Considering Wake Gestures for Smart Assistant Use

Patryk Pomykalski
Lodz University of Technology
Lodz, Poland
p.pomykalski@ubicomp.pl

Mikołaj P. Woźniak
Lodz University of Technology
Lodz, Poland
mpwozniak@ubicomp.pl

Paweł W. Woźniak
Utrecht University
Utrecht, The Netherlands
p.w.wozniak@uu.nl

Krzysztof Grudzień
Lodz University of Technology
Lodz, Poland
kgrudzi@iis.p.lodz.pl

Shengdong Zhao
NUS-HCI Lab, School of Computing
National University of Singapore
zhaosd@comp.nus.edu.sg

Andrzej Romanowski
Lodz University of Technology
Lodz, Poland
androm@iis.p.lodz.pl

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.
© 2020 Copyright is held by the author/owner(s).
ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3383089

Abstract

Smart speakers have become an almost ubiquitous technology as they enable users to access conversational agents easily. Yet, the agents can only be activated using specific voice commands, i.e. a wake word. This, in turn, requires the device to constantly listen to and process sound, which represents a privacy issue for some users. Further, using the trigger word for the agent in a conversation with another human may lead to accidental triggers. Here, we propose using gestural triggers for conversational agents. We conducted gesture elicitation to identify five candidate gestures. We then conducted a user study to investigate the acceptability and effort required to perform the gestures. Initial results indicate that the snap gesture shows the most potential. Our work contributes initial insights on using smart speakers with ubiquitous sensing.

Author Keywords

gesture; gesture elicitation; smart assistant; smart speaker; gestural input

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** User studies;

Introduction

Smart speakers have become part of users' homes. With more and more units sold every year worldwide [2], many users interact with conversational agents embedded in smart speakers on an everyday basis. Past work reported that users tend to develop relationships with their smart speakers and tend to personify them [7].

Yet, despite the wide adoption of such devices, their use is not always hassle-free. Pyae and Joelsson [8] reported that voice interaction with smart speakers often posed many usability challenges. Later research suggested that users would benefit from multimodal interaction with smart assistants, e.g. by combining voice with touch as proposed by Bentley et al. [1]. Further work suggested that using wake words to activate smart speakers was problematic, inter alia due to the risk of undesired activation and suggested using alternate modalities for activation. McMillan et al. [4] showed that desirable experiences could be produced by employing gaze activation for smart speakers.

In this work, we further investigate alternative modalities for activating smart speakers by considering gestural input as an alternative. Anticipating that ubiquitous gestural sensing will soon be available to everyday users [3], we investigate what gestures could be used as wake gestures for smart assistant and if such gestures would be acceptable in home environments. To that end, we first conduct gesture elicitation and then conduct an initial comparative study.

Phase One: Gesture Elicitation

To explore possibilities for wake gesture, we first decided to ask users about suitable gesture designs. We conducted gesture elicitation to obtain a set of user-defined gesture candidates. We employed the method developed by Wobbrock et al. [11].

Study Design

We recruited 20 participants aged $M = 23$, $SD = 15.33$ (11 Male and 9 Female) using leaflets on campus and social media. All participants reported being proficient users of smart speakers and being fully familiar with smart speaker features.

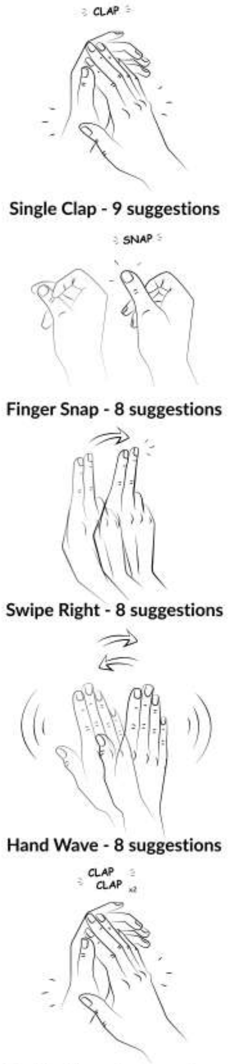
The study started with a quick introduction to the concept of a wake gesture. After obtaining informed consent, we asked the participant to propose at least five gestures suitable for triggering smart speakers. We chose to ask for a minimum of five gestures as a means of reducing legacy bias through production [6]. We encouraged commenting the gestures loudly. During the study, the participant remained seated on a chair with ample space around them. This allowed users to perform movements with different body parts or even whole-body gestures. Study sessions were video recorded, experimenters took notes describing particular gestures and transcribed participant comments.

During pilot studies, we observed that some participants assumed that only hand gestures were allowed. Therefore, during the study introduction, we explicitly defined the gesture as any physical, non-verbal action.

Results

The overall agreement rate, calculated according to the method by Vatavu et al. [10], representing the degree of consensus among participant in a single number, was $\mathcal{AR} = 0.032$. The users suggested a total of 100 gestures, 43 unique gestures were recorded. Figure 1 shows the top five gestures proposed and number of participants who suggested each of the gestures.

We also noted some unique gestures and full-body interactions. One user suggested grabbing their torso with both hands as a potential wake gesture. Others proposed a wink



Single Clap - 9 suggestions

Finger Snap - 8 suggestions

Swipe Right - 8 suggestions

Hand Wave - 8 suggestions

Double Clap - 6 suggestions

Figure 1: Top 5 most frequently suggested gestures with respective number of suggestions received.

of an eye, touching one's ear and drawing a circle with a finger in the air.

Phase Two: Initial Evaluation

Having observed that the agreement rate was low among the participants, we decided to additionally consider acceptability as an aspect that can determine a gesture's suitability as a wake gesture.

Study Design

Through an announcement on the department's website, $n = 11$ participants were recruited. The participants were aged $M = 23, SD = 4.06$ (6 Male and 5 Female). They were all familiar with smart speakers. We first obtained informed consent and introduced the participants to the concept of a wake gesture. The experimenter then assumed the role of a Wizard-of-Oz smart speaker. The participant was then provided with a task sheet, which was an open cloze test that required obtaining four pieces of information about a particular location. The task was to be completed for the five top gestures and five different locations, administered in a Latin-square-balanced order. Figure 3 presents a sample task. The participants thus completed a minimum of $4 \times 5 = 20$ gestures.

The gestures were sensed using a Leap Motion controller placed on a table in front of the participant. The Leap Motion hardware provided a reliable approximation of future sensing technologies. The experimenter could see a screen that showed successfully identified gestures and reacted with 'How can I help you?' each time a gesture was successfully detected. We introduced the sensing technology so that the Wizard-of-Oz study would offer an experience of simulating interacting with a computer and not a scripted conversation.

After the user completed the task using each gesture, we administered questionnaires related to the condition. We measured the anticipated effort of performing each of the gestures using the Rating Scale Mental Effort (RSME), developed by Zijlstra [12]. We chose this scale as it is was designed to be suitable for short actions and atomic tasks. We further investigated social acceptability using scales suggested by Rico and Brewster [9] and Montero et al. [5], asking about acceptability at home, performed alone, with friends and strangers (as these contexts are most typical for smart speaker use based on related work, e.g. [7]). When all conditions were completed, we asked the participants to complete a questionnaire ranking the gestures in order of preference. The participants then participated in a semi-structured interview where we asked about their anticipated experience of wake gestures. We asked users to reflect on the differences between the gestures and possible use in different contexts. The interview was saved on a voice recorder.

Results

We conducted a one-way ANOVA to investigate the effect of gesture performed on RSME. We found no significant effect, $F_{4,40} = 0.57, p > .05$. Figure 2 shows the results. Further, we conducted one-way ANOVAs on align-ranked data (cell frequencies were checked to align-rank transform requirements) to investigate the effect of gesture used on the perceived acceptability at home and in different audiences. We found no significant results. Table 1 presents the results.

Finally, we summed all the ranks (1 to 5) assigned to the gestures by the participants. The Snap and Swipe gestures received the lowest sum of scores, which indicates that they were most highly ranked. Detailed results are shown in Figure 5.

Semi-structured interviews conducted at the end of the

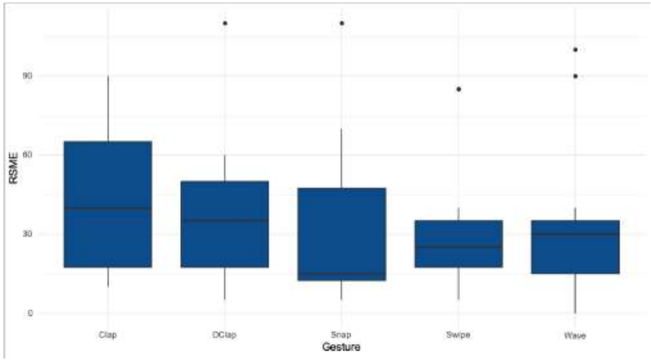


Figure 2: RSME measurements for the respective gestures. Error bars show standard error.

	$F_{4,40}$	p	M	SD
HOME	0.62	$> .05$	5.73	0.73
ALONE	0.17	$> .05$	5.78	0.53
FRIENDS	0.55	$> .05$	4.58	1.42
STRANGERS	0.86	$> .05$	2.40	1.59

Table 1: ANOVA results for align-rank-transformed data for acceptability scales.

Obtain the information needed to complete the gaps (marked —) from the smart speaker.
It would be nice to start the trip in the country's capital city —.
Matt and Anna are going to fly from Frankfurt, which offers connections with Venezuela's largest airport —.
Credit cards are not that popular in Venezuela. It would be nice to have some cash. The currency in Venezuela is called —.
The exchange rate is — for 1 USD.

Figure 3: A sample task to be completed by the participants in Phase Two.

study were transcribed verbatim. Two researchers then identified passages in the data that discussed wake gestures. We then used affinity diagramming to identify two main themes in the data: the social meaning of gestures and modality preferences.

The participants commented extensively on how wake gestures could be perceived by others in different contexts. They were cautious that some gestures could have unintended implications if other users were unaware that they were wake gestures. One participant explicitly stated their reservations:

> *People can misunderstand me if I use the gestures.*

Significant concern was also raised about performing audible gestures. Participants were worried that it may draw undesired attention and be disturbing to others. One participant explicitly pointed to sample situation:

> *Clapping is loud and draws attention. I would come from the other room if I heard someone clapping. I would think that something is going on.*

On the other hand, many users commented that a careful choice of gestures could alleviate many issues and reduce ambiguity. Another participant commented how a hand wave was not suitable as a wake gesture as it carried an inherently social meaning:

> *A hand wave is explicitly a gesture targeted at another person. It would confuse people in my surroundings.*

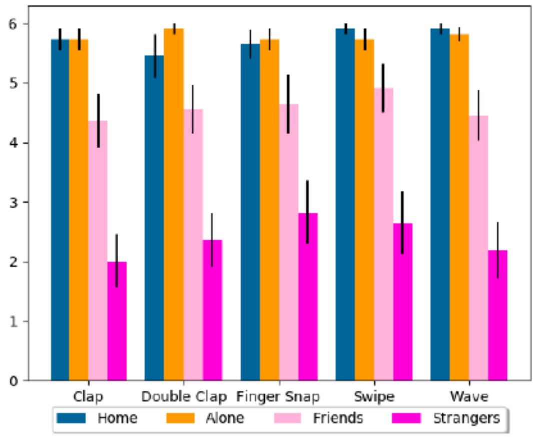


Figure 4: Acceptability scores for the five gestures studied. We analysed acceptance at home, when used alone, among friends and strangers. The questionnaire was based on previous work. [5, 9]

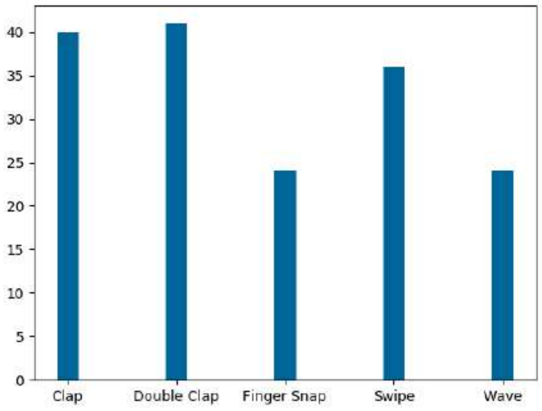


Figure 5: The sum of ranks obtained for the five gestures. The participants were asked to rank the gestures in order of preference. The lower the score, the more highly-ranked the gesture.

The interviews also showed that users welcome multi-modal possibilities for accessing their smart assistants. All the users reported that they would use a wake gesture. The possibility of using alternate modalities was appreciated:

> *The perfect situation would be if you could alternatively use voice or gestures.*

The participants also reflected on the fact that wake gestures offered the possibility to reduce ambiguity and prevent accidental triggers that may happen during conversations:

> *Gestures are potentially better at preventing accidentally using the system, during a conversation, for example.*

Users noted that gestural sensing instead of wake words reduced the need for constant voice recording:

> *In terms of my personal data privacy, using a wake gesture appears to be way more practical than using voice.*

One participant stated, that technology used by the system has a significant influence on perceived privacy of personal data:

> *If the system was detecting my movements, with some kind of radar, I would feel more secure using gestures. On the other hand if I were recorded by some vision-based camera, I would rather use my voice.*

Finally, users preferred the gesture modality if the smart assistant were to be used in public settings. A gesture was perceived as less disrupting and more socially acceptable when in public:

> *If you talk to yourself, people will be scared of you. I prefer the gestures, because they engage the environment around you less (...) this would make me more comfortable.*

Discussion

While our results did not show significant differences between the gestures, we gathered initial feedback that can inform future inquiries. First, our work shows the feasibility and potential for user acceptance for wake gestures. Combined with the developments in motion sensing, our work suggests that future research should investigate multimodal ways of activating smart assistants and allowing users to choose alternative activation modalities based on context of use.

Second, we identified an initial set of gestures that could be used as wake gestures through gesture elicitation. In the qualitative feedback gathered in the second phase of our work, we found that gestures that carry a social meaning may build a negative user experience when used as wake gestures. These findings suggest that future research should investigate gestures with little social meaning. Given how culture-specific many gestures are, culture is likely to be a key design factor in choosing wake gestures for particular user groups. This also highlights a limitation of our study — all 31 participants in the studies presented here were resident in Europe and had an European cultural background.

Third, our work suggests that utilising gesture delimiters, could enhance perceived privacy of personal data, given that underlying recognition technology is not based on visual recognition. Users seem to be less concerned about third parties potentially getting access to their motoric data (from motion sensors, EMG-based or microwave-based devices) than to being eavesdropped.

Future Directions

In this paper, we presented our initial inquiry into wake gestures. Our goal is to understand the possibilities of activating smart assistants using different modalities. Inspired by related work [4], in a future study, we plan to compare voice, gaze and gesture activation for smart assistants.
To that end, we will perform studies in a living lab where users will be able to control smart home, entertainments and communication systems using alternative modalities in different social contexts (groups of users). Our results suggest that users will benefit from changing the wake modality based on context and we plan to investigate if and how such a possibility can enhance interacting with smart assistants.

Conclusion

In this paper, we presented our initial inquiry into wake gestures for smart speaker and smart assistants. We first conducted gesture elicitation with $n = 20$ participants to identify five initial gesture candidates. In a subsequent Wizard-of-Oz study, $n = 11$ users performed a mock smart speaker task with the five gestures. The snap and wave gestures were most preferred by the participants. Qualitative feedback from the study showed that participants preferred gestures that did not have a social meaning. Further, all users identified benefits in using multiple modalities to wake smart assistants. Our work shows the feasibility of using wake gestures and provides a starting point for further inquiries into exploring modalities for smart speaker activation.

REFERENCES

[1] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article Article 91 (Sept. 2018), 24 pages. DOI: http://dx.doi.org/10.1145/3264901

[2] Bret Kinsella. 2020. Smart Speaker Sales to Rise 35% Globally in 2019 to 92 Million Units, 15 Million in China, Growth Slows. (2020). https://voicebot.ai/2019/09/24/smart-speaker-sales\-to-rise-35-globally-in-2019-to-92\-million-units-15-million-in-china-growth-slows/.

[3] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. 2016. Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar. *ACM Trans. Graph.* 35, 4, Article Article 142 (July 2016), 19 pages. DOI: http://dx.doi.org/10.1145/2897824.2925953

[4] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: Tama – a Gaze Activated Smart-Speaker. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article Article 176 (Nov. 2019), 26 pages. DOI:http://dx.doi.org/10.1145/3359278

[5] Calkin S. Montero, Jason Alexander, Mark T. Marshall, and Sriram Subramanian. 2010. Would You Do That? Understanding Social Acceptance of Gestural Interfaces. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI ’10)*. Association for Computing Machinery, New York, NY, USA, 275–278. DOI: http://dx.doi.org/10.1145/1851600.1851647

[6] 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. DOI: http://dx.doi.org/10.1145/2591689

[7] Amanda Purington, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel Hardman Taylor. 2017. "Alexa is My New BFF": Social Roles, User Satisfaction, and Personification of the Amazon Echo. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA ’17)*. ACM, New York, NY, USA, 2853–2859. DOI:http://dx.doi.org/10.1145/3027063.3053246

[8] Aung Pyae and Tapani N. Joelsson. 2018. Investigating the Usability and User Experiences of Voice User Interface: A Case of Google Home Smart Speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI ’18)*. ACM, New York, NY, USA, 127–131. DOI: http://dx.doi.org/10.1145/3236112.3236130

[9] Julie Rico and Stephen Brewster. 2010. Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’10)*. Association for Computing Machinery, New York, NY, USA, 887–896. DOI: http://dx.doi.org/10.1145/1753326.1753458

[10] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. Formalizing Agreement Analysis for Elicitation Studies: New Measures, Significance Test, and Toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1325–1334. DOI: http://dx.doi.org/10.1145/2702123.2702223

[11] Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-Defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. Association for Computing Machinery, New York, NY, USA, 1083–1092. DOI: http://dx.doi.org/10.1145/1518701.1518866

[12] Fred Zijlstra. 1993. Efficiency in Work Behavior: A Design Approach for Modern Tools. (01 1993).