Machine Body Language: Expressing a Smart Speaker's Activity with Intelligible Physical Motion

Mirzel Avdic
miavd18@cs.au.dk
Aarhus University
Aarhus, Denmark

Nicolai Marquardt
n.marquardt@ucl.ac.uk
University College London
London, UK

Yvonne Rogers
y.rogers@ucl.ac.uk
University College London
London, UK

Jo Vermeulen
jo.vermeulen@autodesk.com
Autodesk Research
Toronto, ON, Canada
and Aarhus University
Aarhus, Denmark

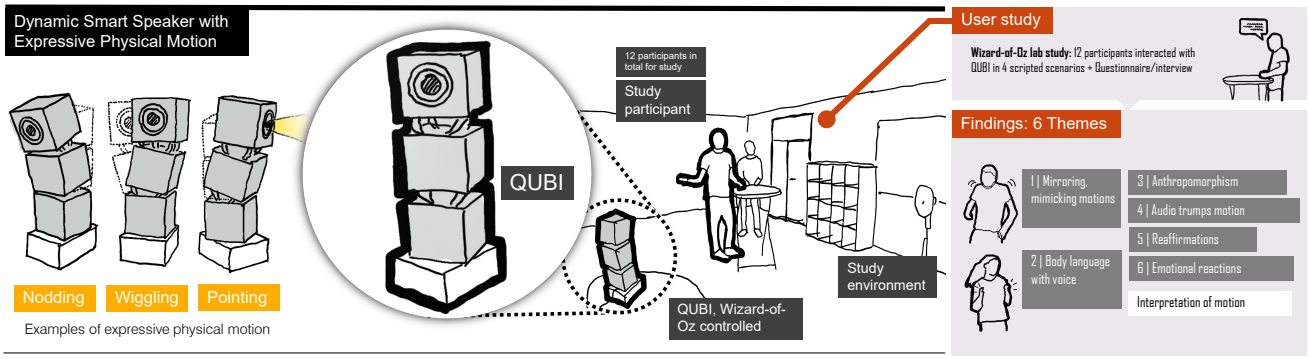

Figure 1: Visual abstract summarizing research of QUBI – a smart speaker design with multiple degrees of freedom for expressive physical motion, and key findings of our user study about people's interaction with QUBI.

ABSTRACT

People's physical movement and body language implicitly convey what they think and feel, are doing or are about to do. In contrast, current smart speakers miss out on this richness of body language, primarily relying on voice commands only. We present QUBI, a dynamic smart speaker that leverages expressive physical motion – stretching, nodding, turning, shrugging, wiggling, pointing and leaning forwards/backwards – to convey cues about its underlying behaviour and activities. We conducted a qualitative Wizard of Oz lab study, in which 12 participants interacted with QUBI in four scripted scenarios. From our study, we distilled six themes: (1) mirroring and mimicking motions; (2) body language to supplement voice instructions; (3) anthropomorphism and personality; (4) audio can trump motion; (5) reaffirming uncertain interpretations to support mutual understanding; and (6) emotional reactions to QUBI's behaviour. From this, we discuss design implications for future smart speakers.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Natural language interfaces**.

KEYWORDS

Smart Speakers, Intelligent Personal Assistants, Voice User Interfaces, Intelligibility, Breakdowns

ACM Reference Format:
Mirzel Avdic, Nicolai Marquardt, Yvonne Rogers, and Jo Vermeulen. 2021. Machine Body Language: Expressing a Smart Speaker's Activity with Intelligible Physical Motion. In *Designing Interactive Systems Conference 2021 (DIS '21), June28-July 2, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3461778.3462031

DIS '21, June 28-July 2, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8476-6/21/06…$15.00
https://doi.org/10.1145/3461778.3462031

1 INTRODUCTION

Internet of Things (IoT) devices have been around for a while, and it is perhaps only a matter of time until they fully transition into

(semi-)autonomous roles in homes [2, 46]. However, it is still an open question as to how more autonomous IoT devices [3] will blend into people's homes. Some IoT devices may start to adopt techniques from robotics [30, 50, 57], enabling them to react to users through physical motions, enriching their expressiveness. Indeed, research has shown that observing objects move elicits emotions in people and interpretations of "intent" [6, 25, 30, 36, 41, 74]. Desmond Morris wrote about *"the mannerism of walking, sitting and moving"* [54, p.260], which is a trait people pick up throughout their lives that helps them read other's intentions, such as asking for permission to talk by raising one's hand in the air. In particular, these gestures help people interpret each other's state of mind, thoughts, and feelings without necessarily speaking. How might we imbue the new generation of smart home technologies with such expressiveness or "machine mannerisms" so that human beings can better understand their intentions?

To answer this question, we turn our attention specifically to one kind of IoT device that has become popular in recent years and found its way into many people's homes: smart speakers (e.g. Amazon Echo [4], Nest Audio [33], Apple HomePod [7]). Smart speakers are an excellent candidate for investigating this research question due to their limited expressiveness. Current smart speaker designs typically use a cylindrical form with some LED status lights appearing on top. This forces users to rely almost entirely on the intelligent personal assistant's (IPA's) voice for interaction. However, smart speakers have limited conversational intelligibility [12, 56, 61]: users experience conversational breakdowns with their smart speakers as the device does not make the cause of the breakdown clear. These breakdowns can potentially also spread to more sophisticated smart home setups as smart speaker users tend to expand their homes with additional smart appliances connected to smart speakers over time [13]. As a result, it is challenging for smart speakers to blend into the fabric of users' everyday lives. Some users desire their smart speakers to be alive with a more animate form, as demonstrated by attempts to decorate the device by placing a doll on top of the device [21]. Similarly, other users show interest in making their IPAs act more as companions [49]. This suggests an opportunity to explore how smart speakers can be made more expressive and lively.

In this paper, we investigate how we can achieve a richer communication with smart speakers through the use of physical expressiveness in a novel smart speaker prototype. In particular, we explore whether adding physical motion to the smart speaker's form can make its conversational interaction and its activities within an IoT ecosystem expressive and intelligible [11]. While smart speakers are not social robots (e.g. Jibo [16, 26]), they can leverage their spatial placement in a room by utilizing techniques from the field of robotics such as pointing and orienting themselves to explicitly communicate their underlying activities. We designed QUBI, a smart speaker that complements its conversational interaction with physical actuation to inform users about its inner state and ongoing activities. QUBI supports nine different motions (Figure 2) for different purposes such as indicating QUBI's current state (ready, idle), whether it has trouble understanding or fulfilling a request, and pointing at other IoT devices. While our expressive motions are inspired by earlier approaches (such as Jibo's side-ways motion [26], through a two-part rotating head), we expand on this through wider range of expressive motions, and – most importantly – we focus on studying their effect for *mediating people's interaction with smart speakers*.

We conducted a qualitative lab study in which participants interacted with QUBI to perform a series of tasks in four scenarios. We wanted to understand people's reactions to physical motions in different situations, which motions complemented speech responses, and which were suitable for non-verbal communication alone, both during breakdowns and during expected interactions. Our findings suggest both opportunities and remaining challenges for expressive physical motion in smart speakers. In summary, we make the following contributions:

- A vocabulary of nine expressive physical motions to complement smart speakers' voice interaction and communicate the inner state, which we demonstrate through a smart speaker prototype, QUBI;
- Six key findings from our lab study investigating people's reactions to expressive physical motion in a smart speaker prototype: (1) mirroring and mimicking motions ; (2) body language to supplement voice instructions; (3) anthropomorphism and personality; (4) audio can trump motion; (5) reaffirming uncertain interpretations to support mutual understanding; and (6) emotional reactions to QUBI's behaviour. From these findings, we distill opportunities and challenges and derive future research directions for smart speaker design.

2 RELATED WORK

2.1 Body Language

People express themselves in various ways and one of these is through the use of body language such as crossing one's arms, frowning, or stomping one's feet. As argued by some [54, pp.14-26], actions individuals perform are either *inborn* (actions we do not have to learn), *discovered* (actions we discover for ourselves), *absorbed* (actions we acquire unknowingly from others), *trained* (actions we have to be taught), or a mixture of those. It is important to emphasise that there are actions that have different meanings depending on factors such as culture and situation. At the same time, people interpret gestures and actions similarly if there is enough context, such as when describing a car and pointing at it [22, pp.243-268], or raising one's hand in a class to get permission to talk. People's experience with gestures, mannerisms, and body postures has also been observed to influence how they interpret intent in moving objects [37]. Heider and Simmel [37] observed people attributing humanlike characteristics to objects due to their motions' trajectory and pattern. This is called anthropomorphism, and as Duffy [27] describes it, *"It is attributing cognitive or emotional states to something based on observation in order to rationalise an entity's behaviour in a given social environment."* This shows that not only does a person's body language inform others about their state of mind or intent, but so does the physical movement that gives rise to the body language or posture. The pace and trajectory of something or someone can shift a person's perception of the object or subject as Heider and Simmel showed with simple geometric shapes [37].

Physical actuation and motion as cues to a system's activities and intentions have been investigated in the field of human–robot

interaction [17]. Robots designed with more humanlike features such as eyes, hands, and mouths, have been shown to be perceived as transparent about their states [17]. In particular, people have attributed "mental" states to robots using the robot's explicit social cues such as nods and deictic gestures, as well as the robot's implicit behaviour such as the eye motions, by either looking at an object it wants to interact with or if it is ready to engage with the user by establishing eye contact [17]. However, as pointed out by Duffy [27], researchers and designers of social robots have to carefully design the robots to not fall into Mori et al.'s "Uncanny Valley" [53]. Beyond a certain threshold, adding more anthropomorphic features can cause people to feel uneasy about interacting with the robot. This also means that people attribute higher expectations of robots with more humanlike features, inevitably leading to disappointment if those are not met [27]. In other words, researchers and designers need to acknowledge the minimal human qualities needed that make social interactions among people work, and apply only those necessary ones to social robots.

By omitting humanlike features such as eyes, ears, and hands, researchers have shown that physical motion, or actuation in the case of objects, can also effectively convey the intention of a social robot [6], an automatic door [41], a piece of furniture [67], or a smart appliance [30, 74]. Non-humanoid robot design has shown to be effective in allowing people to understand both whether the robot is suitable for social interaction [6] and its intent by observing simple motions and draw positive and/or negative emotions from those. This shows that both explicit gestures and implicit behaviours can inform people about a robot's state without the explicit use of human facial traits or limbs. On a similar note, Luria et al. [50] have proposed using a social robot, Vyo, as an interface to control appliances in a smart home using physical icons (phicons). Vyo also leverages physical motion (or gestures) to suggest some of its states to its users (e.g. to indicate when it is listening or requires attention). Researchers have explored the introduction of social robots into people's homes [58, 71] and some studies suggest that many people may be favourable towards such a vision [24]. Furthermore, Garcia et al. [30] found that physical movement can influence people's perception of an autonomous object's performance as well, which shows how physical movement can be associated with tasks, also called *consequential communication* [65] as some movements such as a Roomba's cleaning are inherent to the robot's task. This shows the potential of leveraging human gestures and body language not only in social robots, but also with respect to IoT technologies such as smart speakers, as current smart speakers suffer from a range of challenges.

2.2 Challenges with Smart Speakers

This paper grounds itself in the body of work on how voice-enabled smart speakers, as potential future conversational interfaces between inhabitants and smart homes, hide or show limited information about the smart speaker's internal activities [12, 61]. The fact that smart speakers face conversational issues [12, 61] makes it so that the devices have a difficult time integrating into the fabric of everyday life. Smart speakers have a homogeneous design across different brands, using typically a cylindrical form factor, and a seamless design in which information about the device's activities is restricted or made unavailable upfront. Smart speakers rely on their voice responses and thus are prone to known voice user interface (VUI) limitations such as providing too much information in a voice response, which can be overwhelming or providing little to no information, which then becomes non-informative [55]. While other challenges include unhelpful responses from smart speakers in conversational interactions [12, 61], not every problem is conversational. As smart speaker users tend to expand their smart homes with additional smart appliances over time [13], it is also important to point out the increase in customisations of smart homes with action-trigger service such as IFTTT (If-this-then-that) [73]. These increased trends could lead to breakdowns [77] that revolve around IoT ecosystems in which mispronunciations are not the only type of errors that could happen. In fact, home automation has been reported to be difficult to keep track of, as multiple household members could change settings such as smart lighting timers [46]. Finally, smart speakers are known for fading into the background during events where people immerse themselves in activities, resulting in smart speakers that react to false positives and unintentionally execute misinterpreted requests [76]. As has been argued before [11], it is important to design such sensing systems by making it clearer to users what those systems sense and how they can negate unintentional consequences. Breaking away from the current smart speakers' minimalistic and seamless design might in fact help make smart speakers become transparent about their internal behaviour as well as more visible and engaging.

2.3 Physical Expressiveness in Smart Speakers

More recently, expanding on the core functionality of smart speakers, commercial approaches are emerging that incorporate motion into smart speaker design, for example, using displays that rotate around their base ([5, 26, 28])". A few studies have begun to further explore of how such smart speakers and their intelligent personal assistant's (IPA's) presence could be designed differently than the state-of-the-art. Kim et al. [45] experimented with an IPA becoming more present in the space in which users live and do activities by giving an IPA a virtual human body in augmented reality (AR). Such presence informs the user about their IPA's location and if it is attentive to their requests, which helps the user determine whether the IPA is potentially violating their privacy or not, or if they are attentive, similar to how we perceive another person's presence. Similarly, McMillan et al. [52] designed Tama, a gaze-aware smart speaker, which is invoked through mutual gaze. In combination with head rotation, Tama is capable of establishing mutual gaze and orient itself towards the user with which it is interacting. This subtle body language is used to establish a connection between the user and their smart speaker on a different level compared to the current smart speakers. While both studies point towards an opportunity to strengthen trust and invocation, they also show that if smart speakers externalize some of their underlying activities, the smart speakers' internal behaviour becomes clearer to users.

3 DESIGN PHILOSOPHY

3.1 Inspiration from Body Language, Robotics and Proxemics

Our design philosophy builds on the idea that bodies communicate through gestures, manners and postures, and are contextually grounded [22, 23, 43, 54]. This motivated us to investigate physical motion as a means to inform users about underlying activities

and thereby provide a better understanding and control of other smart home devices through the smart speakers. While smart speakers' minimalistic designs limits their expressiveness, social robots possess the quality of expressing themselves through physical motion [6, 19]. We draw on and are inspired by work on social robots (e.g. Jibo [16, 26]), but in contrast to focusing on evoking emotional responses in social interactions [6, 18, 48], our primary focus is on leveraging social cues and behaviours to mediate interactions with smart appliances. In particular, smart speakers go through a series of underlying activities and states during interaction sessions from being invoked and listening, to interpreting, and to (de-)activating a smart appliance, and it is these kind of underlying states we want to unpack and make explicit through physical motion. This motion can either be inherent to the task a robot is completing [30, 65], or simply a physical motion that is independent of task execution. A smart vacuum cleaner, like iRobot's Roomba, moves around to clean the floor, and as a result communicates its state and progress through this motion inherent to its task. While we do not argue that smart speakers have similar features when executing their tasks, we do emphasise that physical motion is part of human communication, similar to the use of physical expressiveness in communication (e.g. gestures, posture, and subtle physical expressions). Using physical motion as a means to express a robot's intent has been found to be make robots more readable [69] and generally focusing on expressive motions can make robot's less reliant on anthropomorphic designs [38]. Indeed, people anthropomorphize technologies that behave in familiar ways [63], and the same applies to smart speakers due to their VUI [49, 68]. Given enough anthropomorphic features, people may start to overestimate a robot's capabilities [27]. That is why Hoffman & Ju argue for leveraging the potential of physical motion [38]. Another important part of human communication is proxemics [35]: the nuanced ways in which people negotiate interpersonal space and how this spatial proximity maps to social proximity. Proxemics has already been explored in designing people's interactions with technology [9, 34, 51], and we believe it can be useful to consider in terms of designing physical motion for smart speakers as well. Smart speakers are typically located in rooms like kitchens and living rooms [64] where activities can become noisy, possibly making it difficult for the smart speaker to interpret the users' requests. In this case, smart speakers could make use of machine body language that might entice the user to get closer to the device in order for it to interpret the request properly.

3.2 Inspiration from Limitations of Current Smart Speakers

We identified five opportunities of transforming static smart speakers into dynamic ones based on the limitations of current smart speakers.

O1: *Unclear when the device is listening.* There are situations in which smart speakers react to false positives and accidentally complete requests that are not supposed to happen [76]. In current smart speaker designs, it is not always clear when the device is listening or being invoked. Clear delineations of IPAs' presence in rooms has shown to increase users' awareness of their digital assistants "whereabouts" and attentiveness [45]. This suggests there might be opportunities to use physical motion to clearly delineate different states of the smart speaker such as being invoked and ready to listen or inactive.

O2: *Unhelpful responses when a request cannot be completed.* Smart speakers' voice responses can be repetitive, and are not always helpful to guide the user to a solution. It can be unclear to users whether the device heard anything at all, understood the request, or could not go through with the request due to some other issue [12, 61]. This provides an opportunity to investigate whether using physical motion can be used as an additional cue to clarify whether the smart speaker did (not) understand the user as opposed to when it could not perform such a task.

O3: *Lack of access to alternative interpretations.* Smart speakers do not provide much useful information during conversational breakdowns [12, 61].The user's requests that the smart speaker has interpreted are commonly available in the companion app on the user's smart phone. However, in case of an ambiguous request, users typically lack a way to access alternative interpretations. Alternative interpretations could help to repair communication and correct a wrong response from the smart speaker. Yet, offering lists of options via speech (like in telephone menus) can overload people's cognition [55]. It is interesting to explore whether providing this list of alternatives is deemed useful when the voice response is augmented with physical motion cues inspired by how people would present multiple alternatives [59] to provide additional grounding and improve recall [10].

O4: *Unclear when background noise is an issue.* Background noise can cause problems for smart speakers to interpret the users' requests. Prior work found that users tend to actively incorporate silence and use turn taking to reduce background noise and interference [61]. However, in some situations, users may not be aware that background noise poses a problem, for example, when attempting to address the smart speaker at a party or while playing music. The only indication of a potential issue will be the lack of or an incorrect response from the smart speaker. This begs the question whether physical motion cues could be leveraged to indicate that the environment is too loud or that users should get closer to the smart speaker.

O5: *Difficulty in addressing other IoT appliances.* Smart speaker users tend to use their smart speakers with other IoT devices over time [13], by using their smart speaker as a control interface for smart appliances and configuring custom trigger-action rules with services like IFTTT [72]. Having numerous smart appliances connected to a home network with the smart speaker as a central interface to these appliances can be overwhelming for users with respect to knowing what appliances are available and recalling their names [8]. This can make it hard to predict what will happen when the smart speaker is asked to "turn on the lamp". This creates possibilities for exploring the use of physical motion as a means to clarify and disambiguate which appliances the smart speaker is connecting to.

4 QUBI: DESIGNING A PHYSICAL MOTION VOCABULARY FOR SMART SPEAKERS

We designed QUBI, a smart speaker prototype with expressive mannerisms, leveraging gestures, body posture [54], and proxemic

interaction [51]. QUBI is composed of three moving cubes that allow it to transition between nine states (Figure 2) by using its four degrees of freedom (DoF): (1) tilting back- and forth, (2) tilting to the left or to the right, (3) raising and lowering itself, and (4) rotating around its own axis. In addition, QUBI is equipped with a ring of 24 RGB LEDs and a flashlight in the center of the LED ring. Finally, QUBI provides speech output to allow an operator to respond to voice requests and simulate the VUI capabilities of commercial smart speakers.

4.1 The Design of QUBI

Nine Motions. We designed QUBI with a vocabulary of nine distinct expressive physical motions (Figure 2).

- ***M1$_{ready}$***: QUBI rises up from the idle state (***M9$_{idle}$***) and the LED ring turns yellow (Figure 2.M1). This was designed to acknowledge the user's request for attention and form of greeting. As described by Morris [54, pp.88-93], people usually greet each other by a degree of *inconvenience display*: a level of displacement that the greeter takes to show the strength of their friendliness (e.g. standing up from a seated position when a guest enters the room). In addition, Sirkin et al. [67] observed how participants noticed that the raising and lowering of an Ottoman's cushion lid indicated its readiness to go, and so we incorporated this idea that when the prototype was idle, it would shrink and bend over slightly, while when invoked, it would rise up and signal *readiness*. This also strengthens the smart speaker's presence in a room, making users aware that it is listening and allowing them to react in case it was a false positive (**O1**).
- ***M2$_{nod}$***: QUBI nods with the upper most cube by moving it down and up once (Figure 2.M2). This is analogous to how people nod [54, pp.80-83] and has been used in robotics to say *"yes"* [17]. In this case, QUBI uses this motion to acknowledge the user's command and proceeds to complete the request (**O2**).
- ***M3$_{shake}$***: QUBI rotates once around its own axis left to right for 20 degrees, imitating a head shake (Figure 2.M3), to indicate that QUBI cannot comply with the user's request. A headshake is commonly understood by people as a negative response, covering a wide range of *"no's"* such as *"I cannot"*, *"I disagree"*, and *"I do not know"* [54, pp.80-83]. The headshake is usually also understood as a negative response in cultures where other negative reactions are used instead [54, pp.80-83]. We make a distinction between *understanding* a request, but not being able to comply on the one hand, and *not understanding* the request on the other hand (**O2**). For the former, QUBI uses ***M3$_{shake}$*** and for the latter it uses ***M4$_{shrug}$***.
- ***M4$_{shrug}$***: QUBI performs a (shoulder) shrug gesture by raising its middle cube and lowering its upper most cube simultaneously, after which each cube performs the opposite motion to return back to its prior position (see Figure 2.M4). This is to indicate uncertainty about QUBI's interpretation of the user's request. Note that not all cultures adhere to shrugging the shoulders as a sign of uncertainty [40], especially in Eastern cultures the shrug is not necessarily interpreted as a sign of uncertainty, as it is in Western cultures. We envision this motion to be used when none of QUBIs interpretations of what was said are above a set confidence threshold (**O2**), as opposed to when QUBI has two highly likely interpretations (for which we use ***M7$_{wiggle}$***).
- ***M5$_{forward}$***: QUBI leans forward to indicate that it gets closer to the user (Figure 2.M5), inspired by proxemics in which proximity indicates increased engagement [51]. We wanted to investigate whether QUBI could make conversations between users and smart speakers more natural by signalling its difficulties in interpreting the user's request. By leaning forward, QUBI signals to the user that they should get closer [70] by attempting to trigger a common reaction known as body pose mirroring [29].
- ***M6$_{backward}$***: In contrast to the previous motion, when QUBI leans backward, it disengages from what is going on and distances itself from the user (Figure 2.M6). This is designed to indicate that QUBI feels a discomfort with the environment [70], suggesting that high background noise will influence QUBI's reaction to and interpretation of the user's requests (**O4**). By leaning back, QUBI attempts to tell the user to get closer and speak up, if they intend to interact with it. We designed ***M5$_{forward}$*** and ***M6$_{backward}$*** to provide additional information about the sensing capabilities of the smart speaker, and help users realize why the smart speaker does not respond to their requests, possibly even before they are made.
- ***M7$_{wiggle}$***: QUBI tilts once to the right and asks the user if they meant interpretation 'A', followed by a tilt to the left and asking if they meant interpretation 'B', and finally goes back to the ready position (Figure 2.M7). This gesture is inspired by how people sometimes raise their hands with the palms facing up, saying *"On the one hand ... on the other hand"* [59]. This motion is an attempt to address **O3**.
- ***M8$_{point}$***: QUBI points at a particular connected IoT device by orienting itself towards the device, leaning forward, turning on the flashlight, and describing the device (see Figure 2.M8). Pointing is a deictic gesture and refers to the temporality of directing someone's attention to something with respect to the subject of concern [22, pp.243–268]. Pointing has been shown to help people understand robot's "mental" states in a given situation [17], which can influence people's subsequent behaviour and understanding of the robot.
  To address **O5**, we incorporated a flash light into QUBI to allow it to point at other IoT devices and increase transparency in terms of with which other IoT devices QUBI is communicating. This motion explores the potential of a more seamful design [20] approach to the smart speaker's internal behaviour, by explicitly showing the device's communication within an IoT ecosystem.
- ***M9$_{idle}$***: QUBI is in a collapsed and slightly forward bended position and triggers a white slow fading in and out animation of the LED ring (Figure 2.M9). We added this animation as an additional clue for being in standby: inactive but ready to respond if invoked (**O1**). We drew inspiration from Apple's breathing LED light in their laptops when in sleep mode [60].

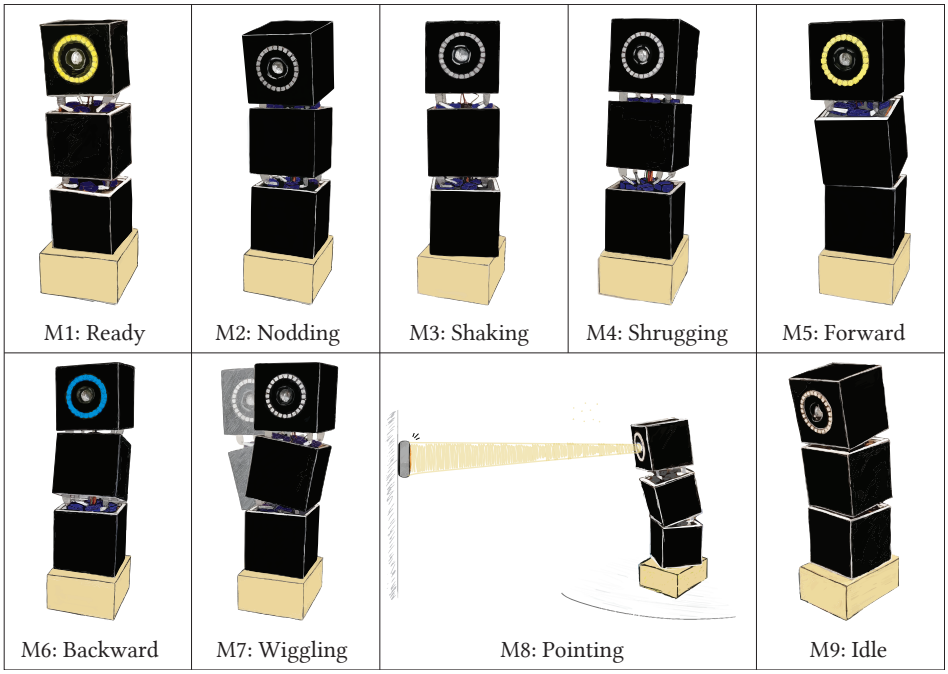

Figure 2: QUBI's Nine Motions.

LED States. The LED ring has four states: the *off state* (i.e. LEDs are off) is either used when the prototype is in a transition between two physical states or pointing at something with the flashlight on. The *attentive state* (i.e. all LEDs turn yellow) indicates that the prototype is ready to attend to the user's requests. Yellow is associated with the 'ready' state, as has been argued and used by McMillan et al. [52]. Additionally, the bright yellow blends well with other light sources in a typical home environment. The *busy state* (i.e. parts of the LEDs turn blue and start moving in a circular motion) indicates that the prototype is busy searching for a response, hence not ready to attend to the user's request. Similar to yellow, we were inspired by McMillan et al.'s use of blue in a *"loss of gaze"* context [52], which we translated here as a loss of attention, i.e. the busy state. QUBI also uses the busy state when the background music is too loud for it to understand requests properly, indicating that the user has to get closer, before QUBI can attend to their request. Finally, the *resting state* turns the LEDs white, while fading in and out.

Speech. We included speech in QUBI to resemble IPAs in available smart speakers and allow an experimenter to respond to voice requests using common smart speaker responses (e.g., *"OK"*, *"I'm not sure how to help"*, and *"You got it"*). We also added a number of additional phrases, which were not available in commercial smart speaker IPAs due to the novelty of some of our gestures such as pointing.

4.2 Implementation

We used off-the-shelf and lightweight materials such as Styrofoam blocks, foam core and cardboard that we attached together with glue and nails to make the cubes. We used a NeoPixel Ring 24 x 5050 RGB LED and a Hausbell 7W mini LED flashlight for the colour LED ring and for pointing respectively. We rapidly prototyped a simple system that we could control remotely. We built a graphical control user interface (Figure 4.b) in Python with Tkinter through which a Wizard-of-Oz operator [42] could send commands to QUBI using serial communication. In addition, the GUI allows the operator to play scripted voice responses to the participants' requests. We also included the ability to improvise with a free-form text field that uses the Google text-to-speech API [32], in case participants would say something that went beyond the standard scenarios used in the study. For QUBI's voice, we used one of Google's available male voices. A microcontroller (Arduino Uno) pushes the commands it receives from the GUI to a PCA9685 16-channel servo controller that controls eight MG90S servos and one HS-322HD servo. While one servo was positioned below the three cubes to rotate them, the other eight were split into two groups (four in each) and placed in between the gaps of the prototype (Figure 3). The servo arms faced each corner of the cubes diagonally to allow the cubes to raise and lower from the corners and also fit tightly into a small surface area. Finally, the Arduino Uno also controls the 24-RGB LED NeoPixel Ring and flashlight.

5 STUDY

We conducted a qualitative lab study, inspired by Sirkin et al.'s Wizard of Oz (Mechanical Ottoman) study [67], to investigate participants' general experience with and reactions to a physically actuated smart speaker in a domestic setup. We were interested in understanding which motions complemented speech responses and which ones were suitable for non-verbal communication, both during breakdowns and during expected interactions. We decided

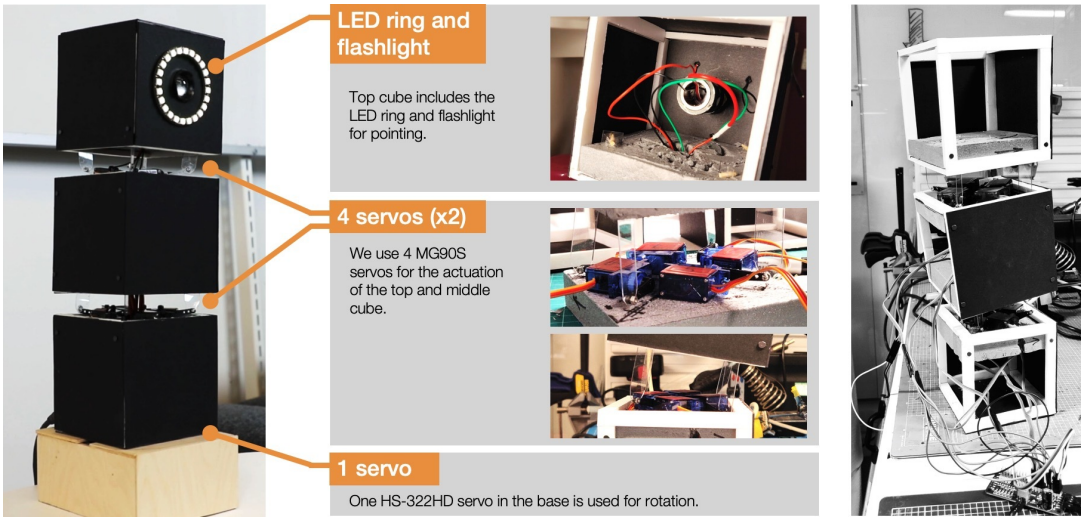

Figure 3: QUBI implementation: placement of LED ring, flashlight, and the nine motors. Inside and construction views of the top cube where the flashlight and LED ring are placed, and close-up of four of the eight MG90S servos and their attachments.

to run a Wizard of Oz study [42] since it allowed us to create and control interaction breakdowns for the purpose of investigating how the participants would react. As we were also interested in understanding what the participants thought during their interactions with QUBI, we asked the participants to think aloud [15].

5.1 Participants

To recruit participants for the study, we announced an open call on social media platforms (e.g. Facebook) and placed posters around public spaces in our city. We also shared the call with colleagues and friends, and approached people in public spaces to ask whether they would be interested in participating. The only requirement for our recruitment was that the participants were capable adults who would voluntarily participate in the study and were comfortable with English due to the prototype only supporting responses in English. We deliberately chose to recruit a broad sample of participants who could consider using a smart speaker across a variety of ages. Twelve people participated in our study (4 female and 8 male). Ages ranged from 20 to 52 with a mean of 32.4. 8/12 participants were students studying political science, natural science, finance, and engineering, while the rest comprised a film producer, a property manager, a marketing manager and an IT consultant. Some participants lived alone (4/12), others with family (6/12) or in a shared accommodation (2/12). The participants' frequency of use of smart speakers or voice assistants ranged from never (5/12), occasionally (4/12) to a few times a day (3/12). The interaction sessions with the prototype lasted 20–38 minutes with an average duration of 29 minutes, and were followed by semi-structured interviews (total duration: 53–103 minutes, average 77 minutes).

5.2 Study Setup

The study took place in a lab space in our university building (Figure 4.a), which was set up to simulate a domestic setting. We placed one GoPro wide-angle camera behind the prototype to record a frontal view of the participant, and a DSLR camera to record a frontal view of the prototype. The participants were standing at a standing table approximately three meters away from QUBI when starting the study. We added subtle masking tape markings to the floor to facilitate monitoring participants' distance to QUBI in later video analysis. We equipped the setup with two loudspeakers for music and voice output, hidden under the tablecloth of the table that QUBI was placed on. We connected two lamps and a fan to smart plugs that the participants could interact with through QUBI. An operator controlled the prototype and the smart plugs from a separate room. QUBI was controlled using TeamViewer connected to the laptop that QUBI was connected to and the smart plugs were controlled via an infrared remote. Another experimenter introduced and guided the participants through the study and was present in the lab space at all times. We verified through pilot studies that the setup was realistic (e.g. that QUBI appeared to operate autonomously and respond to the participant, and that the voice and music output appeared to originate from QUBI).

5.3 Procedure

5.3.1 Demographics and Introduction to QUBI. Every participant took part in the lab study individually. First, the experimenter introduced the participants to the study and asked them to fill in a brief demographics survey. Halfway through this survey, the operator made QUBI greet/address the participants. The use of a simple demographics survey to surprise users with a proactive motion was inspired by Sirkin et al.'s approach [67]. This was to observe the participants' awareness of the prototype's presence and how people would react to a device introducing itself. Participants were unaware that this was part of one out of four scripted scenarios in which participants would engage with QUBI.

Scenario 1: QUBI moves from the idle to the ready state and introduces itself to the participants by saying *"Hello, my name is QUBI, I will be your digital assistant for today"*.

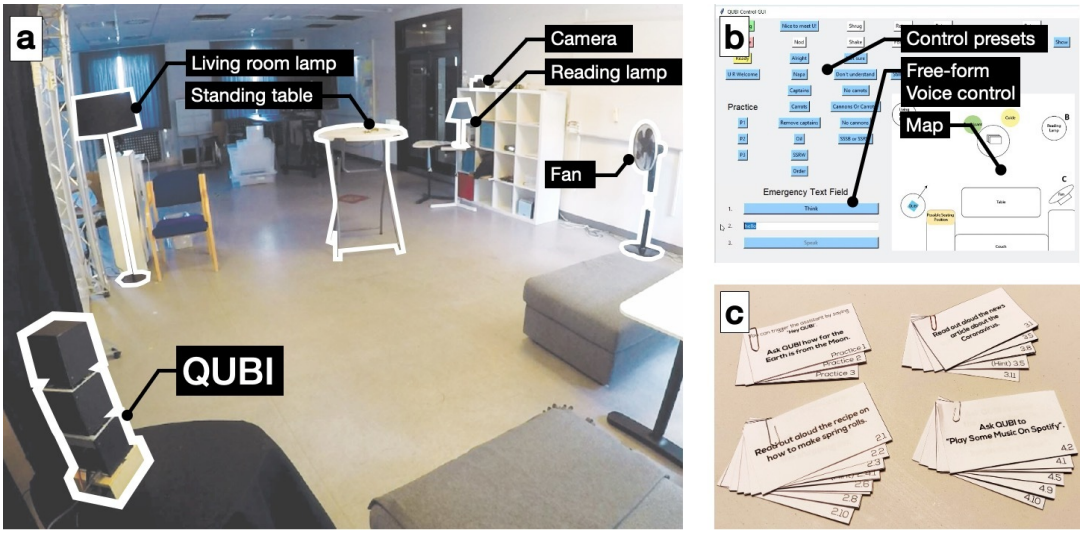

Figure 4: Study setup: (a) lab environment used for conducting the user study, (b) Wizard-of-Oz control interface, and (c) cards used for the practice activities and study tasks for the different scenarios.

5.3.2 Demonstrating QUBI's Motions. Next, we walked the participants through the nine physical motions of the prototype and asked them to write down a *single* word that best describes each motion (we allowed participants to write down more words in case they strongly associated the motions with more things), followed by a sentence to elaborate on the one-word description. We were interested in investigating participants' similarities and differences when interpreting the motions. Additionally, we asked the participants to reflect on their initial descriptions after they interacted with QUBI in the different scenarios, to investigate which motions are more context-reliant [65] than others.

5.3.3 Practice Scenario. Before starting the last three scenarios, the experimenter asked the participants to go through three sample questions with QUBI, breaking the ice and getting them familiar with interacting with QUBI. The sample questions were written as follows on three separate cards: *"Ask QUBI how far the Earth is from the Moon."*, *"Ask QUBI 'Do you know where you are?'"*, and *"Ask QUBI 'Why are you here?'"* QUBI's responses were scripted in a humorous way, to make the participants a bit more comfortable interacting with the prototype.

5.3.4 Scenarios 2-4. After the introductory part, we guided each participant through three remaining scenarios. Each of the tasks within the scenarios were written on a card that we gave the participants as they went along. By giving the participants one task at a time, the experimenter stayed in control of the timing of when the following cards (Figure 4.c) were given so it resembled a more natural flow of interaction rather than a fast-paced procedure. In case the participants got stuck and gave up on a task (e.g. reducing the volume), the experimenter would provide a hint-card (e.g. *"Get closer to QUBI and speak louder"*) to aid the participants in completing the task. The written instructions helped streamline the study by avoiding disruptions and inconsistencies during the scenarios. Finally, participants were asked to engage in a few additional tasks when not interacting with QUBI directly. These tasks were introduced to simulate a realistic setting in which QUBI is used to support ongoing activities in a home setting. The additional tasks would flow with the interactions with QUBI in each scenario: reading a recipe (scenario 2), reading a news article and playing a puzzle game (scenario 3), and drawing on paper (scenario 4).

Scenario 2: We designed this scenario around adding items to a shopping list and placing orders. This scenario had 10 tasks in total, ranging from reading a printed-out recipe of a dish, asking QUBI to add/remove ingredients to/from the shopping list, to placing an order based on the shopping list. We also introduced an intentional breakdown where QUBI would not understand the item *carrots* until the participant approached QUBI. In addition, as the participant would try again after they got closer to QUBI, QUBI would interpret their request for *carrots* as *captains* and add *captains* instead. We included this deliberately to elicit possible moments of frustration and the need to repair a breakdown.

Scenario 3: This scenario was designed around the events of controlling other IoT devices in a home. This scenario had 11 tasks in total, ranging from reading a printed-out news article and doing a puzzle on paper (separately), to asking QUBI to turn on/off two lamps and a fan.

Scenario 4: Lastly, we designed a scenario in which participants had to control the same appliances as in scenario 3, but this time with noisy background music. This scenario had 10 tasks, one of which had to be completed while noisy music was playing. During the noisy music, we scripted a breakdown where QUBI would lean back and turn its LED ring blue as it said *"It's loud in here."* to indicate that QUBI would have a difficult time understanding the participant's request. The participants had to walk closer to reinstate QUBI into a ready position. In this scenario, we also asked the participants to draw and show how they imagined their smart speaker would look and behave if it was listening, uncertain, and wanted their attention.

5.3.5 Semi-Structured Interviews and Closing. We concluded the study by conducting semi-structured interviews with the participants to delve deeper into our understanding of the participants' experiences and their own understanding of QUBI. We focused the interview on the participants' behaviours that we observed during the scenarios and asked why they (thought) they (re)acted the way they did. Furthermore, we aimed to uncover the participants' understanding of QUBI's mannerism (postures and gestures), whether their expectations varied or not, what they expected QUBI to do next, and if they associated QUBI and its motions with something, like a particular gesture or an animate being. We also followed up on the participants initial descriptions of QUBI's motions to see if they understood the motions differently now.

At the end of the interview, we also informed the participants about the Wizard of Oz setup. None of them realized that QUBI was controlled by an operator.

5.3.6 Analysis. The video recordings of the participants' interactions with QUBI were examined using a simplified version of interaction analysis [66, pp. 329–334, Table 9.5], due to our focus on only a single participant's interactions with QUBI. Through our in-depth video analysis, we coded both the verbal and non-verbal interactions between the participants and QUBI. The video analysis resulted in 101 codes, with 74 related to interactions with QUBI. The 101 codes were grouped in 21 high-level codes – e.g. 'proximity', 'appliance' (e.g. fan), and 'head movements'. Examples of subcodes in 'head movements' include 'nod', 'lower head', and 'frowns and nods'. The analysis of the interview transcripts comprised creating and renaming, splitting and merging, and linking and making hierarchies of the codes [14, pp.51-54]. One author coded the data to familiarize himself with the data and then further discussed together with the other authors to develop themes in an *interpretivist semi-structured approach* [14, pp.63–64]. In the interview, we also asked the participants to reflect and compare their initial descriptions of QUBI's motions without speech responses to how they experienced them in the scenarios with speech responses. We then compared these two sets of descriptions.

6 RESULTS

6.1 Interpretations of and Reactions to QUBI's Motions

We asked the participants to describe their impression of QUBI's nine motions (in a *single* word and *one* sentence) before further engaging with QUBI in the scenarios. Table 1 shows these descriptions and participants' reactions to each of the nine motions during the study. Participants' descriptions of the motions that are in line with the intended meaning are highlighted in green (e.g. "agree" for ***M2nod*** or "resting" for ***M9idle***).

6.2 Reflections on the Motions after the Scenarios

After the scenarios, the participants were asked to reflect on these motions again. **P5** reflected on ***M3shake*** and ***M4shrug*** as they seemed the same to him. So he thought that ***M4shrug*** was used to perform a function, but was not sure about it as he initially wrote that shrugging was like *"Dude, I don't know what you are saying."*. It is surprising that after the scenarios he thought that ***M4shrug*** performed a function considering the accompanying voice response *"My apologies. I don't understand"*. All other participants understood ***M4shrug*** after the scenarios. **P11** mentioned that he understood ***M5forward*** initially as if QUBI was focusing on something but changed his view on it as he realised during the interaction that QUBI used it when it did not understand him. While this was what we expected, we also expected that the participants would naturally walk up to QUBI, however that proved not to be the case for any of the participants. **P7** found it odd that QUBI would lean back (***M6backward***) when the music was loud as the music was coming from the device itself. She suggested *"Getting smaller, trying to hide from the music."* as an alternative physical motion. None of the participants reacted to ***M6backward*** as expected, since they found it unclear what QUBI was asking of them. When asked if **P6** knew when QUBI wiggled, he did not recall, yet he immediately said that ***M7wiggle*** meant, *"Did you mean this or that?"* as he would do that himself with his hands. **P10** made a similar comment. We did not expect the wiggle to be clear without the additional voice response, as also evident from participants' initial reactions in Table 1. As both **P10** and **P12** confirmed, QUBI's additional voice response made it clear that QUBI was offering options. After completing the scenarios, all participants understood that ***M8point*** was about pointing at an appliance and controlling it, as was expected. Surprisingly, **P5** found ***M9idle*** unclear due to the repeated fading in and out. He was unsure if QUBI was attentive or not, even after he completed all of the scenarios.

Summary. Most of the motions were clearer to the participants after they had interacted with QUBI in the scenarios. This is perhaps unsurprising, given that the physical motion cues were designed to complement voice interaction. Since we did not ask the participants to systematically go through all of the motions, but rather whether they understood some of the motions differently now, it is difficult to assess how many participants changed their opinion about the motions after their interactions with QUBI. As some of the participants mentioned, they either overlooked or did not remember observing certain motions in the scenarios. Most notably, nearly all participants interpreted ***M7wiggle*** as a dance in their initial descriptions, showing that ***M7wiggle*** is an example of a movement that needs context and audio cues, whereas the participants almost unanimously described ***M4shrug*** as shrugging shoulders or not understanding. Surprisingly, none of the participants interpreted QUBI's forward and backward leaning (***M5forward*** and ***M6backward***) as a sign for the participants to get closer. Participants required additional hints during the scenarios to understand the meaning of the ***M5forward*** and ***M6backward*** motions. ***M5forward*** could have benefited from an additional voice response to make QUBI's intent stronger when it tried to signal the participants to come closer. While ***M6backward*** had a verbal response, *"It's loud in here."*, it was still unclear to the participants that they had to get closer. In fact, because QUBI did not react until the participants would approach QUBI, it is unclear to which extent QUBI's lack of response, the loud music, the card with the hint, or a combination of those were the triggers that made the participants approach QUBI. Finally, in terms of the participants' reactions, ***M4shrug*** seemed to be the most

Table 1: Participants' initial descriptions of QUBI's nine motions with a single word and optional sentence as elaboration, and their verbal and emotional reactions to the nine motions during the scenarios. Green highlights indicate participants' interpretations that were closely aligned with our intended meaning for that particular motion.

QUBI's Motions	Initial Descriptions	Participants' Reactions
M1: Ready	"rising" (5), "stretching" (3), "ready" (2), "opening" (1), "plant on grow spurt" (1).	Chuckle, laugh or smile (3), frown (3).
M2: Nod	"nod" (8), "Yes" (2), "agree" (1), "bend" (1).	Chuckle or smile (4), "ah-ha" (1), "huuh aww" (1).
M3: Shake	"shake" (5), "no" (3), "disagree" (1), "shifting" (1), "looking around or declining" (1) , "rotate" (1).	Chuckle, laugh or smile (5), imitate (2).
M4: Shrug	"shrug" (5), "I don't know" (2), "lifted its torso" (1), "perhaps" (1), "retreat" (1), "confusion" (1), "contract" (1).	Laugh or chuckle (9), imitate (2), *"cute"* (1), *"Jesus"* (1), *"I am associating that thing with an actual human being."* (1)
M5: Leaning Forward	"leaning forward/interested/looking closer/focus" (5), "turning down/lowers" (2), "sitting down/knees" (2), "contraction" (1), "sad" (1), "retreat" (1).	Smile (3), frown (2).
M6: Leaning Backward	"leaning back" (2), "fear" (2), "reserved" (1), "I don't like" (1), "shock" (1) , "adjustment" (1) , "slight bend" (1) , "focus" (1), "sitting" (1), "lowers" (1).	Chuckle or smile (7).
M7: Wiggle	"dance" (8), "swaying from side to side" (2), "Shimmy" (1), "rolling" (1).	Laugh or smile (6), imitate (1), *"It is very synthetic"* (1).
M8: Point	"look(ing) at this" (4), "focus" (3), "projection/film" (3), "turning to the right" (1), "quarter turn" (1).	Observed the projection (9), smile (2), frown (2).
M9: Idle	"resting/standby/charging/shutting back down/finished" (7), "contract together/going down" (2), "shrinking downwards" (1), "owner needs help" (1), "lighting up" (1).	Smile (2), *"hmmm"* (1).

surprising motion, as 9 participants either laughed or chuckled compared to the other motions.

7 OVERALL FINDINGS

After coding our interviews and analysing the video recordings, we discussed our combined findings and converged them into six key themes.

7.1 Theme 1: Mirroring and Mimicking Motions

We observed that participants often imitated QUBI's motions during the study. To make it easier to differentiate between instances when participants copied QUBI's motions in the moment during their interactions and when they reconstructed the motions during the interview, we will use the terms *mirroring* and *mimicking* respectively.

In the beginning, when we introduced the participants to QUBI's nine motions, we noticed that three participants mirrored QUBI's movements. **P2** mirrored ***M2nod***, ***M4shrug***, and ***M7wiggle***, while **P5** and **P8** only did it once with ***M3shake*** and ***M4shrug*** respectively. During both scenario 4—in which participants had to draw, show, or describe how they imagined their digital assistant—and the interviews, we observed eight participants mimic a variety of motions they recalled QUBI doing or imagined that their own version of QUBI would do. When we asked **P5** about what QUBI did when it did not understand him, he thought that it clearly indicated that it did not understand him but could not remember what it said. In addition, **P5** did not notice himself move his hands to either side when he said *"But it was nice that it also suggested two options and said 'Did you mean this or that?'"* In fact, even though he mimicked QUBI's wiggle motion, he did not recall that QUBI tilted to both sides to indicate two options, instead, he referred to ***M7wiggle*** as a dance, thinking it happened during the music. In addition to **P5**, five other participants also did not remember ***M7wiggle*** when QUBI suggested two options, however, they made the connection as they thought that the two suggestions mapped well with ***M7wiggle*** while mimicking the motion with their hands. Out of those eight participants, **P10** and **P12** were the only ones who remembered the wiggle and mimicked the motion themselves with their hands, and as **P10** said: *"I think they fit very well with where the mind would be. It feels very natural to what a person would do…like the body language."* **P12** also mimicked the shrug and thought that his version of QUBI would also tilt its head when it was uncertain. In other instances, six other participants nodded right after observing QUBI nod as it completed a task. While some participants only did it once or twice such as **P4**, **P6** and **P10**, others did it between three and six times like **P2**, **P3**, and **P5**.

7.2 Theme 2: Body Language to Supplement Voice Instructions

The video analysis also showed that participants used deictic gestures such as pointing at appliances or redundantly using their fingers to indicate the number of the choice they made. During a scenario where **P9** requested QUBI to turn on the living room

lamp, QUBI tilted to each side once while asking, *"Did you mean the reading lamp or the living room lamp?"* **P9** responded with *"Living room lamp."* while tilting his head in the direction of the living room lamp. This could both be viewed as mirroring the direction QUBI was tilted towards as it said, "living room lamp" and as body language to supplement **P9**'s verbal answer *"living room lamp."* **P11** was also observed doing this when he requested QUBI to turn off the fan. Apart from **P9** and **P11**, eight other participants were also observed using gestures supplementing their speech, seemingly to clarify what that they were communicating to QUBI. We noticed three participants lean over to QUBI to assure that QUBI would hear their request clearly, during some instances where QUBI had previously not understood their request. On a slightly similar note, **P12** nodded his head upwards once as he said *"Hey QUBI"*. Three participants also made use of their hands when they wanted to either emphasize a word such as *"squared spring roll wrappers"*, explain to QUBI what a carrot is and what it looks like, and indicate which appliance(s) the participant was referring to in the request. Finally, five participants made use of their fingers. Three used it to confirm the task completion to QUBI with a thumbs up gesture. Other reasons included pointing at the recipe while asking for carrots, gesturing in mid-air with an emphasis on a particular word, redundantly making a gesture for the number of the choice they made, and instructing QUBI which task to complete first. **P8** also imagined if she was frustrated about QUBI playing music that is too loud, she would curse at it while raising her voice and index finger to indicate seriousness about the request. Furthermore, if she needed QUBI to complete a series of tasks quickly, she would point and say, *"Please turn on the fan, turn off the light, open the fridge, now the coffee machine."* all while pointing in different directions. Finally, we also observed a moment in which **P7** apologised to QUBI, as she said in the practice scenario *"Hey QUBI, do you know who you are? ... where you are? Sorry."* As she said *"where you are?"*, she pointed at QUBI.

7.3 Theme 3: Anthropomorphism and Personality

We noticed that the participants anthropomorphised QUBI to a varying extent throughout the sessions. We asked the participants if they associated QUBI with anything. Several participants associated QUBI with characters from popular movies. Both **P1** and **P12** associated QUBI with the Star Wars robot characters BB-8 and R2D2, while **P9** thought that it would be fun to add Darth Vader's voice to QUBI. Similarly, **P7**, **P10** and **P11** associated QUBI with the little robot Wall-E from the movie with the same name. The rest of the participants associated QUBI with either some other specific technologies, games, movies, or generally humanlike behaviour. Moreover, **P1**, **P3**, and **P12** consistently referred to QUBI as 'he', even after **P3** and **P12** caught themselves do that, they continued afterwards again. **P3** even said *"...I am saying 'he', perhaps it's a 'her'."* while **P12** said *"He is pleasant. Generic 'he'. It's pleasant..."* **P7** and **P8** occasionally referred to QUBI as 'he' as well.

P5 viewed QUBI as a *"helper"* and a *"servant"*. Conversely, **P5** also talked about how he viewed the interactions with QUBI as a collaboration between the two of them. He was not doing these tasks just for himself but also for QUBI. On a similar note, **P3** said, *"When you get to [know] him, it becomes almost like a relationship even though it might sound awkward, but you do when you are working together."* Similarly, **P12** found it more collaborative to interact with QUBI than with commercially available smart speakers: *"The motions made me more easily forget that it was a machine and more easily forgiving it. I would go through the motions as if I was interacting with a person."* and elaborating that: *"It's a matter of making the experience more humanlike and understandable. [...] But he is actually teaching me. If I was an elderly person, he could teach me what he can do by turning and pointing at the appliances."*

For **P12**, this was a refreshing experience and he thought that the added motions were a step in the right direction because *"... it has a presence in the room. It feels like it's interacting more with the user and trying to convey a feeling of 'I have more than a rudimentary knowledge of the room.'"* As he pointed out earlier, he believed that elderly individuals would have an easier time interacting and understanding QUBI than existing smart speakers, and he thought that this would apply to children as well, since both groups would benefit more from tactile and visual cues than just auditory cues. On a similar note, **P1** also felt that QUBI had a stronger presence in the room due to its motions and personality, as she compared Google Home Mini with QUBI, she said: *"While Google Home Mini looks more nice and compact, I would still prefer QUBI because it would be nice to have a sentient robot in my house. I feel like it has much more of a personality than an Alexa or something like a Google Home Mini. The voice does a lot too in terms of personality but with added movements it's just even more."* These associations of personality and projecting human-like characters onto QUBI resonates with previous research on people's interactions with social robots [6, 18, 67, 75].

On the other hand, **P8** did not pay attention to QUBI's motions because she viewed QUBI as an inanimate object with no need to express its feelings, like a human or animal. While she did associate character and traits in QUBI's motions during the introduction of the motions, she had a clear image of QUBI as a synthetic and artificial design. She would have preferred that QUBI was more humanlike with similarities to a doll that has more details than mere cubes.

Participants found it generally fun that QUBI had some humour and banter as part of some of its responses, in particular during the practice scenario, and when it was mixing up carrots, captains, and cannons, as the two latter ones are not edible. **P7** was among the most disappointed participants when the experimenter told her that it was a Wizard of Oz study, as she responded with *"Nooo. Oh my god, this is so sad. I got so connected with QUBI."*

7.4 Theme 4: Audio Can Trump Motion

During the participants' interactions with QUBI, we observed that five participants would look at QUBI, wait for a verbal response, but then miss the physical motion that QUBI would perform as they would look away due to the verbal confirmation. **P2** did this when he added napa cabbage to the shopping list, since QUBI's nod was slightly delayed after it said *"Ok"* in *"Ok, I have added napa cabbage to your shopping list."* We observed this seven times across other similar instance with five participants. Six out of the seven times, the participants would miss a nod, while the last one was **P1**

missing a shrug due to QUBI not understanding *"Squared spring roll wrappers"*.

On a different note, **P2** preferred audio to visual cues and said in the interview that he technically did not find any of QUBI's motions important, as that would tether him to the room where QUBI is located. He said that *"Audio is the most important, but if you want to add movements, then they need to follow an audio cue."* Similarly, **P5**, **P9** and **P10** all thought that it would have been great if QUBI would have told them to come closer in addition to when it leaned forward and wanted the participants to approach it. **P5** suggested, *"I am not sure, can you speak up?"* to indicate to the user to speak up, or get closer to QUBI if needed.

7.5 Theme 5: Reaffirming Uncertain Interpretations to Support Mutual Understanding

During the scenarios, each participant experienced QUBI offering two alternative interpretations three times. All participants liked that QUBI offered alternatives for parts of their requests, to help them identify what QUBI was uncertain about understanding. As **P10** said when she commented on QUBI suggesting captains and carrots: *"I knew that it knew that we were talking about two different things. That I was talking about 'carrots' and it about 'captains', so it asked 'Wait, did you mean captains or carrots?' It would have been less aware if it immediately suggested carrots, then it would not be as clear that it was aware of there being two different things."* She pointed out how clearly and transparently QUBI processed her request.

P4, **P7**, **P8**, and **P12** all had comments on QUBI's contextual awareness of the requests. Both **P4** and **P7** thought that QUBI should know the context and only offer options relevant to the context such as recipes. As **P7** said, *"When I talk about shopping lists, it could perhaps learn and associate words or typical items that I would buy, to the shopping list."* On the other hand, **P8** and **P12** found it humorous that QUBI offered captains as an option even though captains are not edible as are carrots, but still proceeded to complete the task. Along the same thought as **P4** and **P7**, **P6** found it important that QUBI would offer his choice, to avoid any unnecessary repetition. It would be less helpful, if QUBI offered two options, none of which he wanted.

Interestingly, when QUBI asked whether **P12** meant *"squared spring stone bladders"* or *"squared spring roll wrappers"*, **P12** first thought about using a number to indicate a choice, but instead decided to test if QUBI accepted less common words, such as saying *"the latter"*. **P12** said: *"(…) He offered me two options and instead of risking him misunderstanding me again, I wanted to see if he could differentiate between the numbers. But then I used 'the latter' being somewhat of a not commonly used [word], I should think."* Contrary to these opinions, **P11** thought that it was unnatural to say *"Option 1"* in a conversation, and since repeating the same word could potentially not be understood by the digital assistant, he found it difficult to imagine that this functionality would work. **P4** disagreed with this, as he said, *"If it suggests two things, it would be nice to get to the suggestions [faster], by selecting either the first or last thing without saying the name of it."* **P4** thought it felt more like a dialogue in that case. For the three times that QUBI offered alternatives in the study, **P11** switched between repeating the item suggested by QUBI and the ordinal numeral corresponding to their choice (e.g. first or second), while **P4** only used the full name of the item. In 12 out of 36 cases, participants used phrases such as *"The second one"*, *"The latter"* and *"No. 2"*, of which five were a mixture of those phrases and the name of the item requests, such as *"I meant carrots, the second: carrots."* The 12 instances were spread across eight participants.

7.6 Theme 6: Emotional Reactions to QUBI's Behaviour

We observed how some participants reacted emotionally towards QUBI's behaviour, both positively and negatively.

Note that in scenario two, in which participants had to add carrots to the shopping list, QUBI did not understand the request until the participants approached it. This was a potentially frustrating moment for all participants. **P7** spent nearly 2 minutes requesting QUBI to add carrots to the shopping. After four attempts, **P7** said, *"So I will just write it down myself."* As she was then reminded that the tasks had to be completed, she tried two more times, one of which she spelled out the word *"Carrots"*, before saying *"I am confused"*. As she was given the hint about getting closer to QUBI, she experienced one more obstacle, which was that QUBI added *captains* to the shopping list instead of *carrots*. She quickly replied, *"No, no, not captains, please add carrots."* followed by QUBI offering her the option to choose between *cannons* and *carrots*. As QUBI confirmed her choice, *carrots*, she turned around, threw her fists in the air, and dropped them while whispering *"Yes"*, expressing her relief. All participants either frowned, raised their eyebrows, looked disappointed and confused, or scratched their jaws in a variety of instances, but predominantly when QUBI did not understand carrots and when it added captains to the shopping list instead.

In regard to positive reactions, ten participants either smiled, chuckled or laughed during some of the interactions with QUBI. More notably, **P2** and **P4** laughed when QUBI added captains instead of carrots, while **P6** chuckled, and similarly **P9** chuckled when QUBI responded to **P9**'s request of removing captains from the shopping list with *"Sorry, I can't find any carrots in your shopping list."* As we will describe further down, some participants mentioned how QUBI's humour made frustrating moments such as misunderstandings, more forgiving.

Furthermore, **P3** was quite enthusiastic about some moments in which QUBI completed a task after struggling a few times. Throughout the scenarios, **P3** would throw his hands in the air six times and say things such as *"You are brilliant!"*, *"Perfect!"*, and *"Thanks!"*

During a request in which QUBI had to turn off the living room lamp and turn on the reading lamp, **P8** did not pay attention to QUBI's motions nor the appliances and stared, like she was lost in thought. Suddenly she looked at the experimenter, and whispered, *"It almost feels rude to not say 'thank you'."* and laughed a bit. After the experimenter told her that she could express herself however she felt in the interactions with QUBI, she looked at QUBI and said *"Thank you QUBI, you are so nice."* She explained in the interview: *"I felt really bad because I am usually very thankful. When someone or something is doing what I am asking, it is compelling to me to say 'Thank you'."*

Another aspect was QUBI's *"You're welcome"* reply, which was triggered by the operator when the participants said *"Thank you"* to QUBI. With three participants, it was clear that they appreciated

QUBI's politeness with either a smile and/or a nod. This was in contrast to **P8**'s opinion as she found QUBI to not be polite. Note, QUBI never said *"You're welcome!"* because the operator did not hear her *"Thank you."*, QUBI was not in ready mode when she said it, and she said it rarely. Similar to the above-mentioned examples of emotional reactions, prior work has also observed emotional reactions in people's interactions with other abstract robots [6, 18] as well as furniture-like robots [67].

7.7 Limitations

Our study is limited in a number of ways. We deliberately recruited participants from different backgrounds, but they were all recruited from one city. It may be that people from other parts of the world would view and perceive QUBI differently, depending on how gestures and body languages are construed. For example, gestures in South American countries may be understood differently to those from Northern European countries. Given that participants went through a set of four scripted scenarios, they were not able to interact with and experience QUBI in a fully open-ended way, which may have influenced our results. The setup of the room and the standing table, may have influenced people's movements in the room. Similarly, our use of hints when people got stuck may have biased people in responding in a certain way. More specifically, when participants missed some of QUBI's motions, a firmer approach to obtaining the participants' gaze [31, 47] prior to QUBI's important motions could have allowed participants to observe those motions, and maybe have reduced the missed observations. While none of the participants realized that QUBI was remote controlled by an operator, some commented on QUBI being slow to respond. This was likely due to the slight delay in the operator selecting the correct command to send to QUBI to respond to the participants' requests. Finally, our nine motions should not be considered an exhaustive list but rather be seen as a starting point to investigate more and different motions to augment smart speakers.

8 DISCUSSION

Our findings have demonstrated how the use of small movements in a smart speaker prototype can complement existing voice interactions in a meaningful way. Our initial vocabulary of nine dynamic and expressive physical motions showed promise in engaging and disengaging with users, emphasizing the smart speaker's physical presence, providing additional information about the smart speaker's state and ongoing activities, providing contextually appropriate responses, and even delighting individuals with cues that are reminiscent of our own social protocols. This approach suggests benefits of having expressive smart speakers with limited humanlike body movements, like nodding and pointing, that can overcome some of the limitations of relying solely on voice interaction and can help resolve misunderstandings. Next, we discuss three future research directions in the context of smart speakers.

8.1 Machine Mannerisms: Towards Machine Body Language

Our study has shown that smart speakers can borrow limited humanlike behaviour and translate that into machine body language that can assist users in communicating with the smart speaker and better understanding its intent. Our participants did not report a negative experience with QUBI's motions. While some said they would be fine if QUBI did not use physical motion, the others felt that QUBI's physical motions added more value to the whole experience of interacting with it. The $M1_{ready}$, $M9_{idle}$, $M2_{nod}$, $M4_{shrug}$, $M3_{shake}$, $M7_{wiggle}$ and $M8_{point}$ motions were the most relateable and understandable motions, while $M5_{forward}$ and $M6_{backward}$ were the most difficult to understand.

This begs the question where we go from here. Our findings show promise in the use of physical motion in smart speakers. QUBI's initial physical motion vocabulary and our findings are a starting point that other smart speaker researchers can build upon to further explore this rich space. An interesting direction would be to expand on the vocabulary of physical motions, by exploring additional motions, either to convey additional information or to investigate alternative designs to the physical cues we designed for QUBI. In particular, exploring alternative designs for QUBI's leaning forward and backward motions that do have the intended effect would be worthy of further study. Perhaps a more ambitious goal would be to investigate whether it is possible to develop a compendium of machine mannerisms that people could readily learn. Additional studies could investigate the mechanisms that underlie the intelligibility of these physical cues. Moreover, physical motion cues could also be expanded from smart speakers to other IoT devices alike. Existing work in this area has explored the use of physical motion for smart thermostats [74], autonomous furniture [67], and robot vacuums [30]. Could the use of machine body language make future IoT devices more enjoyable to interact with, more transparent, and more engaging?

8.2 Mutual Body Language Awareness

While existing smart speakers make little to no use of physical motion or "machine body language", they equally do not tend to take into account aspects of the user's body language and rely mostly on the user's voice input. One exception is Amazon's Echo, which senses the direction of the user's voice and indicates with a blue LED ring from which direction it heard a request. However, our study indicates that many participants used body language to complement their voice instructions and even mimicked and mirrored QUBI's physical motions. Participants either used body language to add additional or redundant information to their request [22], which was particularly evident during conversational breakdowns. This provides an interesting opportunity for smart speakers to sense and respond to users' body language. Body language could be used as a resource to help disambiguate what users request when they interact with the device (e.g. pointing at the lamp they want to turn on) and to anticipate when users encounter unexpected responses. Our study contributes further evidence to the potential of research in this area, such as smart speakers that can respond to users' facial expressions [78] and gaze [52].

We also observed that some participants failed to see the physical motions as they stopped looking at QUBI once they had made their request. This suggests that QUBI's body language may not necessarily get the right amount of attention and time to convey its intent to the user. However, it may again show that there would be benefits to *QUBI* being more attentive to the users' body

language, their gaze and their orientation towards QUBI. Similar to a conversation between two people, where it is crucial to maintain eye contact and establish mutual shared attention, QUBI could take into account the attention of its user(s).

It would be an interesting avenue to investigate to which extent a user's and smart speaker's mutual gaze [52] could make physical motions stand out during interactions between users and smart speakers. While we acknowledge that users should not feel forced to look at QUBI at all times, we do believe that QUBI would have an advantage in making decisions with respect to when to use gestures and body language by also reading the user's body language: *mutual body language awareness*. And in case users fail to notice QUBI's gestures and body language, QUBI could supplement the body language with additional speech to clarify its intent. This could be done in a situation where QUBI is in the midst of pointing at an appliance or leaning forward, and it notices that the user is not paying attention, QUBI could then supplement its physical motion on the spot with a voice response, such as saying *"Over there."* and *"I can't hear you, could you please get closer?"* respectively.

8.3 Teaching Machine Mannerisms through Direct Manipulation

As was pointed out by one of our participants, QUBI's stronger physical presence compared to existing smart speakers provided more visual cues based on familiar interactions. Examples mentioned include how QUBI points at other devices, how it indicates uncertainty with its shrug motion and how it offers suggestions to repair the ongoing communication. QUBI's motions can be interpreted within a particular spatial context such as when it is pointing at a particular appliance, which provides additional information about its intent. Some participants speculated on the possibility of directly manipulating QUBI to make it turn on a lamp, or program it so it knows where the lamp is. This could be a potential research direction to explore further with respect to spatial interaction [39]. This also points to exciting possibilities in making QUBIs movements configurable by its users through the use of direct manipulation. This would allow users to record and demonstrate motions that QUBI can then play back and use in the appropriate situation, similar to Topobo [62] or ClicBot [44]. While today's smart speakers merely use buttons for limited interaction, QUBI might be a source of inspiration in opening up alternative interaction possibilities through physical and spatial interaction. Such alternative physical and spatial interactions could potentially be used to make smart speakers more accessible to non-primary users such as guests and possibly also first-time users who tend to be less familiar with smart speakers [1, 8].

9 CONCLUSION

Our research has shown how adding physical motion to a smart speaker's form can make its conversational interaction more expressive and make its activities within an IoT ecosystem intelligible. The physical motion of QUBI – a smart speaker prototype – was perceived to be meaningful and helpful in understanding its underlying behaviour. A Wizard-of-Oz study in which participants were able to experience interacting with QUBI, revealed six key findings in terms of participants' behaviours: (1) mirroring and mimicking motions; (2) body language to supplement voice instructions; (3) anthropomorphism and personality; (4) the ability of audio to trump motion; (5) reaffirming uncertain interpretations to support mutual understanding; and (6) emotional reactions to QUBI's behaviour. The research suggests that it is possible to derive an initial "machine mannerism" vocabulary for smart speakers, consisting of a set of expressive physical motions that are readily distinguishable and complement and enrich the existing conversational interaction. This kind of machine etiquette is intended to inspire researchers to think about alternative approaches to designing our interactions with smart technology.

ACKNOWLEDGMENTS

We thank our participants for their contribution in making this work happen. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 740548). This project was approved by the Institutional Review Board at Aarhus University (serial number: 2020-12).

REFERENCES

[1] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J. Lee. 2020. Tangible Privacy: Towards User-Centric Sensor Designs for Bystander Privacy. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 116 (Oct. 2020), 28 pages. https://doi.org/10.1145/3415187

[2] M. R. Alam, M. B. I. Reaz, and M. A. M. Ali. 2012. A Review of Smart Homes—Past, Present, and Future. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 1190–1203. https://doi.org/10.1109/TSMCC.2012.2189204

[3] Alper T. Alan, Enrico Costanza, Sarvapali D. Ramchurn, Joel Fischer, Tom Rodden, and Nicholas R. Jennings. 2016. Tariff Agent: Interacting with a Future Smart Energy System at Home. *ACM Trans. Comput.-Hum. Interact.* 23, 4 (Aug. 2016), 25:1–25:28. https://doi.org/10.1145/2943770

[4] Amazon. 2021. Amazon Echo. https://www.amazon.com/dp/B085HK4KL6?ref=MarsFS_AUCC_lr

[5] Amazon. 2021. Amazon Echo Show 10. https://www.amazon.com/dp/B07VHZ41L8?ref=ods_ucc_aucc_ta_nrc_ucc

[6] L. Anderson-Bashan, B. Megidish, H. Erel, I. Wald, G. Hoffman, O. Zuckerman, and A. Grishko. 2018. The Greeting Machine: An Abstract Robotic Object for Opening Encounters. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 595–602. https://doi.org/10.1109/ROMAN.2018.8525516

[7] Apple. 2021. Apple HomePod. https://www.apple.com/homepod/

[8] Mirzel Avdic and Jo Vermeulen. 2020. Intelligibility Issues Faced by Smart Speaker Enthusiasts in Understanding What Their Devices Do and Why. In *OZCHI'20: Proceedings of the 32nd Australian Conference on Human-Computer-Interaction (OzCHI '20)*. ACM, New York, NY, USA. https://doi.org/10.1145/3441000.3441068 (in press).

[9] Till Ballendat, Nicolai Marquardt, and Saul Greenberg. 2010. Proxemic Interaction: Designing for a Proximity and Orientation-Aware Environment. (2010).

[10] Geoffrey Beattie. 2003. *Visible Thought: The New Psychology of Body Language*. Routledge.

[11] Victoria Bellotti and Keith Edwards. 2001. Intelligibility and Accountability: Human Considerations in Context-Aware Systems. *Human–Computer Interaction* 16, 2-4 (Dec. 2001), 193–212. https://doi.org/10.1207/S15327051HCI16234_05

[12] Erin Beneteau, Olivia K. Richards, Mingrui Zhang, Julie A. Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication Breakdowns Between Families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, 243:1–243:13. https://doi.org/10.1145/3290605.3300473 event-place: Glasgow, Scotland Uk.

[13] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottrjdge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (Sept. 2018), 1–24. https://doi.org/10.1145/3264901

[14] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative HCI Research: Going Behind the Scenes*. Morgan & Claypool. https://ieeexplore.ieee.org/document/7450876

[15] T. Boren and J. Ramey. 2000. Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication* 43, 3 (Sept. 2000), 261–278. https://doi.org/10.1109/47.867942

[16] Cynthia Breazeal. 2017. Social Robots: From Research to Commercialization. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/2909824.3020258

[17] C. Breazeal, C. D. Kidd, A. L. Thomaz, G. Hoffman, and M. Berlin. 2005. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 708–713. https://doi.org/10.1109/IROS.2005.1545011

[18] Mason Bretan, Guy Hoffman, and Gil Weinberg. 2015. Emotionally expressive dynamic physical behaviors in robots. *International Journal of Human-Computer Studies* 78 (2015), 1–16. https://doi.org/10.1016/j.ijhcs.2015.01.006

[19] E. J. Carter, M. N. Mistry, G. P. K. Carr, B. A. Kelly, and J. K. Hodgins. 2014. Playing catch with robots: Incorporating social gestures into physical interactions. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 231–236. https://doi.org/10.1109/ROMAN.2014.6926258

[20] Matthew Chalmers. 2003. Seamful Design and Ubicomp Infrastructure. *Proceedings of Ubicomp 2003 workshop at the crossroads: The interaction of HCI and systems issues in Ubicomp* (2003).

[21] Minji Cho, Sang-su Lee, and Kun-Pyo Lee. 2019. Once a Kind Friend is Now a Thing: Understanding How Conversational Agents at Home Are Forgotten. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. ACM, New York, NY, USA, 1557–1569. https://doi.org/10.1145/3322276.3322332 event-place: San Diego, CA, USA.

[22] Herbert Clark. 2003. Pointing and Placing. In *Pointing: where language, culture, and cognition meet*, Sotaro Kita (Ed.). L. Erlbaum Associates, Mahwah, N.J, 243–268.

[23] Herbert H. Clark and Susan E. Brennan. 1991. Grounding in Communication. In *Perspectives on Socially Shared Cognition*, Lauren Resnick, Levine B, M. John, Stephanie Teasley, and D. (Eds.). American Psychological Association, 127–149.

[24] K. Dautenhahn, S. Woods, C. Kaouri, M. L. Walters, Kheng Lee Koay, and I. Werry. 2005. What is a robot companion - friend, assistant or butler?. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 1192–1197. https://doi.org/10.1109/IROS.2005.1545189

[25] Ruta Desai, Fraser Anderson, Justin Matejka, Stelian Coros, James McCann, George Fitzmaurice, and Tovi Grossman. 2019. Geppetto: Enabling Semantic Design of Expressive Robot Behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI '19*. ACM Press, Glasgow, Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300599

[26] NTT Disruption. 2021. Jibo. https://jibo.com

[27] Brian R. Duffy. 2003. Anthropomorphism and the social robot. *Robotics and Autonomous Systems* 42, 3-4 (March 2003), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

[28] Emotech. 2021. Olly. https://www.indiegogo.com/projects/olly-the-first-home-robot-with-personality#/

[29] Luis A. Fuente, Hannah Ierardi, Michael Pilling, and Nigel T. Crook. 2015. Influence of Upper Body Pose Mirroring in Human-Robot Interaction. In *Social Robotics*, Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland, and Mehdi Ammi (Eds.). Vol. 9388. Springer International Publishing, Cham, 214–223. https://doi.org/10.1007/978-3-319-25554-5_22 Series Title: Lecture Notes in Computer Science.

[30] Pedro Garcia Garcia, Enrico Costanza, Sarvapali D. Ramchurn, and Jhim Kiel M. Verame. 2016. The Potential of Physical Motion Cues: Changing People's Perception of Robots' Performance. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 510–518. https://doi.org/10.1145/2971648.2971697 event-place: Heidelberg, Germany.

[31] Charles Goodwin. 1980. Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. *Sociological Inquiry* 50, 3-4 (1980), 272–302. https://doi.org/10.1111/j.1475-682X.1980.tb00023.x

[32] Google. [n.d.]. Cloud Text-to-Speech. https://cloud.google.com/text-to-speech/

[33] Google. 2021. Nest Audio. https://store.google.com/com/product/nest_audio

[34] Saul Greenberg, Nicolai Marquardt, Till Ballendat, Rob Diaz-Marino, and Miaosen Wang. 2011. Proxemic interactions: the new ubicomp? *interactions* 18, 1 (Jan. 2011), 42. https://doi.org/10.1145/1897239.1897250

[35] Edward T. Hall. 1966. *The hidden dimension*. Vol. 609. Garden City, NY: Doubleday.

[36] John Harris and Ehud Sharlin. 2011. Exploring the affect of abstract motion in social human-robot interaction. In *2011 RO-MAN*. IEEE, Atlanta, GA, USA, 441–448. https://doi.org/10.1109/ROMAN.2011.6005254

[37] Fritz Heider and Marianne Simmel. 1944. An Experimental Study of Apparent Behavior. *The American Journal of Psychology* 57, 2 (April 1944), 243. https://doi.org/10.2307/1416950

[38] Guy Hoffman and Wendy Ju. 2014. Designing Robots with Movement in Mind. *J. Hum.-Robot Interact.* 3, 1 (Feb. 2014), 91–122. https://doi.org/10.5898/JHRI.3.1. Hoffman

[39] Eva Hornecker and Jacob Buur. 2006. Getting a Grip on Tangible Interaction: A Framework on Physical Space and Social Interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 437–446. https://doi.org/10.1145/1124772.1124838 event-place: Montréal, Québec, Canada.

[40] Kristiina Jokinen and Jens Allwood. 2010. Hesitation in Intercultural Communication: Some Observations and Analyses on Interpreting Shoulder Shrugging. In *Culture and Computing*, Toru Ishida (Ed.). Vol. 6259. Springer Berlin Heidelberg, Berlin, Heidelberg, 55–70. https://doi.org/10.1007/978-3-642-17184-0_5 Series Title: Lecture Notes in Computer Science.

[41] Wendy Ju and Leila Takayama. 2009. Approachability: How People Interpret Automatic Door Movement as Gesture. *International Journal of Design* Vol. 3(2), Design and Emotion (Aug. 2009), 15.

[42] J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)* 2, 1 (Jan. 1984), 26–41. https://doi.org/10.1145/357417.357420

[43] Adam Kendon. 1988. How gestures can become like words. In *Cross-cultural perspectives in nonverbal communication*, Fernando Poyatos (Ed.). Hogrefe, Toronto ; Lewiston, N.Y, 131–141.

[44] Ltd. KEYi Technology Co. 2020. Clicbot. https://clicbot.keyirobot.com

[45] Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy N. Bailenson, Gerd Bruder, and Gregory F. Welch. 2018. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 105–114. https://doi.org/10.1109/ISMAR.2018.00039

[46] Tiiu Koskela and Kaisa Väänänen-Vainio-Mattila. 2004. Evolution towards smart home environments: empirical evaluation of three user interfaces. *Personal and Ubiquitous Computing* 8, 3 (July 2004), 234–240. https://doi.org/10.1007/s00779-004-0283-x

[47] Hideaki Kuzuoka, Karola Pitsch, Yuya Suzuki, Ikkaku Kawaguchi, Keiichi Yamazaki, Akiko Yamazaki, Yoshinori Kuno, Paul Luff, and Christian Heath. 2008. Effect of Restarts and Pauses on Achieving a State of Mutual Orientation between a Human and a Robot. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work* (San Diego, CA, USA) *(CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 201–204. https://doi.org/10.1145/1460563.1460594

[48] Iolanda Leite, Andre Pereira, Carlos Martinho, and Ana Paiva. 2008. Are emotional robots more fun to play with?. In *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*. 77–82. https://doi.org/10.1109/ROMAN.2008.4600646

[49] Irene Lopatovska and Harriet Williams. 2018. Personification of the Amazon Alexa: BFF or a Mindless Companion. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval - CHIIR '18*. ACM Press, New Brunswick, NJ, USA, 265–268. https://doi.org/10.1145/3176349.3176868

[50] M. Luria, G. Hoffman, B. Megidish, O. Zuckerman, and S. Park. 2016. Designing Vyo, a robotic Smart Home assistant: Bridging the gap between device and social agent. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 1019–1025. https://doi.org/10.1109/ROMAN.2016.7745234

[51] Nicolai Marquardt and Saul Greenberg. 2015. *Proxemic Interactions from Theory to Practice*. (1 ed.). Morgan & Claypool, San Rafael; Plymouth. http://VH7QX3XE2P.search.serialssolutions.com/?V=1.0&L=VH7QX3XE2P&S=JCs&C=TC0001578585&T=marc&tab=BOOKS OCLC: 1066604696.

[52] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: Tama – a Gaze Activated Smart-Speaker. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 1–26. https://doi.org/10.1145/3359278

[53] M. Mori, K. F. MacDorman, and N. Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics Automation Magazine* 19, 2 (2012), 98–100.

[54] Desmond Morris. 2002. *Peoplewatching* (rev. ed ed.). Vintage, London. OCLC: 248883463.

[55] Christine Murad, Cosmin Munteanu, Benjamin R. Cowan, and Leigh Clark. 2019. Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing* 18, 2 (April 2019), 33–45. https://doi.org/10.1109/MPRV.2019.2906991

[56] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–7. https://doi.org/10.1145/3173574.3173580

[57] Diana Nowacka, Katrin Wolf, Enrico Costanza, and David Kirk. 2018. Working with an Autonomous Interface: Exploring the Output Space of an Interactive Desktop Lamp. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction - TEI '18*. ACM Press, Stockholm, Sweden, 1–10. https://doi.org/10.1145/3173225.3173227

[58] K. Park, H. Lee, Y. Kim, and Z. Z. Bien. 2008. A Steward Robot for Human-Friendly Human-Machine Interaction in a Smart House Environment. *IEEE Transactions on Automation Science and Engineering* 5, 1 (2008), 21–25. https://doi.org/10.1109/TASE.2007.911674

[59] Allan Pease and Barbara Pease. 2006. *The Definitive Book of Body Language*. Routledge.

[60] Avital Pekker. 2016. A closer look at Apple's breathing light. https://avital.ca/notes/a-closer-look-at-apples-breathing-light

[61] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice Interfaces in Everyday Life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, 640:1–640:12. https://doi.org/10.1145/3173574.3174214

[62] Hayes Solos Raffle, Amanda J. Parkes, and Hiroshi Ishii. 2004. Topobo: A Constructive Assembly System with Kinetic Memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 647–654. https://doi.org/10.1145/985692.985774

[63] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press, New York, NY, USA.

[64] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I. Hong. 2018. "Hey Alexa, What's Up?": A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 857–868. https://doi.org/10.1145/3196709.3196772

[65] Leon Segal. 1995. Designing Team Workstations: The Choreography. In *Local applications of the ecological approach to human-machine systems*, Peter A. Hancock (Ed.). L. Erlbaum Associates, Hillsdale, N.J, 392–415.

[66] Helen Sharp, Jennifer Preece, and Yvonne Rogers. 2019. *Peoplewatching* (5th ed ed.). Wiley.

[67] David Sirkin, Brian Mok, Stephen Yang, and Wendy Ju. 2015. Mechanical Ottoman: How Robotic Furniture Offers and Withdraws Support. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*. ACM Press, Portland, Oregon, USA, 11–18. https://doi.org/10.1145/2696454.2696461

[68] Marie Louise Juul Søndergaard and Lone Koefoed Hansen. 2018. Intimate Futures: Staying with the Trouble of Digital Personal Assistants through Design Fiction. In *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*. ACM Press, Hong Kong, China, 869–880. https://doi.org/10.1145/3196709.3196766

[69] Leila Takayama, Doug Dooley, and Wendy Ju. 2011. Expressing Thought: Improving Robot Readability with Animation Principles. In *Proceedings of the 6th International Conference on Human-Robot Interaction* (Lausanne, Switzerland) *(HRI '11)*. Association for Computing Machinery, New York, NY, USA, 69–76. https://doi.org/10.1145/1957656.1957674

[70] Deborah L. Trout and Howard M. Rosenfeld. 1980. The Effect of Postural Lean and Body Congruence on the Judgment of Psychotherapeutic Rapport. *Journal of Nonverbal Behaviour* (1980).

[71] C. Tsai, S. Hsieh, Y. Hsu, and Y. Wang. 2009. Human-robot interaction of an active mobile robotic assistant in intelligent space environments. In *2009 IEEE International Conference on Systems, Man and Cybernetics*. 1953–1958. https://doi.org/10.1109/ICSMC.2009.5346049

[72] Blase Ur, Elyse McManus, Melwyn Pak Yong Ho, and Michael L. Littman. 2014. Practical trigger-action programming in the smart home. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*. ACM Press, Toronto, Ontario, Canada, 803–812. https://doi.org/10.1145/2556288.2557420

[73] Blase Ur, Melwyn Pak Yong Ho, Stephen Brawner, Jiyun Lee, Sarah Mennicken, Noah Picard, Diane Schulze, and Michael L. Littman. 2016. Trigger-Action Programming in the Wild: An Analysis of 200,000 IFTTT Recipes. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. ACM Press, Santa Clara, California, USA, 3227–3231. https://doi.org/10.1145/2858036.2858556

[74] Anke van Oosterhout, Miguel Bruns Alonso, and Satu Jumisko-Pyykkö. 2018. Ripple Thermostat: Affecting the Emotional Experience through Interactive Force Feedback and Shape Change. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. ACM Press, Montreal QC, Canada, 1–12. https://doi.org/10.1145/3173574.3174229

[75] Steve Whittaker, Yvonne Rogers, Elena Petrovskaya, and Hongbin Zhuang. 2021. Designing Personas for Expressive Robots: Personality in the New Breed of Moving, Speaking, and Colorful Social Home Robots. *J. Hum.-Robot Interact.* 10, 1, Article 8 (Feb. 2021), 25 pages. https://doi.org/10.1145/3424153

[76] Kyle Wiggers. 2018. Alexa recorded a woman's private conversation and sent it to a random contact. https://venturebeat.com/2018/05/24/alexa-recorded-a-womans-private-conversation-and-sent-it-to-a-random-contact/

[77] Terry Winograd, Fernando Flores, and Fernando F. Flores. 1986. *Understanding Computers and Cognition: A New Foundation for Design*. Intellect Books. Google-Books-ID: 2sRC8vcDYNEC.

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