes the finger position at the beginning of the mid-air gesture. The difference between *flat hand* and *open hand* is, that the fingers are together at *flat hand* and spread at *open hand*. The *movement* dimension distinguishes if the participant includes relevant finger movement (*movement*) or no finger movement (*no movement*). The *complexity* dimension describes, as for voice commands, if the gesture consists of a single or a composition of gestures. *Compound* gestures can be decomposed into *simple* ones.

Table 6: The most common voice commands for each task

Task	German	English	Frequency
1	lauter	louder	53.8 %
	Musik lauter	music louder	30.8 %
2	Nächster Kanal	next channel	46.2 %
	weiter	next	23.1 %
3	Buntwäsche 60°	multi-colored wash 60°	30.8 %
	Starte Buntwäsche 60°	start multi-colored wash 60°	15.4 %
4	Ofen aus	oven off	84.6 %
	ausschalten	turn off	7.7 %
5	Wärmer	warmer	30.8 %
	Raumtemperatur erhöhen	increase room temperature	30.8 %
6	Rollladen öffnen	open roller shutter	69.2 %
	Rollladen auf	roller shutter up	23.1 %
7	Licht an	light on	100.0 %
8	Licht dimmen	dim light	30.8 %
	Licht dunkler	light darker	15.4 %
9	Fenster schließen	close window	53.8 %
	Fenster zu	window closed	46.2 %
10	Haustür absperren	lock front door	61.5 %
	Tür zu	door closed	23.1 %
11	Kamera an	camera on	61.5 %
	Kamera einschalten	switch on camera	23.1 %

Table 7: Average agreement scores and fitness ratings

	Agreement Score	Fittness
Voice	M=0.41, SD=0.26	M=5.7, SD=1.4
Display	M=0.13, SD=0.06	M=6.0, SD=1.4
Gestures	M=0.19, SD=0.21	M=3.9, SD=1.9

3.2 Interaction Sets

After developing the taxonomies, we classified the suggested commands for each modality using the corresponding taxonomy. For

Table 8: The most common display interactions for each task

Task	Display Interaction	Frequency
1	vertical slider	23.1 %
	horizontal slider	23.1 %
2	button with arrow to the right	46.2 %
	horizontal swipe	15.4 %
3	button with washing program + button with 60 degree	23.1 %
	(all others)	7.7 %
4	button with "off"	38.5 %
	vertical slider	15.4 %
5	vertical slider	30.8 %
	button with arrow to the top	23.1 %
6	button with arrow to the top	46.2 %
	horizontal slider	15.4 %
7	button with "on/off" symbol	38.5 %
	button with "on"	23.1 %
8	vertical slider	46.2 %
	(all others)	7.7 %
9	button with window and "close"	23.1 %
	button "close window"	23.1 %
10	button with lock	23.1 %
	button with key	15.4 %
11	button with camera	30.8 %
	button with "on/off" symbol	15.4 %

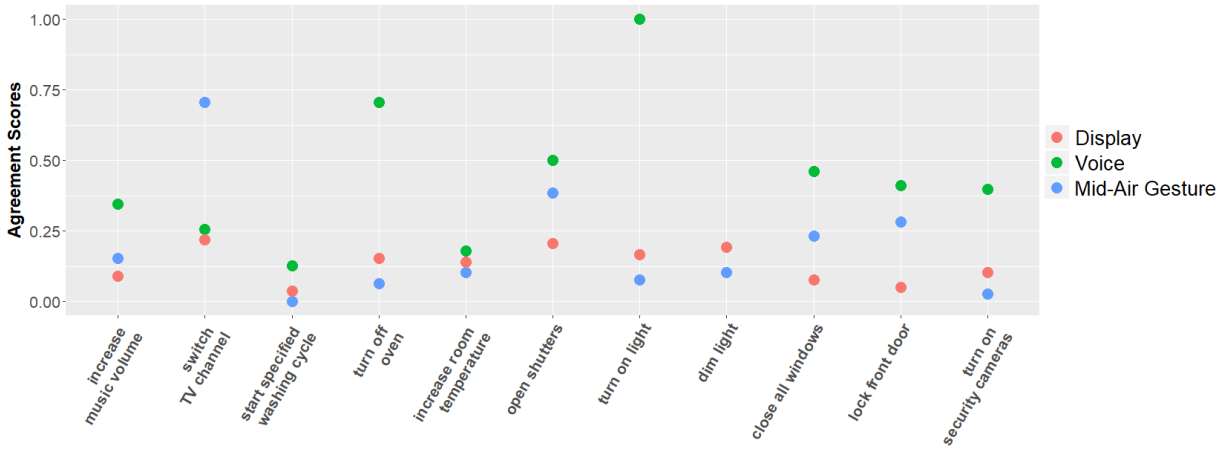

Figure 2: Agreement scores for each task with the three interaction modalities

each task and modality, we grouped identical suggestions. The group with the largest size was chosen to be the representative for the corresponding task and modality. To determine the degree of consensus among the participants, we computed the *agreement score* $A_{m,t}$ (Equation 1) proposed by Vatavu and Wobbrock [33].

$$A_{m,t} = \frac{|P_{m,t}|}{|P_{m,t}|-1} \sum_{P_i \subseteq P_{m,t}} \left(\frac{|P_i|}{|P_{m,t}|}\right)^2 - \frac{1}{|P_{m,t}|-1} \quad (1)$$

In equation 1, m is one of the three modalities, t is a task from the set of tasks T, $P_{m,t}$ is the set of suggestions for m and t, and P_i is a subset of identical suggestions from $P_{m,t}$. *Increase the volume of the music*, for example, received four groups of identical voice commands with sizes of 7, 4, 1 and 1. Accordingly, the agreement score for *increase the volume of the music* using voice control is:

$$A = \frac{13}{12}\left(\left(\frac{7}{13}\right)^2 + \left(\frac{4}{13}\right)^2 + \left(\frac{1}{13}\right)^2 + \left(\frac{1}{13}\right)^2\right) - \frac{1}{12} = 0.346 \quad (2)$$

After classifying and grouping all commands, we used the group sizes to compute the respective agreement scores (see Table 7 and Figure 2). A one-way repeated measure ANOVA revealed a significant difference between the modalities ($F = 6.2$, $p = 0.008$). Pairwise comparison with Bonferroni correction shows that the average agreement score for voice control was significantly higher than for display control ($p = 0.017$).

We derived a set of user-defined voice commands for the eleven tasks. Table 6 shows the most common and the second most common phrases with their respective frequency. As the study was conducted in German, we also provide English translations. We repeated the procedure for the display controls and the mid-air gestures. The respective sets are shown in Table 8 and Table 9.

3.3 Comparing the Modalities

Using participants' ratings, we compared the interaction modalities. Figure 3 shows how fitting participants consider the modalities for each task. On average display control received the highest ratings, followed by voice control and mid-air gestures (see Table 7). A two-way repeated measure ANOVA revealed a significant effect of the modalities ($F = 28.6, p < 0.001$) and task ($F = 3.2, p = 0.023$) on the ratings. There also was a significant interaction between the modalities and tasks ($F = 5.9$, $p < 0.001$). Pairwise comparison of the modalities using Bonferroni correction showed that voice control ($p = 0.001$) and display control ($p < 0.001$) are rated significantly higher than mid-air gestures.

Participants also rated the goodness, ease, enjoyment and social acceptance of their suggestions using 7-point Likert items (Table 10). For goodness, ease, and enjoyment 7 is the best score while for social acceptance 1 is the best score. While on average voice and display control were similarly rated, mid-air gestures received the worst scores for all measures.

3.4 Qualitative Results

We analyzed the semi-structured interview using a lightweight qualitative analysis. We randomly selected three interviews that were coded by three authors. We compared the resulting codes and

Table 9: The most common mid-air gesture for each task

Task	Display Interaction	Frequency
1	one hand, vertical motion	38.5 %
	one hand, hand rotation	15.4 %
2	one hand, horizontal motion	84.6 %
	(the two others)	7.7 %
3	(all different)	7.7 %
4	one hand, horizontal motion	23.1 %
	both hands, cross in front of body	15.4 %
5	one hand, vertical motion	23.1 %
	one hand, thump up, vertical motion	23.1 %
6	one hand, vertical motion	61.5 %
	both hands, vertical motion	15.4 %
7	one hand, flicking with the finger	23.1 %
	one hand, open closed hand	15.4 %
8	one hand, vertical motion	30.8 %
	one hand, hand rotation	15.4 %
9	one hand, hand rotation	38.5 %
	one hand, pushing motion	30.8 %
10	one hand, hand rotation	53.8 %
	both hands, brought together in front of the body	15.4 %
11	one hand, open closed hand	15.4 %
	one hand, vertical motion	15.4 %

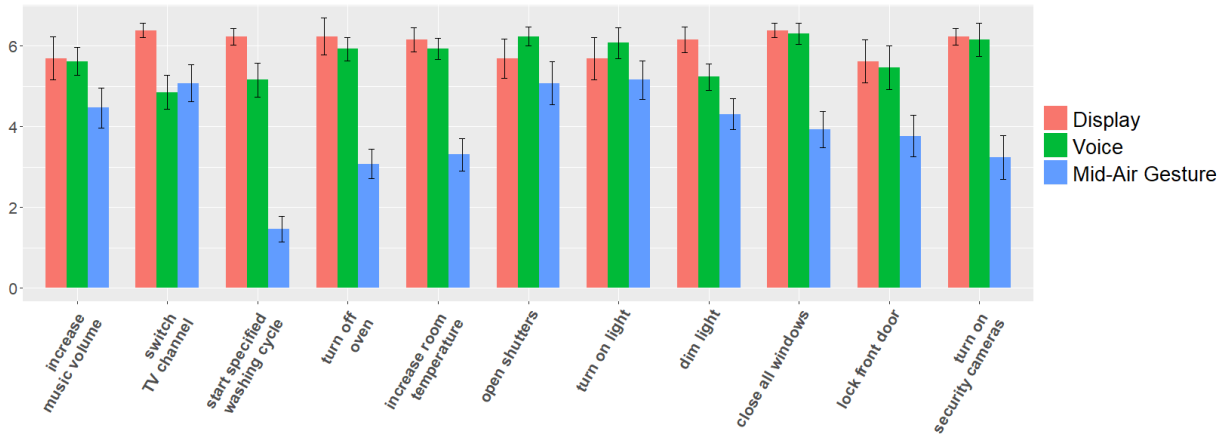

Figure 3: Modalities rated on a scale from 1="not fitting at all" to 7="very fitting". Error bars show the standard error.

Table 10: Ratings of the modalities across all tasks

	Goodness	Ease	Enjoyment	Social Acceptance
Voice	M=6.6, SD=0.7	M=6.3, SD=1.0	M=5.0, SD=1.4	M=1.7, SD=1.1
Display	M=6.3, SD=0.9	M=6.4, SD=1.0	M=5.2, SD=1.6	M=1.2, SD=0.8
Gesture	M=5.4, SD=1.3	M=5.7, SD=1.4	M=4.7, SD=1.6	M=2.4, SD=1.5

derived a unified codebook that was used to code all interviews. The interviews confirm the participants' preferences for voice and display control already expressed in their ratings of the modalities. Both modalities mainly received positive statements, while mid-air gestures received mostly negative comments. Participants explicitly stated why they would not choose mid-air gestures for most tasks. Expressed reasons include that they are counterintuitive, complicated to use, require long-term memory and lengthy actions. That mid-air gestures are not already used for other interfaces also contributed to participants negative assessment. On the contrary, that voice and display control are already used for other interfaces was positively highlighted by participants.

When specifically asked about choosing either voice or display control over the other, participants opinions were almost evenly split. Many stated their indifference over choosing one, which is mirrored in the almost equal quantitative ratings of the modalities. Instead, participants recommended implementing both modalities if this is not hindered through technical difficulties or costs as they prefer to have both modalities available. This mirrors the quantitative results that also do not justify clear recommendations of one modality over the other. When asked about future interfaces, participants suggested to include additional modalities, such as haptic feedback. They also recommended implementations that require as few interaction as possible. Instead, they would prefer autonomous assistance that learns about the users' intentions.

4 DISCUSSION & CONCLUSION

Quantitative and qualitative results clearly show participants' preferences for voice and display control over mid-air gestures. Voice and display control consistently received higher ratings compared to mid-air gestures. Similarly, participants consider voice and display control to be more socially acceptable than mid-air gestures. The results are supported by the qualitative feedback. Participants suggested that voice and display controls could be combined while they see little value in mid-air gestures. We conclude that future smart home systems should provide voice and display controls while support for mid-air gestures seems much less important.

Agreement scores for mid-air gestures are in line with previous work [2, 6, 36]. Probably due to the task's open nature, agreement scores for display interaction are fairly low. Agreement scores for voice commands are significantly higher than for the other modalities. Consequently, results suggest that elicitation studies, originally designed for gesture-based interfaces are a viable option to design voice interaction. This is also noteworthy as elicitation studies for voice commands are much easier to analyze than elicitation studies for gesture-based interfaces. Previous work warned that speech-based interfaces could increase gender bias (e.g. [10, 13, 25]). As empirical work on gender bias caused by speech-based interfaces is limited [8, 9], user-defined voice commands might also be away to prevent effects caused by developers' biases.

Not surprisingly, all participants had experience with touch-based interaction and most participants had experience with voice control. Clearly, the elicited commands are likely affected by legacy bias [21, 28]. Previous work proposed approaches to reduce legacy and performance bias [12, 21, 28]. Legacy bias can, however, also be advantageous as it helps users to adopt new interfaces [15] In line with Vogiatzidakis and Koutsabasis [34] we, therefore, also call for further analysis of elicited gestures and other commands to measure their usability and fatigue. In particular, we believe that future work should find ways to incorporate constraints induced by, for example, physiological characteristics [4, 16, 18], social acceptability [26, 30], effects on the interaction with others [20], and ways to communicate the commands to users [19]in elicitation studies.

In the study, we used a specific set of 11 actions and recruited a sample of a specific population. While we intentionally selected representative tasks, the eleven tasks are still only a subset of actions required to control smart homes. Nonetheless, future work that intends to design complete action sets can benefit from our work in two ways. First, our results provide a starting point by providing a subset of useful actions and suggest that future work should focus on two out of three modalities. Second, we provide novel taxonomies especially for voice commands which can help to analyze the results of future elicitation studies.

REFERENCES

[1] Christopher Ackad, Andrew Clayphan, Martin Tomitsch, and Judy Kay. 2015. An In-the-wild Study of Learning Mid-air Gestures to Browse Hierarchical Information at a Large Interactive Public Display. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 1227–1238. https://doi.org/10.1145/2750858.2807532

[2] Shaikh Shawon Arefin Shimon, Courtney Lutton, Zichun Xu, Sarah Morrison-Smith, Christina Boucher, and Jaime Ruiz. 2016. Exploring Non-touchscreen Gestures for Smartwatches. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3822–3833. https://doi.org/10.1145/2858036.2858385

[3] Andrea Bellucci, Andrea Vianello, Yves Florack, Luana Micallef, and Giulio Jacucci. 2019. Augmenting objects at home through programmable sensor tokens: A design journey. *International Journal of Human-Computer Studies* 122 (2019), 211 – 231. https://doi.org/10.1016/j.ijhcs.2018.09.002

[4] Joanna Bergstrom-Lehtovirta and Antti Oulasvirta. 2014. Modeling the Functional Area of the Thumb on Mobile Touchscreen Surfaces. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1991–2000. https://doi.org/10.1145/2556288.2557354

[5] Richard A. Bolt. 1980. Put-that-there: Voice and Gesture at the Graphics Interface. *SIGGRAPH Comput. Graph.* 14, 3 (July 1980), 262–270. https://doi.org/10.1145/965105.807503

[6] Tilman Dingler, Rufat Rzayev, Alireza Sahami Shirazi, and Niels Henze. 2018. Designing Consistent Gestures Across Device Types. In *Engage with CHI*, Regan Mandryk and Mark Hancock (Eds.). The Association for Computing Machinery, New York, New York, 1–12. https://doi.org/10.1145/3173574.3173993

[7] Yasmin Felberbaum and Joel Lanir. 2018. Better Understanding of Foot Gestures. In *Engage with CHI*, Regan Mandryk and Mark Hancock (Eds.). The Association for Computing Machinery, New York, New York, 1–12. https://doi.org/10.1145/3173574.3173908

[8] Florian Habler, Marco Peisker, and Niels Henze. 2019. Differences Between Smart Speakers and Graphical User Interfaces for Music Search Considering Gender Effects. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia (MUM 2019)*. ACM, New York, NY, USA, 7. https://doi.org/10.1145/3365610.3365627

[9] Florian Habler, Valentin Schwind, and Niels Henze. 2019. Effects of Smart Virtual Assistants' Gender and Language. In *Proceedings of Mensch Und Computer 2019 (MuC'19)*. ACM, New York, NY, USA, 469–473. https://doi.org/10.1145/3340764.3344441

[10] Charles Hannon. 2016. Gender and Status in Voice User Interfaces. *Interactions* 23, 3 (April 2016), 34–37. https://doi.org/10.1145/2897939

[11] Niels Henze, Andreas Löcken, Susanne Boll, Tobias Hesselmann, and Martin Pielot. 2010. Free-hand Gestures for Music Playback: Deriving Gestures with a User-centred Process. In *Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia (MUM '10)*. ACM, New York, NY, USA, Article 16, 10 pages. https://doi.org/10.1145/1899475.1899491

[12] Lynn Hoff, Eva Hornecker, and Sven Bertel. 2016. Modifying Gesture Elicitation: Do Kinaesthetic Priming and Increased Production Reduce Legacy Bias?. In *Proceedings of the TEI '16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction (TEI '16)*. ACM, New York, NY, USA, 86–91. https://doi.org/10.1145/2839462.2839472

[13] Alison Duncan Kerr. 2018. Alexa and the Promotion of Oppression. In *Proceedings of the 2018 ACM Celebration of Women in Computing (womENcourage '18)*. ACM, New York, NY, USA. https://womencourage.acm.org/2018/wp-content/uploads/2018/07/womENcourage_2018_paper_54.pdf

[14] Christine Kühnel, Tilo Westermann, Fabian Hemmert, Sven Kratz, Alexander Müller, and Sebastian Möller. 2011. I'm home: Defining and evaluating a gesture set for smart-home control. *International Journal of Human-Computer Studies* 69, 11 (2011), 693 – 704. https://doi.org/10.1016/j.ijhcs.2011.04.005

[15] Anne Köpsel and Nikola Bubalo. 2015. Benefiting from Legacy Bias. *interactions* 22, 5 (Aug. 2015), 44–47. https://doi.org/10.1145/2803169

[16] Huy Viet Le, Sven Mayer, Benedict Steuerlein, and Niels Henze. 2019. Investigating Unintended Inputs for One-Handed Touch Interaction Beyond the Touchscreen. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19)*. ACM, New York, NY, USA, Article 34, 14 pages. https://doi.org/10.1145/3338286.3340145

[17] Andreas Löcken, Tobias Hesselmann, Martin Pielot, Niels Henze, and Susanne Boll. 2011. User-centred process for the definition of free-hand gestures applied to controlling music playback. *Multimedia Systems* 18 (2011), 15–31. https://doi.org/10.1007/s00530-011-0240-2

[18] Sven Mayer, Perihan Gad, Katrin Wolf, Paweł W. Woźniak, and Niels Henze. 2017. Understanding the Ergonomic Constraints in Designing for Touch Surfaces. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, New York, NY, USA, Article 33, 9 pages. https://doi.org/10.1145/3098279.3098537

[19] Sven Mayer, Lars Lischke, Adrian Lanksweirt, Huy Viet Le, and Niels Henze. 2018. How to Communicate New Input Techniques. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction (NordiCHI '18)*. ACM, New York, NY, USA, 460–472. https://doi.org/10.1145/3240167.3240176

[20] Sven Mayer, Lars Lischke, Paweł W. Woźniak, and Niels Henze. 2018. Evaluating the Disruptiveness of Mobile Interactions: A Mixed-Method Approach. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 406, 14 pages. https://doi.org/10.1145/3173574.3173980

[21] 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 2014), 40–45. https://doi.org/10.1145/2591689

[22] Meredith Ringel Morris, Jacob O. Wobbrock, and Andrew D. Wilson. 2010. Understanding Users' Preferences for Surface Gestures. In *Proceedings of Graphics Interface 2010 (GI '10)*. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 261–268. http://dl.acm.org/citation.cfm?id=1839214.1839260

[23] Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of Pre-designed and User-defined Gesture Sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1099–1108. https://doi.org/10.1145/2470654.2466142

[24] Michael Nielsen, Moritz Störring, Thomas B Moeslund, and Erik Granum. 2003. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In *International gesture workshop*. Springer, 409–420. https://doi.org/10.1007/978-3-540-24598-8_38

[25] Chidera Obinali. 2019. The Perception of Gender in Voice Assistants. In *Proceedings of the Southern Association for Information Systems Conference*. 6. https://aisel.aisnet.org/sais2019/39

[26] Julie Rico and Stephen Brewster. 2010. Gesture and Voice Prototyping for Early Evaluations of Social Acceptability in Multimodal Interfaces. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10)*. ACM, New York, NY, USA, Article 16, 9 pages. https://doi.org/10.1145/1891903.1891925

[27] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-defined motion gestures for mobile interaction. In *Conference proceedings and extended abstracts / the 29th Annual CHI Conference on Human Factors in Computing Systems*, Desney Tan, Geraldine Fitzpatrick, Carl Gutwin, Bo Begole, and Wendy A. Kellogg (Eds.). ACM, New York, NY, 197. https://doi.org/10.1145/1978942.1978971

[28] Jaime Ruiz and 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)*. ACM, New York, NY, USA, 3347–3350. https://doi.org/10.1145/2702123.2702583

[29] Thomas Schlömer, Benjamin Poppinga, Niels Henze, and Susanne Boll. 2008. Gesture Recognition with a Wii Controller. In *Proceedings of the 2Nd International Conference on Tangible and Embedded Interaction (TEI '08)*. ACM, New York, NY, USA, 11–14. https://doi.org/10.1145/1347390.1347395

[30] Valentin Schwind, Niklas Deierlein, Romina Poguntke, and Niels Henze. 2019. Understanding the Social Acceptability of Mobile Devices Using the Stereotype Content Model. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 361, 12 pages. https://doi.org/10.1145/3290605.3300591

[31] Radu-Daniel Vatavu. 2012. User-defined Gestures for Free-hand TV Control. In *Proceedings of the 10th European Conference on Interactive TV and Video (EuroITV '12)*. ACM, New York, NY, USA, 45–48. https://doi.org/10.1145/2325616.2325626

[32] Radu-Daniel Vatavu. 2013. A Comparative Study of User-defined Handheld vs. Freehand Gestures for Home Entertainment Environments. *Journal of Ambient Intelligence and Smart Environments,* 5, 2 (March 2013), 187–211. https://doi.org/10.3233/AIS-130200

[33] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo (Eds.). ACM Press, New York, New York, USA, 1325–1334. https://doi.org/10.1145/2702123.2702223

[34] Panagiotis Vogiatzidakis and Panayiotis Koutsabasis. 2018. Gesture Elicitation Studies for Mid-Air Interaction: A Review. *Multimodal Technologies and Interaction* 2, 4 (2018). https://doi.org/10.3390/mti2040065

[35] Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A. Myers. 2005. Maximizing the Guessability of Symbolic Input. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1869–1872. https://doi.org/10.1145/1056808.1057043

[36] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined gestures for surface computing. In *CHI 2009 - digital life, new world*, Dan R. Olsen, Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott Hudson, and Saul Greenberg (Eds.). ACM, New York, NY, 1083. https://doi.org/10.1145/1518701.1518866

[37] Ionuţ-Alexandru Zaiţi, Ştefan-Gheorghe Pentiuc, and Radu-Daniel Vatavu. 2015. On free-hand TV control: experimental results on user-elicited gestures with Leap Motion. *Personal and Ubiquitous Computing* 19, 5-6 (2015), 821–838. https://doi.org/10.1007/s00779-015-0863-y