

Designing with Gaze

Tama – a Gaze-Aware Smart Speaker Platform

DONALD MCMILLAN, DSV, Stockholm University, Sweden

BARRY BROWN, DSV, Stockholm University, Sweden

IKKAKU KAWAGUCHI, University of Tsukuba, Japan

RAZAN JABER, DSV, Stockholm University, Sweden

JORDI SOLSONA BELENGUER, DSV, Stockholm University, Sweden

HIDEAKI KUZUOKA, The University of Tokyo, Japan



Fig. 1. The Tama Gaze-Aware Smart Speaker Platform

Recent developments in gaze tracking present new opportunities for social computing. This paper presents a study of Tama, a gaze actuated smart speaker. Tama was designed taking advantage of research on gaze in conversation. Rather than being activated with a wake word (such as “Ok Google”) Tama detects the gaze of a user, moving an articulated ‘head’ to achieve mutual gaze. We tested Tama’s use in a multi-party conversation task, with users successfully activating and receiving a response to over 371 queries (over 10 trials). When Tama worked well, there was no significant difference in length of interaction. However, interactions with Tama had a higher rate of repeated queries, causing longer interactions overall. Video analysis lets us explain the problems users had interacting with gaze. In the discussion, we describe implications for designing new gaze systems, using gaze both as input and output. We also discuss how the relationship to anthropomorphic design and taking advantage of learned skills of interaction. Finally, two paths for future work are proposed, one in the field of speech agents, and the second in using human gaze as an interaction modality more widely.

CCS Concepts: • **Human-centered computing** → **Interaction techniques**; *User studies*; *Interaction design*;

Additional Key Words and Phrases: Smart Speaker; Voice Assistant; Gaze Interaction; Gaze Detection

Authors’ addresses: Donald McMillan, DSV, Stockholm University, Stockholm, Sweden, donald.mcmillan@dsv.su.se; Barry Brown, DSV, Stockholm University, Stockholm, Sweden, barry@dsv.su.se; Ikkaku Kawaguchi, University of Tsukuba, Tsukuba, Ibaraki, Japan, kawaguchi@cs.tsukuba.ac.jp; Razan Jaber, DSV, Stockholm University, Stockholm, Sweden, razan@dsv.su.se; Jordi Solsona Belenguer, jordi@dsv.su.se, DSV, Stockholm University, Stockholm, Sweden; Hideaki Kuzuoka, kuzuoka@acm.org, The University of Tokyo, Bunkyo-ku, Tokyo, Japan.

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

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2573-0142/2019/11-ART176 \$15.00

<https://doi.org/10.1145/3359278>

ACM Reference Format:

Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: *Tama* – a Gaze-Aware Smart Speaker Platform. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 176 (November 2019), 26 pages. <https://doi.org/10.1145/3359278>

1 INTRODUCTION

Speech agents such as Alexa, Siri, or Google Assistant have become increasingly prevalent in our daily life. While these systems have many strengths, it is clear that they also suffer from high error rates, and a host of interaction problems when they are used in real-world settings [66, 73]. One issue concerns their lack of understanding of gaze or gesture. Conversation between humans makes use of gaze for many essential features, such as the regulation of turns in talk [59], speaker selection [50], as well as a host of explicit communicative features such as deixis [31]. Gaze is not an additional communication channel that is overlaid upon, or that is additional to talk. Rather talk and gaze work in concert as a way of both speakers and listeners to co-ordinate their attention, actions, and communication. Indeed, it has been suggested that the loss of this resource is one reason why audio-only conference calls are more fractured, particularly around managing attention and turn-taking [74, 82].

In this paper, we discuss our attempts to address some of these through adding gaze awareness and feedback to a speech agent system — building ‘Tama’, a gaze and voice aware speech agent. This system reacts to the gaze of a user or users, and instead of relying on a ‘wake word’ (such as ‘Alexa’ or ‘Hey, Google’) it makes use of gaze to trigger and confirm that a command was being spoken. It then uses the commercial Google Assistant service to provide an appropriate voice response. Through a movable head with an approximation of eyes, mutual gaze can be established and maintained with a user. The system can then be looked at, look back, and establish mutual gaze as part of speech interaction.

We studied the use of this system in both an experimental single-user, and semi-experimental multi-user setting. This allowed us to test the system in a realistic usage situation consisting of a variety of user-generated questions, non-system directed speech, and interaction between co-present others and the system. In tests, users could successfully interact and ask Tama questions, as well as receive answers. The system had a low rate of false positives (where it triggered when talk was not being directed at Tama). Yet overall, the performance of the system was less than that of an original speech agent. Users took around 29% percent longer on their queries and had to repeat themselves to Tama 22% percent of the time. So while there were successes overall, the performance of our gaze enabled system was not as fast in use as a wake-word triggered agent.

This shows that adding gaze as a modality to a speech agent, certainly to replace the initial wake word, is harder than we first expected. While we had designed for gaze, drawing on the work of conversation and interactional analysis, problems with our design combined with problems with the accuracy of cameras, sensors and motors.

Close analysis of the interaction reveals interesting subtleties in how gaze is used to manage talk in interaction. We identified three problems users faced using gaze with our system. First, users would fail to establish initial gaze with the system, with prolonged ‘pre-’ phases where they attempted to establish mutual-gaze before speaking a command. In other words, users would not speak a command until mutual-gaze was established. Second, during talk with the system, the system would look away if it failed to detect the user’s continued engagement, causing the user to restart or abandon their query. While this is something that is often done in person to person speech to gain or maintain the attention of a speaker, in this case, it significantly slowed down interaction with the system.

Analysis of videos of interaction explains how these problems were experienced and adapted to by users. In particular, we document the adaption of gaze not as a natural ‘seen but unnoticed’ part of interaction by users, but something that came to be much more *explicitly* managed by users. Interestingly, this echoes results from how speech comes to be explicitly managed by users [66] in an attempt to get speech systems to work.

Documenting these problems generates a range of interesting issues more broadly for the design of gaze as an input method. In conclusion, we discuss some future directions for incorporating gaze into interactive systems subtly and advantageously. While there is much potential for gaze, as with speech itself, it cannot simply be bolted onto existing systems. Human interaction, gaze, and speech all work together in a total communicative system – understanding and designing for this offers exciting new possibilities for CSCW system design.

2 BACKGROUND

Here we focus on three areas of research, upon the nexus of which the work presented here resides. Gaze has long been seen as a rich resource for developing social robots, and in improving human-robot interaction. Gaze in human-human conversation has also been a longstanding area of research in conversation analysis and related fields. In our work, we combine the technical use and detection of gaze and its sociological understanding in interaction with speech agents, a field which combines a long history of research with a current focus on their use in the wild.

2.1 Gaze and Social Robotics

In the field of Human-Robot Interaction (HRI), researchers have highlighted the importance of non-verbal cues such as gaze and gesture when interacting with, and through, technology [10, 43]. Gaze has been extensively used to understand user behaviour in human-computer interaction [40, 64], to augment other input modalities [41], improve accessibility [48], as well as to enhance the social interaction of communicative social robots [1, 70].

Harnessing different roles that gaze plays in conversation and using that for system output has been the focus of a number of studies in HRI, including enabling agents to show attention to conversational partners and objects [e.g. 9]. Szafir and Mutlu [79] built an embodied agent that monitored peoples’ attention and adapted its behaviour to user engagement. Gaze in HRI has been explored using both virtual agents and physically embodied robots, Admoni et al. [1] categorised social gaze research in HRI into three categories based on goals and methods: a human-centred approach, a designed-centred approach, and a technology-centred approach. They show that HRI as a field focuses primarily on design-focused research, using physical appearance and gaze behaviour to improve interaction with people. Most studies with virtual agents and social robots have used eye gaze as a signal of attention [37], demonstrating engagement [11, 68, 76], and increasing conversational fluidity [17, 53]. Many of the studies on social gaze conducted in HRI, as concluded by Admoni et al. [1], isolate gaze behaviour in a controlled way in order to understand the characteristics of gaze in a particular situation rather than use it in interaction.

Another area of focus in social robotics has been embodied gaze – designing more human-like interactive agents [9, 70], and striving to achieve human-like facial expressions and eye movements. XU et al. [83], for example, investigated the ways that a robot should look at a users’ face in response to a user gaze. Other research has focused on controlling the eye gaze of virtual embodied conversational agents [62]. Andrist et al. [4] presented a conversational gaze aversion robot able to generate and combine head motions to perform mutual gaze. Their study has shown that gaze aversion can be used to demonstrate cognitive efforts, modulate intimacy, and mediate turn-taking. It shows that the direction of gaze plays an important role in shaping conversational participant roles, which can be effective for interacting with virtual agents [7, 10, 58].

That said, several studies in this area suggest that the gaze of a robot is interpreted differently than human gaze [1]. Taking advantage of this difference, as well as the similarities, during conversation could help in creating a more holistic vision of human-robot interaction with gaze.

2.2 Speech Agents

Speech dialogue systems that can perform talk in a relatively natural way with users have been a goal for designers, engineers, and researchers ever since Bell Labs' Homer Dudley developed the Vocoder [21], the first electronic machine that produced human speech in the 1930s. The technical challenges include transcribing audio into speech, extracting meaning and intent from transcribed speech, logically deciding on resultant system actions, producing a response, and producing a synthetic voice coherent with context and use. Many researchers have studied, and continue to study, these individual aspects [60]. Early CSCW research also pioneered studies of speech interaction, drawing on social science studies of talk and interaction, particularly from conversation analysis [15, 24, 27].

Recently, there has been a growth of research that investigates interactions with speech agents. Looking at Smart Speakers use in particular, Sciuto et al. [73] investigate the experience of households with conversational agents. From analysing the logs of 75 Alexa users, they provide detail on how people initially use Alexa, the physical placement of the devices, the daily patterns of conversational usage, and how children interact with them. Porcheron et al. [65] identify the characteristics of interaction with VUIs on mobile devices and how such interactions unfold in multi-party social settings. In further work [66], they examine of the ways in which users of smart speakers practically and interactionally situate the device's talk within their ongoing conversational setting at home, and how users' formulate and direct queries to the device. Moon et al. [54] conducted a between-subjects experiment to investigate the relationship between users' personality and a number of voice technologies for smart home environments, and how it affects users' social responses to these systems. Luger and Sellen [52] highlighted the limited functionality of existing commercially-available voice interfaces and how it causes a gulf between their capabilities and the users' expectations. These studies highlight a number of similar problems in Speech Agent interaction, including; troubles situating the interaction in ongoing conversation, troubles activating and verifying activation, and troubles learning how to formulate and enunciate queries for these devices.

The necessity for human-like communication strategies with systems has been the focus of many studies, as that may help people to understand and to interact with these systems, resulting in more trust and more frequent system use [20, 46]. Therefore, many studies have been conducted to investigate the use of gaze cues in conversational agents [18, 25, 36, 67, 76]. Exploration of the effect of gaze cues in turn-taking in two- and multi-party discourse has been a popular area of research [9, 23, 49, 78, 81].

This work points to the need to understand speech agents in more complex use scenarios, that include multi-party discourse, as well as the opportunities present to counter some interactional challenges with speech agents in use by harnessing other modalities of interaction.

2.3 Gaze in Conversation

Verbal communication is, of course, not isolated from non-verbal communication. Gaze, head pose, and body orientation all play an important role in interaction [80, 82]. In social science, researchers have been exploring gaze interaction since the mid-sixties. Much of the early work on gaze focused on the role of gaze in conversation [5, 44]. During conversation eye gaze can be used to convey information, regulate social intimacy, manage turn-taking. Goffman [28] observed that the direction of eye gaze plays a crucial role in the initiation and maintenance of social encounters. Kendon [44]

conducted a detailed exploration of the function of gaze in face-to-face conversation. He classified looking, or avoiding to look, at the conversational partner as an indication of monitoring, regulating, concentrating, or expressing emotion. Kendon summarised attentive gaze in conversation by saying that people tend to look at the other participant more when listening than when speaking and that the speaker's glances at the other person tend to be shorter than those observed during listening [44]. Duncan et al. [22] highlighted the connection between eye gaze and turn-taking.

Goodwin [32] documented in detail the different aspects of the use of gaze between listeners and speakers. In particular, Goodwin examined the vocal actions of the speaker and the non-vocal actions of the hearer through in-depth analysis of recordings of actual conversations in natural settings [31]. Restarts – points where a sentence fragment is followed by a coherent sentence do not occur by chance but rather are a regular product of the procedures constructing the actual talk. In particular, restarts provides a speaker with the ability to begin a new sentence at the point where the recipient gaze is obtained, or alternatively to request a gaze from the hearer. In the study of the organisation of summons-answer Schegloff [72] proposes that the occurrence of a first item in a sequence, such as a summons establishes the relevance of the next item. Thus, the absence of an answer to a summons might be noted by the repetition of the summons, until an answer is obtained, which then allows the summons to move on to further talk. So, if a recipient fails to gaze at a speaker after an initial restart, that can cause the production of a new restart, which will affect the repeating of the summons.

Since we are designing a system that makes use of gaze to initiate conversation we were specifically interested in the arrangement of gaze between speakers and listeners. Goodwin [31] proposed two 'rules' between the speaker and the hearer in face-to-face conversation. The first is that a speaker should obtain the gaze of his recipient during the course of a turn at talk. Goodwin noted that *"when the speaker has the gaze of the recipient, a coherent sentence is produced. To have the gaze of a recipient thus appears to be preferred over not having such gaze, and this preference appears to be consequential for the talk the speaker produces within the turn"*. In this way gaze is an important cue which indicates that the hearer is listening to the speaker. This is consistent with the possibility that gaze is one means available to recipients for displaying to a speaker whether or not they are acting as hearers. The second rule states that a hearer should be gazing at the speaker when the speaker is gazing at the hearer. The speaker can look away from the recipient, but the recipient should not look away from a speaker who is looking at them. Obtaining the gaze of a recipient within the turn is relevant to the speaker. Yet, even casual inspection of a visual record of conversation quickly reveals that hearers do not gaze continuously at speakers.

Goodwin's work draws heavily upon conversation analysis (CA), a tradition of research familiar to CSCW through the work of authors such as Heath and Luff [35]. Conversation analysis, offers a number of potential resources for the design of speech systems [55, 66]. This influence is not straightforward – Conversation analysis is strongly related to ethnomethodology [26], and holds that talk in interaction is not formalisable or rule-bound, but is locally managed and negotiated in and through production [15]. Clearly, this can conflict with designing for formal rule-bound machines. Accordingly, referring to conversational 'rules' Goodwin did not intend this to mean some sort of forced or mechanical arrangement which speakers and listeners *had* to follow. Indeed, in his writing, he includes examples that deviate from those rules. Rather, Goodwin was looking for (and documenting) systematics in interaction – organisational forms that had some regularity and that could be seen and used by speakers themselves in interaction. Although Gilbert et al. [27] argued that conversation analysis could be employed to improve system interactions, frameworks such as Goodwin's, or the turn-taking systematics [71], do not map directly onto design. This said, as Moore points out, concepts such as recipient design, repair and so on, are clearly relevant to design [55].

3 DESIGNING TAMA

The motivation for building gaze into Tama was four fold. First, research on the use of speech agents in home settings has demonstrated that overlapping speech can be a serious problem. If two users speak at once, a system can have trouble differentiating between which speaker to listen to [66]. Potentially, gaze could be used to differentiate between speakers. Second, we hypothesised that gaze could be more natural than using a wake word – after all, we do not use a wake word when we speak to other people (although we do use names in distinctive ways). Third, we hoped that a gaze system might potentially be more accurate than a wake word activated system. While wake word systems have advanced quickly, as any user will account, they often fail to activate or activate accidentally. Even if gaze was less accurate on its own, in combination it could be potentially more accurate. Fourth, in some settings gaze might be more appropriate than a wake word. In particular, issues of privacy could mean that users would prefer a ‘double lock’ on their speech agent – with it only listening when both gaze and a wake word are detected. These possibilities convinced us that it was worthwhile experimenting with gaze.

In designing Tama, we wanted to explore the idea of how a smart speaker could be brought into a conversation much the same way as another human conversational partner [31, 71], using the offer of mutual gaze and its reciprocation. We designed a smart speaker device, somewhat like the Amazon Alexa or Google Home, but responding not to a wake word (like ‘Hey, Siri’ or ‘Ok, Google’) but rather the gaze of the user on the device. Secondly, we sought to have a device that could respond to being gazed at, and returning its gaze. To provide gaze feedback, we designed a retractable (Figure 1-left) spherical head containing two full-colour LED eyes with 180 degrees of movement laterally, and 60 degrees vertically. This was mounted on a tapered cylindrical body section which housed two OMRON HVC-P2 gaze detection cameras, a 7- microphone array (ReSpeaker v2.0) with 12 full-colour LEDs, a speaker, and a Raspberry Pi v.3B.

This configuration provided three conditions of activation and feedback for us to examine. With the head retracted and the activation trigger set to the voice assistant’s wake-word the device performed as a regular smart speaker, we call this the **wake-word** condition. Changing the activation method to be the detection of gaze, but keeping the head retracted and using the LED light ring for feedback allowed us to isolate the gaze input from output, in a condition we call **gaze-activated** here. Finally, with the head providing gaze feedback and the camera’s providing gaze activation we had the **mutual-gaze** condition. The mutual-gaze condition is our ‘full’ Tama system – but having a system without the head output let us explore if having output actually helped users interact with the system.

The interaction design described here includes the improvements designed to handle multiple simultaneous users, added after the single-user trial described in Section 3.1. We used the same light feedback for the conditions wake-word and gaze-activated for consistency, the default colour scheme found in the API from Google Home or Amazon Alexa. The mutual-gaze condition did not use the led ring, instead it uses the colour of the eyes and its position for feedback. These colours were chosen around keeping the green colour consistently meaning ‘listening’ as in the other conditions, and avoiding using red eyes as this is traditionally associated with anger and aggression. As such, yellow was chosen as the ‘ready’ state (mirroring the yellow of a traffic light), blue to indicate loss of gaze, and pink for an ongoing reply. The user interaction model in all three conditions involved Tama moving between three states; Idle, Listening, and Responding.

Idle: Tama is waiting for interaction. In mutual-gaze condition, Tama looks straight forward with the eyes displaying yellow status. In wake-word and gaze-activated no visible feedback is presented.

Listening: Tama is actively listening to a query. In the mutual-gaze condition, Tama looks towards the detected gaze with the eyes displaying a green status. If the system detects two or more users gazing at Tama it will iterate between the gazes, unless it could use audio to differentiate that one was speaking, in which case it would 'lock' on that user until another started speaking. In wake-word and gaze-activated conditions the status lights show a solid green light.

Responding: Tama is retrieving a response from the server or playing the audio of that response. In mutual-gaze condition, the eyes will be pink and will look towards any detected gaze. In wake-word and gaze-activated conditions the status lights display a rotating green animation found in the default smart speaker software.

3.1 Initial Tests

First, we conducted a laboratory study of single user interaction with Tama, with the hypothesis that mutual gaze interaction would result in the participants having a better impression of the interaction [47, 84] and be drawn away from their ongoing task less [13].

This was a within-participants experiment with counter-balanced conditions. In each condition, the participant sat at a table with a printed list of 14 questions and Tama to their left. The experiment was captured by two video cameras. 12 participants were recruited from the local university by word-of-mouth (Two females and ten males, average age was 25.2). For this experiment, we focused on the error rates in activation, the effect of the articulated head, and the users' usability and flow answers. We defined the error as Tama not being activated by an attempted wake word or gaze, taking more than two seconds to detect the participants looking towards Tama or once activated, not answering the query or the user having to query more than one time. As a result, the average error rates were 17.6% (SD=13.9), 20.1% (SD=25.2), and 12.2% (SD=12.1) for the wake word condition, gaze activation condition, and mutual gaze condition respectively. One factor ANOVA did not show any difference between conditions, suggesting similar performance. The participants filled both System Usability Scale (SUS) [12] and Flow questionnaires [85]. A one-way ANOVA on SUS showed a significant effect for the condition ($F(2, 22)=7.68$, $p=.003$,) on perceived usability in favour of the mutual-gaze condition, as did a one-way ANOVA on the Flow responses ($F(2, 22)=25.2$, $p=.000$) suggesting that participants perceived the gaze conditions to be less intrusive of their ongoing task, in the follow-up interview this was reiterated as eight out of 12 participants mentioned that they felt the repeated use of the wake-word cumbersome in this interaction scenario.

This experiment clearly showed that in a controlled setting, Tama could detect users gaze, and respond to queries from users. As users were on their own, however, there was no (or very little) speech that was not directed toward Tama, making any measure of false-positive activation difficult. This was revealed as a larger problem than anticipated when we conducted an initial test of Tama in a multi-party setting. For this pilot test as with the single-user test we placed Tama on a desk, with users sitting at the other end discussing a travel task together. In the version of Tama we tested, the head moved towards each new gaze that was detected by the camera, triggering the listening state. With two users, this resulted in frequent accidental activation of the assistant that had to be cancelled or waited out by the users. Secondly, if both users were looking at Tama, the head would move between two users, moving between each head quickly, disturbing the user who was talking.

3.2 Improving the Interaction

We addressed the accidental activation problem by moving from Amazon's Alexa to Google's Voice Assistant. This provided API functionality necessary to cancel a query in process, and through tracking the user that initiated the interaction we could then cancel an utterance if their gaze was

not detected in a timeout period. This meant that if gaze was detected but not maintained in any way, then any speech which was detected would be discarded and not sent to our speech system (in this case Google). The result of this was the possibility that some valid queries might be discarded, but it did deal with most accidental triggers ('false positives').

The second problem concerned Tama looking towards one user as the other user was speaking to it. We tackled this by taking advantage of the directional microphone array. When the two users engaged with gaze, Tama would change mutual gaze direction depending on the direction of arrival (DOA) of the voice detected and lock it to the user speaking until both gaze and voice were detected from another direction.

4 SEMI-NATURALISTIC SYSTEM TRIAL

For our main system trial, we employed a within-subjects design, recruiting eleven pairs of participants to perform each of the three conditions. The trials lasted around 45 minutes, with the longest lasting 53 and the shortest 19 minutes. This paper focuses on the analysis of 10 of the pairs of participants – the eleventh pair of participants is not included in this data as they were unable to be recognised by the system in either gaze conditions. We suspect that the 'off the shelf' machine vision chips in our cameras had a skewed training set which led them to be not able to detect gaze in this group. Of the 20 participants, 13 identified as male, 1 owned smart speakers, and a further 12 reported using voice assistants on other devices regularly. All but one of the participants were postgraduate students. Each participant received a \$10 gift card.

4.1 Experimental Setup

We wanted to test the system in a semi-experimental setting that would partially approximate how speech agents are used in settings with multiple people and ongoing, non-system directed speech. Moreover, as our system was designed to integrate into conversation in a different way than the existing 'wake word' model, we were interested in how system use could be incorporated into ongoing talk. Interactions between users and the system, between each other, and between each other about the system were important to study. Accordingly, we designed a task where two participants were to decide and agree on a holiday destination, asking questions about different cities. This task featured talk between the participants, but also queries with the system.

Each of the trials took place with the same setup. The participants were facing each other across a table with the smart speaker equidistant from them, creating a conversational triad. Participants were asked to keep within view of the cameras on Tama. Each trial was recorded by two GoPro Hero 3 cameras, one positioned behind the smart speaker, and one positioned facing the smart speaker – both provided a view of Tama and both participants. The trial was also streamed via webcam to the authors in a separate room, who monitored the trial and started and stopped the conditions on Tama remotely. One author entered the room to explain each condition and set up the task for the users, leaving them alone to continue their discussion and use of the system.

We had one control condition (using the familiar wake word activation method: wake-word), and two test conditions (gaze-activation and mutual-gaze). Each condition started with the participants being asked to perform three scripted training queries each, both to familiarise themselves with the interaction technique being tested and the constraints of successful speech agent query formulation. Each pair started with the wake-word activation condition, followed by our two test conditions, which were balanced in order. Starting with the wake-word condition was intended to provide a baseline of gaze at a traditional smart-speaker in this context of use, which may have been confounded had the participants been informed of, or had experience of, the system being gaze-activated previously. While our analysis here does not follow this initial path, we believe that this had a minimal impact on the reported results. All participants had used a wake-word based speech

	Mean (count)			Median (st dev)		
	Wake-Word	Gaze-Activation	Mutual-Gaze	Wake-Word	Gaze-Activation	Mutual-Gaze
Trial 1	7.70 (17)	24.81 (4)	10.45 (33)	6.59 (0.32)	24.81 (13.13)	8.04 (7.18)
Trial 2	8.18 (9)	7.75 (26)	5.76 (48)	8.37 (1.96)	5.99 (6.11)	5.74 (1.55)
Trial 3	6.76 (15)	8.73 (17)	8.33 (18)	6.5 (1.99)	5.52 (7.38)	6.50 (7.18)
Trial 4	8.92 (26)	20.27 (20)	17.52 (29)	8.50 (2.29)	18.00 (14.97)	11.00 (14.08)
Trial 5	8.12 (21)	9.99 (24)	10.48 (30)	6.76 (4.22)	8.85 (5.18)	6.98 (9.44)
Trial 6	6.59 (32)	20.15 (8)	13.25 (26)	6.19 (2.80)	20.26 (10.90)	8.43 (13.37)
Trial 7	7.67 (26)	14.33 (15)	10.97 (16)	7.95 (2.48)	11.62 (11.05)	7.41 (8.67)
Trial 8	6.61 (21)	10.42 (20)	11.70 (20)	7.00 (1.32)	8.00 (6.24)	7.50 (9.51)
Trial 9	9.46 (17)	N/A	11.37 (16)	8.09 (4.84)	N/A	8.25 (6.36)
Trial 10	8.37 (21)	21.11 (15)	13.12 (28)	7.28 (3.77)	15.29 (14.76)	8.06 (9.87)
All Trials	7.78 (205)	13.15 (149)	10.89 (264)	7.00 (3.14)	9.00 (10.83)	7.04 (9.52)

Table 1. Overview of Interaction Lengths (Seconds)

agent previously (with 13/20 being regular users) and were therefore familiar with this form of interaction, and an analysis of variance of Flesch–Kincaid calculations on the transcribed queries showed that across all three conditions there was no change in the complexity of language used to talk to the voice assistant ($F(2, 632)=2.47, P=0.85$) indicating that query formulation was not significantly influenced by repetition of the discussion task [8, 52]. In one trial (trial 9) there is no data for the gaze-activation condition due to failure of the recording equipment.

4.2 Analysis Methods

The two video angles were combined and coded by the authors in group coding sessions, coding the video for each system activation, its length and various aspects of its performance. Each attempted interaction with the device in each condition was coded with a time taken from the moment the participant started an attempt (from their initial attempt to direct their gaze towards the device before initiating a query, or when they started to say the wake-word). The end of the interaction was determined by either the abandonment of the query attempt (signalled by a return to the ongoing conversation with the other participant or a change in the query), or the time at which the assistant started to play its answer to the query. The number of repetitions necessary to get to this point was counted, along with determinations of whether the query was misspoken by the participant, or if the attempt at achieving gaze lasted more than 2 seconds (which we called a *prolonged pre*-). This resulted in a total of 618 interaction attempts, with 205 in the wake-word, 149 in the gaze-activation condition, and 264 in the mutual-gaze condition. We wanted to take a mixed-method approach and understand what was happening with our gaze system beyond the system statistics. As our earlier designs of the system had been informed by the work of interaction analysis and conversation analysis, we employed this approach in turn in analysing how the system came to be used. So, alongside this analysis of the timings of usage of the system, we extracted individual video clips of each question asked for further analysis.

5 RESULTS

5.1 System Performance

Table 1 shows an overview of the average lengths of interaction in each condition in seconds. In general, we can see that the wake-word condition provided the shortest interactions, with the least variation. The wake-word condition is followed by the mutual-gaze condition with an average of 3.11 seconds (28.5%) longer spent to achieve feedback, with the gaze condition taking on average

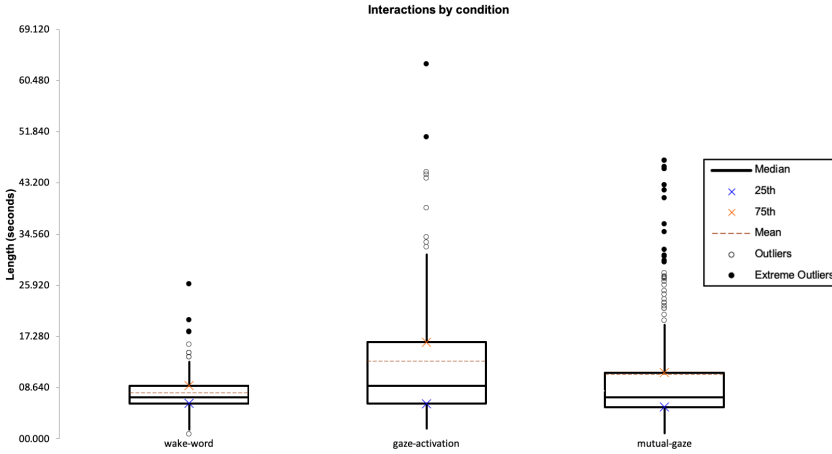


Fig. 2. Distribution of Interaction Lengths

	Mean length			Count (% of total)		
	Wake-Word	Gaze-Activation	Mutual-Gaze	Wake-Word	Gaze-Activation	Mutual-Gaze
Successful Interactions	7.42	6.98	7.54	192 (94%)	76 (51%)	190 (72%)
Single Queries	7.42	7.58	7.64	192 (94%)	85 (57%)	199 (75%)
Repeated Queries	14.3	22.47	23.07	10 (5%)	56 (37%)	57 (21%)
Abandoned Queries	*	*	*	3 (1%)	8 (5%)	8 (3%)
Prolonged-Pre Queries	*	22.32	21.93	0(0%)	26 (17%)	21 (8%)
All	7.78	13.15	10.89	205 (33%)	149 (24%)	264 (43%)

Table 2. Successful and Problematic Interactions

5.34 seconds (40.9%) longer. This means that the condition where queries started by ‘Ok, Google’ (wake-word) was fastest, followed by the condition where users looked at the device and the device looked back (mutual-gaze), while the gaze activated but no mutual gaze condition (gaze-activated) was worst of all. However, looking at the box-plot in Figure 2 and the standard deviation and the per-trial averages in Table 1 we can see that this was not uniform. The longer interactions were epitomised by the occurrence of repeated queries. As can be seen in Table 2, these repetitions were much more prevalent in the gaze conditions. Overall, a Welch’s ANOVA (Levene’s $P=0.000$) showed that the effect of condition on length was significant ($F(2, 254.24)=24.43$, $P=0.000$) as did a Kruskal-Wallis Non-parametric ANOVA ($H=12.156$, $P=0.002$), suggesting that gaze interaction in our test conditions was significantly slower for the users.

The source of the extra time taken by the participants was split between two problems, which will be examined in detail later. The first was difficulty in achieving and maintaining gaze with the system to initiate the interaction, which we label ‘Prolonged Pre’. This draws on the concept of a ‘pre-beginnings’ (shortened to ‘pre-’ in this paper) in conversation, where recipients who want to become next speakers claim a turn before speaking, sometimes through physical rather than spoken means [38]. The second source was repetitions of the actual query, where some problem (usually the response or non-response of Tama in some way) prompts the participant to repeat their initial query.

Table 2 shows the number of queries in each case and the difference in average interaction times. We choose a cut off of two seconds for our prolonged-pre as this is around the time it takes

to speak the Google Home wake-word ‘Ok, Google’ and receive a confirmation beep. This can be seen in comparison to the pre-beginnings seen in smoother interaction sequences, where the user was able to activate and use the system without hesitation and or repetition taking longer than that. Both gaze conditions also had many more repeated queries. The mutual-gaze condition, however, did have a lower repeat rate than gaze-activation, demonstrating that having the feedback of the movable eyes to establish mutual gaze seemed to improve the efficiency of activation and interaction. It is interesting to note that the median interaction time between the wake-word and the mutual-gaze condition are only 0.05 of a second different overall when there are no problems with the interaction. Also, when the repeated queries and those with prolonged pre’s are excluded (as labelled in Table 2 as Successful) there is no significant difference between conditions (Kruskal-Wallis, $H=2.735$, $P=0.255$), which can be taken as an indication that when the interaction went well it was as good as using the wake-word. As shown by both the SUS and Flow results from the single user trial, this form of interaction has the potential to be significantly better in terms of usability and fit with ongoing tasks without sacrificing speed of interaction.

Troubles in starting the interaction were infrequent in the wake-word condition, even those cases where they forgot to say the wake-word were repaired by the participants in under 2 seconds. In the mutual-gaze condition, 8% (21) of the queries suffered from ‘prolonged-pre’. This increased in the gaze activation condition, to the point where 17% (26) of the interactions suffered from ‘prolonged-pre’ as the participants struggled to either activate the system, or recognise that it has been activated and begin to voice their query.

There were situations when the participants successfully initiated the agent and began querying the system yet still had a number of repeats. We can see from Table 2 that these were evenly split between the two gaze conditions in both number and the length of time the participants took to eventually get an answer from the system. In these situations the main reason observed for this repetition was the interaction designed to prevent the assistant responding to glances at the devices without the intent to trigger a response. This meant that if the participants looked away for more than the timeout (average 3 seconds) while the query was being voiced the query never received an answer and, if the participants chose to persevere with this same question, their next attempt counted as a repeat. While this will be examined in detail in later sections, these design decisions resulted in 167 cancels by the system during the trials, reducing the number of false activation to 7. This met our goal of limiting false activation, but at the expense of cancelling potentially valid queries made by users forcing them to repeat.

5.2 Video Analysis of Tama in Use

While Tama is not quite as fast or reliable as a wake-word based system, as a new type of speech agent these results are at least promising. So as to improve this system further, we used the video recordings to examine cases where Tama had problems in its interactions with users. Drawing on interaction analysis, and informed broadly by the conversation analysis literature, we examined the recorded queries to look for revealing cases of interaction. From recordings of Tama’s use, we extracted a corpus of 175 videos, distributed across the ten different trials. In focused analysis sessions, we watched all the video clips of each trial twice, to gain an overview. Informed from our developing analysis, and the quantitative analysis, we extracted a set of 28 video clips for more in-depth analysis. Each of these video clips was selected for illustrating either smooth use of the system, different problems that were emerging from our analysis, or other behaviours that we felt were interesting. These were analysed in more focused data analysis sessions. Our goal in these sessions was to explore the individual differences in the use of the system, both for individual users and trials, but also how the system worked or failed overall. We also sought to understand the

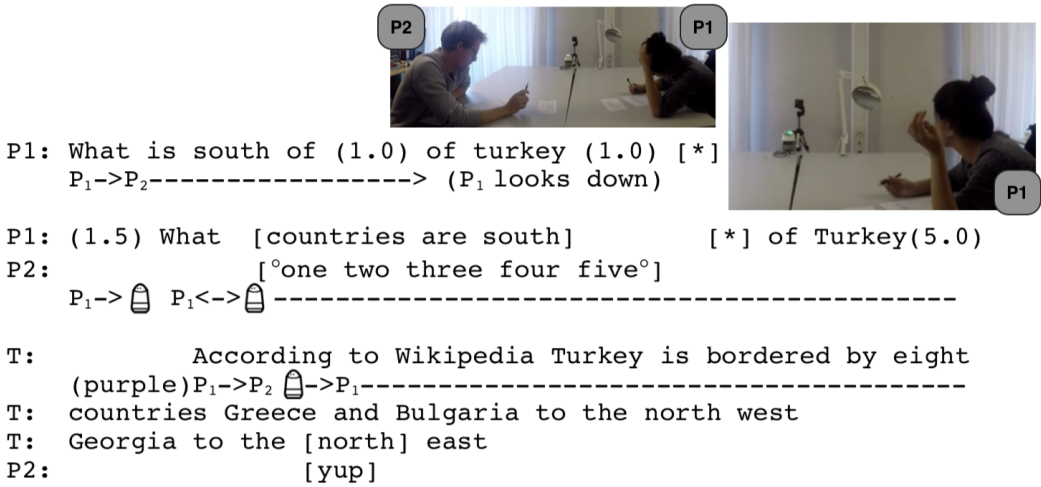


Fig. 3. A relatively smooth query with Tama

different behaviours and uses of the system overall, as well as how different problems emerged in use.

We will look at five examples of Tama being used, four from the mutual-gaze condition, and one from the gaze-activated condition. Our first example (figure 3) is of a relatively smooth usage of Tama, our second is of an extended ‘pre-’ sequence, where a user has to work to get Tama’s attention, while our third and fourth examples are of problems with the use of gaze. Lastly, we look at an example of a user ‘jumping the gun’ and repeating a question before Tama has answered.

5.2.1 ‘What countries are south of turkey?’ Figure 3 gives an example of a relatively unproblematic interaction with Tama. We have added a ‘glance track’ under the transcript of what was spoken. As gaze is one of the most important aspects of Tama it is important to be able to see it in time with the talk and Tama’s replies. The arrows indicate gaze – such as $P_1 \rightarrow P_2$ for participant one looking at participant two, with mutual gaze indicated by a double-headed arrow, e.g. $P_1 \leftarrow \rightarrow P_2$. An extended gaze is shown by an extended line until the point in the transcript when the gaze was broken. Tama is indicated by the small Tama icon. For the transcripts themselves, we have made use of a limited form of Jeffersonian transcription [34, appendix a][63]. The numbers in (brackets) indicate pauses, and we have broken these out into multiple (items) where the gaze changes (indicated on the gaze track). The transcripts also show overlapping talk by the use of square brackets [around the speech that overlaps]. We have also added photographs of the interactions, with their place in the sequence indicated by an * in the transcript.

Returning to figure 3 you can see that the video starts with participant 1 looking at her partner and saying “what is south of turkey”. She then looks down at the assignment card (which does not mention Turkey), and then turns to look at Tama. After 1.5 seconds, Tama moves its head and establishes mutual gaze. She then asks her question of Tama: “What countries are south of Turkey”. She uses a slightly different wording when she asks Tama, compared to when she is talking to her partner, as she adds the word “countries”. As she is asking the query, her partner counts (whispering) the neighbouring countries that they have already written down. She keeps looking at Tama during the five seconds when they are waiting for it to answer, and five seconds after she finished asking her question the lights turn pink, and it gives its answer to the question. As Tama



P1: [\downarrow](2.2) °Look at me°(0.9)
 P₁-> ->P₂ P₁<->

P1: Wh. What sound does a lion make (2.5)
 P₁<-> ----->P₂<->

P2: Sorryff=
 T: * =This is a lion
 P₁-> P₂->P₁ ->P₂----- (tama looks between P₁ and P₂)



Fig. 4. Extended pre- to gain attention

answers, she turns to look at her partner, but Tama keeps looking at her. It is notable how there is less than a one-second break between talking to her partner and talking to Tama, and how after the answer they quickly move back to looking at each other, with the second participant saying “yup” as Tama lists the countries. Notably, as Tama speaks, the participants look away and look at each other, not needing to maintain mutual gaze with Tama after their query is answered. This is a clear example of the smooth use of Tama that was hinted at from our single-user test — a query asked and answered without the need for a wake word.

5.2.2 Problems with Pre-. Unfortunately, using Tama was often not quite as smooth. One of the issues that we had identified while coding the videos, and that we counted in Table 2 above, were cases where there was some sort of ‘prolonged pre-’, where users struggled to get the attention of Tama. In a prolonged pre, a user needs to make extra efforts over two seconds to initiate the query with Tama. In the extract above after looking at Tama it takes 1.5 seconds to respond. But in many cases, Tama would take longer, or require extra actions, before it would respond and the participant would start their query. Indeed, 11 % of queries for Tama featured a ‘prolonged pre-’ of over 2 seconds. But why did these prolonged pre- sequences happen?

In Figure 4 P1 is starring at Tama for 2.2 seconds without Tama breaking its gaze from the other partner P2. She (P1) quietly says ‘look at me’ and this causes Tama to turn and look at her — with mutual gaze finally established as she says ‘me’, after which she waits 0.9 seconds before then speaking her query (with a slight dis-fluency at the beginning - the “Wh.”). Here Tama starts by being ‘locked’ on the wrong conversational partner, who is not speaking, and appears to fail to notice the attempt by participant 1 to gain Tama’s attention. We had not told participants that making noise would cause Tama to change whom it looked at, and so sometimes they would accidentally discover (as in this case) that making a noise would cause Tama to turn to them. In some other cases the cameras on Tama would be responsive to one participant over another, and so would seem to be locked to one participant despite the efforts of the other to respond to them. Participants would make exaggerated gestures with their head to try and ‘trigger’ Tama (somewhat like the speech patterns users use with speech agents). After establishing mutual eye contact, the participant asks her question, and as she does she glances down. Tama remains looking at her, however, and it is not until she is finished her question that it quickly glances at the other

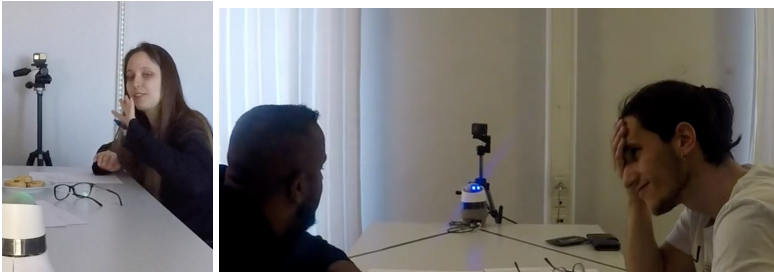


Fig. 5. Waving or hiding face to manage Tama's attention


participant before moving back to look at her. This prompts the other participant to say sorry and start to laugh, just before Tama answers the question.



As we discussed above in the background, Goodwin's analysis of gaze in talk gives a number of examples where speakers start their turn in talk before gaze is obtained. Talk can then either display disfluencies, such as multiple restarts, or can restart proper when mutual gaze is obtained. In this way, talk acts as a sort of 'pre-' attempting to gain the attention of the other speaker before proper communication commences. In this case here, the participant does gain mutual gaze, through saying 'look at me', and only when she has established mutual gaze does she start her query properly. Indeed, she starts her query with "wh" pausing slightly then speaking her query. These are known as false starts, and are relatively common at the start of new turns where there has been a problem with gaining the attention of a speaker [16]. While the participant does then successfully ask her query and get an answer, Tama actually turns to look at the other participant just after she has finished speaking her query. This leads the second participant to say "Sorry" and start to laugh just as Tama replies with the "This is a lion", with Tama looking back and forth between participants at this point.

The prolonged-pre here takes just over 3 seconds, before the participant and Tama achieve mutual gaze and she starts her query. It is important to note that in this case the problem here was not one of inaccuracy in the recognition of gaze. With the participants looking at Tama (photo top left, figure 4) it is ambiguous which participant Tama should look at. This ambiguity is solved when participant 1 speaks, and Tama then turns to look at (and listen to) her.

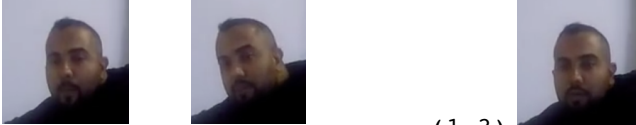
The apology from participant 2 here is interesting as an example of a participant 'explicitly' talking about, or manipulating their gaze with the system — to the point of accounting for accidentally attracting Tama's gaze. While we had designed Tama as a 'natural' speech agent, the problems with its use meant that at some times participants *explicitly* discussed or engaged with their gaze in using the system. In another example, a participant moved their hand over their face to block their gaze from Tama so it would look at the other participant. Clearly, it seems that our trial participants felt some sort of joint responsibility in using Tama, or at least in not disturbing another users' query. Both the 'look at me', and the 'sorry' in this extract are cases where the management of gaze became an explicit issue for the participants, something that had to be managed as an additional aspect of using the system.



Similarly, in some cases a participant would struggle to establish mutual gaze at the beginning of a query, and would have to resort to restarting queries, repeating a query, or even more explicit efforts like waving or moving their head (figure 5). Doing any of these would slow down the query overall, leading to a longer mean query time.



P1: (1.2)(0.5) What is the national dish of Bulgaria
 P₁->  P₁<--->  -----*

P1: (1.3) (1.0) (1.3) What is the national dish of Bulgaria
 -----(BLUE)-----



P1: (0.4) (1.3) (1.3)
 ----> P₁->  *((P₁ tilts head* then back*) P₁<--->  -----

P1: What is the national dish of Bulgaria? (4.6)


T: On the website kashtical dash tourist dot com they say
 ----->  -> P₁

Fig. 6. Speaker loses mutual gaze during query

5.2.3 Problems with Gaze. In successful cases, the participants used gaze and speech to maintain Tama's 'attention', and they continue talking until they have completed their query, after which Tama then fetches a response from the voice agent service. However, in some cases participants had problems with Tama not responding to a query forcing a participant to repeat or even abandon their query.

One cause of problems would be if Tama lost mutual gaze long enough to cause it to abandon the query and stop listening. Due to system failures of the gaze cameras, for some participants this could happen multiple times — even if the participant was actually maintaining their gaze with the system. When Tama lost gaze, the LEDs would flash blue to indicate that it was about to cancel a query. Then, around two seconds later, if it still detected no gaze it would abandon the query and move its gaze away.

In figure 6 we can see a participant ask about the national dish of Bulgaria, having to ask the same question three times before getting Tama to reply. Shortly after the participant asks the question the first time, Tama stops detecting the participants gaze — presumably because of a problem with the gaze cameras. It indicates this with a short blue flash. This prompts the participant to repeat the question. Despite this second attempt, Tama still does not seem to have detected the participants gaze, and it actually looks away and stops listening at the end of the second query.

After repeating their question, and losing Tama's gaze, the participant here moves their head back and forth, a motion which seems to trigger Tama to regain mutual gaze finally. After briefly waiting, he then repeats his query for the third time, and then he finally gets his answer. Goodwin discusses the loss of mutual gaze between speakers he gives examples of cases of participants

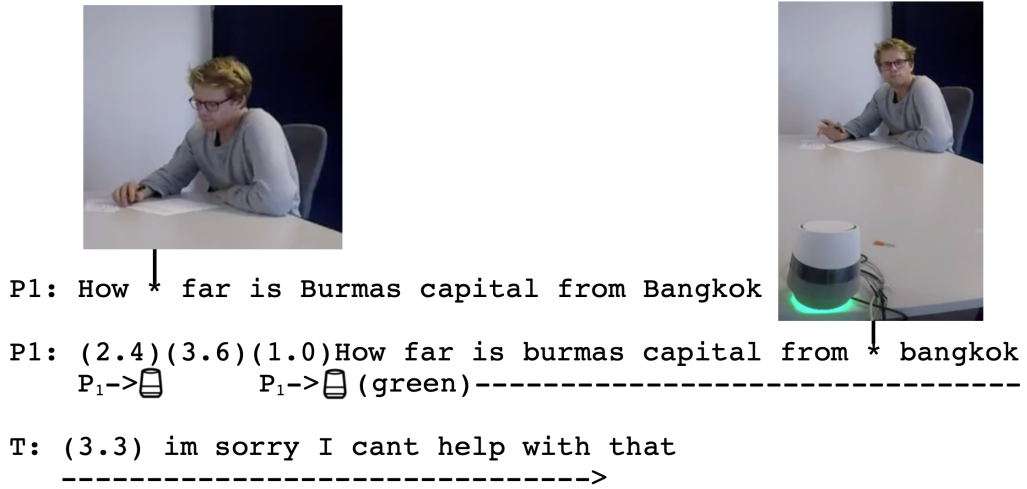


Fig. 7. Speaker does not establish mutual gaze during query (gaze activation)






moving their body or gesturing so as to ‘regain the attention’ of the listener, and this seems to be the case here.

While this was not a hugely successful use of Tama (17.4 seconds for one query) Tama does at least give some feedback that it has not successfully heard a query. This is perhaps what leads to the rapid query repeats — with a wait of 2.3 seconds after the blue flash (indicated as “BLUE”) in Figure 6), and 2.6 seconds after losing Tama’s gaze the second time, before the user retries. After the successful query he waits 4.6 seconds, suggesting more confidence in the query — and he, in fact, does get an answer in the end.

While this is an example of Tama failing to detect gaze, participants also had some problems themselves with maintaining gaze throughout their query. Users’ ‘loss of gaze’ involved users not establish or maintain gaze with Tama as they spoke. In particular, when participants looked down to their notes, or to the card describing the test, they would break eye contact with Tama. If they did this for long enough Tama would assume that this was not a query and would stop listening. This lead to episodes where participants would stop speaking, and an attempt to re-establish eye contact before continuing with their query, or participants repeating their query as they heard no response from Tama.

In figure 7, the participant starts his question without actually looking at Tama, and so Tama does not activate. In fact, he actually looks at Tama right as he finishes his turn — something common in talk as we can ‘select the next speaker’ [71] by looking at them. Unfortunately, Tama’s interaction design means that it only starts listening — triggered by gaze — exactly as he stops speaking. After waiting 2.4 seconds he seems to notice his mistake, looks down then back up again and after 1 second looking at Tama, the system lights go green and he starts his query again. Unfortunately, while the system responds, in this case the speech agent itself seems unable to answer the question.

This extract is from the ‘gaze activation’ condition for our test of Tama — so the system did not look back at the user, but instead flashed different colours to confirm that it had ‘seen’ the user. As the quantitative results show, clearly this had some implications for users successfully getting Tama to respond to their queries. In the gaze activation condition data, we observed that users would fail to establish their gaze or drop their gaze while talking to Tama, (as in figure 7). It appears

P1: (0.2) ww hh. how far away. i.s. mars *(3.0)
 P₁<->  ->P₁ P₁<-> -----> ->P₁ P₁<-> ---->


P1: How f[ar]
 T: [fift]y six million kilometres
 P₁<-> ---->



Fig. 8. Speaker repeats query while Tama answers

that the lack of appropriate feedback from Tama in the ‘gaze activation’ condition caused users to themselves make less use of gaze when interacting with the system, which led to higher error rates with this condition compared to the ‘mutual gaze’ condition.

One challenge we faced in designing Tama was that it requires (in order to distinguish incidental glances from attempts at interaction) some sort of maintenance of gaze to be triggered to completion. Goodwin actually points out that *speakers* need not look at listeners throughout their utterance – it is usual for them to look away after they have established mutual gaze at the beginning of a turn. While listeners usually should be expected to hold mutual gaze with a speaker, this is only *if* that speaker is looking at them. Yet as we had designed Tama it required gaze at least once every 3 seconds regardless of the conversational context. In this way, Tama’s design actually conflicted with Goodwin’s description of how gaze works amongst humans. Of course, we were not designing a system that had anything like the sophistication of a human – and so perhaps these sort of issues are to be expected, but clearly designing gaze into a system is more than simply ‘simulating’ what a human would do in this setting, and perhaps following this could at times be misleading (a point we return to in our discussion). In the mutual gaze condition this was less problematic – although our instructions described Tama as a ‘gaze activated’ system we did not tell them they needed to look at Tama throughout their query, but users on the whole successfully maintained eye gaze, or learned very quickly that this was how to make Tama work.

5.3 Jumping the Gun

After the speaker has finished their query they need to wait for a small period of time before getting an answer from Tama. This was usually around 1 to 2 seconds, but it could be longer due to network and processing delays. At this point, in most cases, Tama speaks an answer to them. Yet sometimes Tama has not heard them correctly, and they need to repeat their query. How long should a speaker then wait before asking the question again? If the participant started to voice the query again while Tama was processing Tama could answer while they were speaking, while waiting too long if their query had not been detected would result in them having to initiate the interaction from the beginning.

It seems like participants were good at judging how long to wait – in only ten cases (2%) did their repeated query overlap with Tama’s response (figure 8). In this case, the participant ‘jumps the gun’ and tries to repeat their query while Tama is still processing. As he is speaking the query, this participant looks down to the card which describes the task. This question is one of the three ‘sample questions’ that we had provided, and the participant seems to look down on the card and read the question as they are speaking it. As he gets to ‘is mars’ he looks down and then back up again as he finishes the sentence. However, these pauses were all quite short and Tama keeps looking at the participant, even though they are not looking at it. In the end, Tama does answer correctly, although overlapping what it says with the participant.

This example suggests that some sort of feedback about when Tama had received a query and was ‘thinking’ (specifically, waiting for feedback from Google). This would have indicated to users that they did not need to repeat their query and that the system was about to answer. This sort of ‘received’ behaviour is noticeable in conversation with people by ‘holding the turn’ behaviour, such as ‘err’ or other non-verbal-noises, or gestures [16].

Lastly, when Tama actually answers, our participants listen to the answer, although in some cases it can become apparent that an incorrect answer is being given (such as figure 7). Sometimes this was not due to a problem with Tama, but either an ill phrased question by the participant, or a limitation in the Google Assistant service. In about 26 percent of Tama questions an answer was given that we considered not to follow the intent of the listener, with about 2 percent being the speaker mis-phrasing their question, and about 24 percent an issue with Google, or Tama in understanding their question.

In some cases, the participants turn their attention away from Tama when they realised that the wrong answer has been given, turning their attention to each other and discussing the error almost immediately after it is detected. Interestingly, this might seem as a discourtesy to a human speaker, but this does not seem to affect participants when it is a speech agent they are dealing with. Again, this is suggestive of the point that while participants were successful in using gaze to interact with Tama to some extent, they did not treat it exactly the same as a human conversationalist.

6 DISCUSSION

While our results are focused specifically on using gaze as part of a speech agent, we can also draw lessons more broadly for designing with gaze as part of an interactive system. In our discussion, we first focus on specific design lessons for how gaze might be used in system design. Second, we discuss the relationship between human-human interaction and human-system interaction in design. Lastly, we discuss more broadly how gaze might be adopted in interactive systems such as social robots.

As an overall reflection, our semi-experimental setting was in some ways more challenging than the familiar home environments that speech agents are generally used in. First, the system had to deal with multiple users during most of its use; for some of the queries users read from the instructions we had given them; and the queries were focused on a specific task (planning a holiday) rather than the usual mix of home automation and media control that home speech agents are used for [2]. However, in home environments there would also likely be more overlapping speech, less clearly spoken commands, not to mention questions about lighting and if gaze was actually detectable [19]. As a non-naturalistic test, this experiment focused on the design issues around gaze, when and how gaze interaction with speech agents might actually fit within the complex and shifting contexts of real-world use is an interesting avenue of research for future work.

6.1 Design Lessons

Our experiments and design iterations with Tama show that as a method of interacting with a speech agent gaze has promise, and that at a basic level, conversational interfaces can be successfully augmented with gaze. This initial implementation allowed us to surface the interesting cases in which users adapted to, and made use of, the mutual gaze that Tama could support. Clearly, giving feedback to users through the movable eyes improved the accuracy of the system. In part, this could be down to the head encouraging users to maintain gaze with the system when they spoke their query prevented accidental cancellation. This also resulted in a lower rate of prolonged pre-sequences, which could point to users being more quickly or easily interpret the mutual-gaze as feedback that Tama was listening compared to the bottom light ring. Prolonged pre-sequences did still slow the system down somewhat – partly because Tama could ‘lock’ onto the wrong user.

When Tama did not gaze back at the user who intended to start an interaction, or Tama looked at the other participant instead, they felt obliged to negotiate, or to make some noise or gestures, until they managed to secure the desired mutual gaze, even though this was entirely unnecessary.

While our video data has not been coded to completely measure the accuracy of the system with respect to it discounting glances towards the system that were not intended to result in a query (we have the false positives, but no indication of the false negatives), on the surface the head seemed to improve this somewhat. One interesting supposition around this could be that, following human-human conventions [e.g. 4, 45] on gaze-aversion, our participants were more surreptitiously in their incidental glances towards the device when it had the possibility of making eye-contact and this behaviour made it more difficult for the cameras to detect these incidental gazes. This all points to the issues that the complexities of human communicative behaviour can bring when harnessing them for system interaction. These are nuanced, contextually and socially dependent behaviours that require careful abstraction – and feedback that exposes the nature of that abstraction. This surfaces a design opportunity where mutual gaze from Tama can determine who the speaker is, and provide instruction through the withholding or providing mutual gaze as to whether the speaker should proceed with their interaction.

One challenge in integrating gaze into systems is the question of how to deal with multiple users [69, 77]. We had specifically designed Tama so that it could support multiple users together, with a relatively flexible shift in interaction for users between talking to each other, and to the system. Gaze then offers the potential for dealing with a system that can dynamically switch between users without the need for explicit commands. This could also be combined with face or speaker recognition to support systems that can dynamically shift between user contexts, providing personalised information or even withholding personalised information when others are detected.

In our experiments, we explored replacing one part of speech interaction (a wake word) with gaze. This suggests that gaze can be used as an effective way of initiating interaction with a system. Indeed, there is an analogy here with recent smartphone designs, which also make use of gaze detection to start (and authenticate the user) [6, 33]. Gaze at different parts of a conversation has different meanings (at least in human-to-human interaction). Drawing on Goodwin, we also discussed how gaze differs for speakers and listeners, who are of course not fixed as a conversation unfolds. However, moving into other parts of the conversation will require unpacking the speech agent to allow an ongoing understanding of the conversational context to be exposed – most current instances separate the transcription from the action, only processing for context after the utterance is complete. With this finer control over the interaction, however, there are a number of possible behaviours that could be included. One would be to control for query interruption by detecting a change in speaker and then discounting ongoing speech for a short period of time until gaze is re-established and a restart that can be contextually matched with the interrupted query is detected. We have unpacked such restart behaviour in the results section above, and leveraging this for interaction would provide a measure of robustness to the interactions.

Lastly, while we have designed and discussed Tama as a speech agent, rather than as a social robot, it clearly has some features that are similar to those deployed in social robotics. One interesting observation from our results that has resonance with the design of socially oriented interactions is the way in which our participants accounted in conversation for the actions of Tama. In our design, Tama was purely reactive meaning that the actions of the system were fairly easily (if not always accurately) visibly accountable to our participants in relation to their own bodily and conversational interactions with the system. This was used especially around more problematic interactions, easing any interpersonal component to the troubled interaction with the system and, to some extent, signalling collaboration and support in the shared challenge of the interaction. In the design of social robots, such accountability may become more difficult, at the same time it could

be posited that as the complexity and the sociality of the interaction with the system increases such accountability will be more valuable when the interaction runs into trouble. While surfacing the internal state of the system for all actions would be interactionally burdensome, we suggest that where such a social robot changes interaction partner, topic, or other states important to the ongoing social context that some accountability cues be provided to ease the human interactions happening alongside the system ones. Another lesson we take for the design of robots is the importance of understanding and designing for conversation in robot interaction. Drawing on conversation analysis gave us particular resources for understanding how gaze unfolds, and when and how Tama could behave and make use of gaze. Understanding the different phases of conversation — and the different roles that a system and users might play have been useful in this case. This said, we do not mean to suggest that only with a holistic understanding of the context and intent of the user can such interaction be successful. Designing systems that can act in conversation will depend upon a subtle understanding of not only talk in interaction as it happens, but also the co-development of systems that while they will always likely have a partial and faulty understanding and behaviour, can at least work in tandem with humans to a workable extent.

6.2 Minimally Anthropomorphic Design

A more general point concerns how we went about our design of Tama in the first place. One initial inspiration for designing the gaze interaction with Tama was to take advantage of how individuals manage conversation. While Tama has some limited robotic features, it is clearly not an anthropomorphic robot like Pepper¹. The concept of anthropomorphism is often talked of as a complex combination of physical design, interaction design, and users simplifying and projecting human-like intents to make sense of the system behaviour and mould their interaction with it [42, 61]. Here our goal was to perform that simplification and projection in design, producing an interaction that was not human-like in its behaviour, but would take advantage of learned skills in human interactions. We sought a system that would use, and enable the use of, gaze in a similar way to a human co-conversationalist, yet while clearly not looking like a person or performing the range of nuance and meaning that a human actor is capable of. In some ways, this echos Levillain and Zibetti's *behavioural objects* and the move away from humanoid or zoomorphic design [51] while performing human-readable behaviours. Yet in our case, the focus is squarely on functionality. We did not aim to give Tama behaviours that made it seem human, or alive, but to limit the interaction and expression to a *minimally anthropomorphic design* in order to reduce the expectations of nuance and complexity such interaction may engender in users. Our expectation was that people would use their interactions with others as a resource for understanding and shaping their behaviour with Tama. Clearly, as the qualitative analysis above shows, this was at least partially the case.

In some ways, we are echoing many of the early CSCW debates around speech systems, and their relationship to conversation analysis, and it is fascinating to see this resurface with more contemporary technology [14, 15]. One early CSCW debate around speech systems was that in attempting to follow 'rules' of conversation, any speech system would necessarily follow a much simplified model of how human to human interaction works. The 'rules' that we have discussed from Goodwin, for example, are not some sort of hard and fast instruction for action. As with rules more broadly in conversation analysis, they are better thought of not as production rules, but rather as templates that can be used to understand the action of others, and in turn how others would see the actions of oneself. It is not that listeners need to always manage their gaze in correspondence with Goodwin 'rules' but rather that actions will be seen *in light of* these interactional structures.

¹<https://www.softbankrobotics.com/us/pepper>

If we look at our data we can see that when Tama behaves in ways that have a similarity to human interaction, it can be responded to and understood by its users. But also that when things go wrong users can dynamically rethink how they are using (in our case) gaze. This is a wonderful example of the flexibility of interaction. When Tama goes wrong, it is not of particularly interesting in an of itself that the use of the system ‘broke down’, but rather that users adopted different strategies to try and get their queries answered (such as waving their hands at Tama, or hiding their face). We take this as evidence that drawing on conversation analysis – and the idea of the system in interaction more broadly – can be productive for design. Yet we must always be aware that users themselves are dynamic and flexible, ready to abandon and try different approaches if need be.

We must also understand and expect that many of the conventions of human conversation examined in the conversation analysis literature are not desirable in the design of most interactive systems. In the same way that the conceptual paradigm of ‘direct manipulation’ settled, in most interfaces, a considerable distance from tangible [39, 57, 75], most interactive systems would not benefit from the considerable burdens that the complexities of human communication encompass. Interactions with a virtual agent need not be managed to ensure future civility [3], nor be interspersed with rituals to manage face [29], or carefully disengaged with an so as not to cause offence [30]. This echoes Moore & Arar [56] in their selection of patterns of interaction from Conversation Analysis which can be readily applied to conversational user interfaces, which they see as something which future research in both CA and HCI can add to. The understandings of human interaction drawn from the social sciences should be seen as a pallet for design in HCI, rather than a template or goal.

6.3 Future Designs

In future work we plan to continue the development of Tama along two paths. The first is to continue to experiment with interaction for speech agents, working to increase the accuracy and interaction in the response to what we have found out in the trials reported on in this paper. This initially will involve experimenting with both more complex models of gaze interaction to better fit with the conversational agent’s place in social context, as well as feedback methods to allow users to better understand the state and interaction with both gaze and voice. The second is to develop Tama as a platform to better understand human gaze as a sociological, communicative function of human interaction. By providing the tools to script, record, and adapt interactions with gaze we hope to enable a deeper understanding of just how and why groups of people use gaze in certain ways.

7 CONCLUSION

In this paper, we present Tama, a gaze-enabled voice assistant, and exploratory trials in both single and multiple user scenarios. We used the video recordings of 10 trials to analyse users interaction with Tama in three conditions, showing that gaze interaction resulted in a better user experience when it included gaze output as well as input. Our study suggests that gaze can be used to augment, or even replace, the wake-work in initiating interaction with speech agents. We examined in detail exactly how gaze interaction can benefit of ongoing co-present conversation, for example providing a clear signal of intent and intention while interacting with the system, and situations where it caused problems, such as users pausing and restarting their queries in response to eye-contact as they would when in conversation with a human to the detriment of their interaction with the system.

We begin here to unpack the complexities of designing with gaze for multiparty conversationally situated interaction. As voice interfaces and speech agents continue to be deployed into more varied contexts, we hope that this work will inspire and guide future research in this area.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI grant number 18H06473, Oki Electric Industry Co., Ltd., Vetenskapsrådet grant 2016-03843, and the Swedish Foundation for Strategic Research project RIT15-0046.

REFERENCES

- [1] Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. *J. Hum.-Robot Interact.* 6, 1 (May 2017), 25–63. <https://doi.org/10.5898/JHRI.6.1.Admoni>
- [2] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3 (April 2019), 17:1–17:28. <https://doi.org/10.1145/3311956>
- [3] Lynne M. Andersson and Christine M. Pearson. 1999. Tit for Tat? The Spiraling Effect of Incivility in the Workplace. *Academy of Management Review* 24, 3 (July 1999), 452–471. <https://doi.org/10.5465/amr.1999.2202131>
- [4] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational Gaze Aversion for Humanlike Robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI '14)*. ACM, New York, NY, USA, 25–32. <https://doi.org/10.1145/2559636.2559666>
- [5] Michael Argyle, Luc Lefebvre, and Mark Cook. 1974. The Meaning of Five Patterns of Gaze. *European journal of social psychology* 4, 2 (1974), 125–136. <https://doi.org/10.1002/ejsp.2420040202>
- [6] Hind Baqeel and Saqib Saeed. 2019. Face Detection Authentication on Smartphones: End Users Usability Assessment Experiences. In *2019 International Conference on Computer and Information Sciences (ICCIS)*. 1–6. <https://doi.org/10.1109/ICCISci.2019.8716452>
- [7] Janet Beavin Bavelas, Linda Coates, and Trudy Johnson. 2002. Listener Responses as a Collaborative Process: The Role of Gaze. *Journal of Communication* (2002), 15. <https://doi.org/10.1111/j.1460-2466.2002.tb02562.x>
- [8] 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>
- [9] Maren Bennewitz, Felix Faber, Dominik Joho, Michael Schreiber, and Sven Behnke. 2005. Integrating Vision and Speech for Conversations with Multiple Persons. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2523–2528. <https://doi.org/10.1109/IROS.2005.1545158>
- [10] Dan Bohus and Eric Horvitz. 2010. Facilitating Multiparty Dialog with Gaze, Gesture, and Speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10)*. ACM, New York, NY, USA, 5:1–5:8. <https://doi.org/10.1145/1891903.1891910>
- [11] Cynthia Breazeal. 2005. Socially Intelligent Robots. *Interactions* 12, 2 (March 2005), 19–22. <https://doi.org/10.1145/1052438.1052455>
- [12] John Brook. 1996. SUS-A Quick and Dirty Usability Scale. In *Usability Evaluation In Industry*, Patrick W. Jordan, B. Thomas, Ian Lyall McClelland, and Bernard Weerdmeester (Eds.). CRC Press, 189–194.
- [13] Allison Bruce, Illah Nourbakhsh, and Reid Simmons. 2002. The Role of Expressiveness and Attention in Human-Robot Interaction. In *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, Vol. 4. 4138–4142 vol.4. <https://doi.org/10.1109/ROBOT.2002.1014396>
- [14] Graham Button. 1990. Going Up a Blind Alley: Conflating Conversation Analysis and Computational Modelling. In *Computers and Conversation*, Paul Luff, Nigel Gilbert, and David Frohlich (Eds.). Academic Press, London, 67–90. <https://doi.org/10.1016/B978-0-08-050264-9.50009-9>
- [15] Graham Button and John R. E. Lee. 1987. *Talk and Social Organisation*. Multilingual Matters.
- [16] Donald Carroll. 2004. Restarts in Novice Turn Beginnings: Disfluencies or Interactional Achievements. *Second language conversations* (2004), 201–220. <https://doi.org/10.5040/9781474212335.0014>
- [17] Justine Cassell, Timothy W. Bickmore, Mark N. Billinghurst, Lee W. Campbell, K. Chang, Snorri Hjörvar Vilhjálmsson, and Hao Yan. 1999. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, New York, NY, USA, 520–527. <https://doi.org/10.1145/302979.303150>
- [18] Justine Cassell, Catherine Pelachaud, Norman Badler, Mark Steedman, Brett Achorn, Tripp Becket, Brett Douville, Scott Prevost, and Matthew Stone. 1994. Animated Conversation: Rule-Based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents. In *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '94)*. ACM, New York, NY, USA, 413–420. <https://doi.org/10.1145/281831.281838>

[//doi.org/10.1145/192161.192272](https://doi.org/10.1145/192161.192272)

- [19] Eunji Chong, Katha Chanda, Zhefan Ye, Audrey Southerland, Nataniel Ruiz, Rebecca M. Jones, Agata Rozga, and James M. Rehg. 2019. Detecting Gaze Towards Eyes in Natural Social Interactions and Its Use in Child Assessment. *arXiv:1902.00607 [cs]* (Feb. 2019). <https://doi.org/10.1145/3131902> arXiv:cs/1902.00607
- [20] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. 2014. SimSensei Kiosk: A Virtual Human Interviewer for Healthcare Decision Support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1061–1068.
- [21] H. Dudley. 1940. The Vocoder—Electrical Re-Creation of Speech. *Journal of the Society of Motion Picture Engineers* 34, 3 (March 1940), 272–278. <https://doi.org/10.5594/J10096>
- [22] Starkey Duncan. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology* 23, 2 (1972), 283–292. <https://doi.org/10.1037/h0033031>
- [23] Mary Ellen Foster, Andre Gaschler, Manuel Giuliani, Amy Isard, Maria Pateraki, and Ronald P.A. Petrick. 2012. Two People Walk into a Bar: Dynamic Multi-Party Social Interaction with a Robot Agent. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction (ICMI '12)*. ACM, New York, NY, USA, 3–10. <https://doi.org/10.1145/2388676.2388680>
- [24] D Frohlich and P Luff. 1990. Applying the Technology of Conversation to the Technology for Conversations. In *Computers and Conversation*, P. Luff, G. N. Gilbert, and D. Frohlich (Eds.). Academic Press, London.
- [25] Maia Garau, Mel Slater, Simon Bee, and Martina Angela Sasse. 2001. The Impact of Eye Gaze on Communication Using Humanoid Avatars. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '01)*. ACM, New York, NY, USA, 309–316. <https://doi.org/10.1145/365024.365121>
- [26] Harold Garfinkel. 1967. *Studies in Ethnomethodology*. Prentice Hall.
- [27] Nigel Gilbert, Robin Wooffitt, and Norman Fraser. 1990. Chapter 11 - Organising Computer Talk. In *Computers and Conversation*, PAUL Luff, NIGEL Gilbert, and DAVID Frohlich (Eds.). Academic Press, London, 235–257. <https://doi.org/10.1016/B978-0-08-050264-9.50016-6>
- [28] Erving Goffman. 1964. The Neglected Situation. *American Anthropologist* 66, 6_PART2 (Dec. 1964), 133–136. https://doi.org/10.1525/aa.1964.66.suppl_3.02a00090
- [29] Erving Goffman. 1967. *Interaction Ritual: Essays on Face-to-Face Interaction*. Aldine, Oxford, England.
- [30] Erving Goffman. 2009. *Relations in Public*. Transaction Publishers.
- [31] Charles Goodwin. 1980. Restarts, Pauses, and the Achievement of a State of Mutual Gaze at Turn-Beginning. *Sociological Inquiry* 50, 3-4 (July 1980), 272–302. <https://doi.org/10.1111/j.1475-682X.1980.tb00023.x>
- [32] Marjorie Harness Goodwin and Charles Goodwin. 1986. Gesture and Coparticipation in the Activity of Searching for a Word. *Semiotica* 62, 1-2 (1986), 51–76.
- [33] A. Hadid, J. Y. Heikkilä, O. Silven, and M. Pietikainen. 2007. Face and Eye Detection for Person Authentication in Mobile Phones. In *2007 First ACM/IEEE International Conference on Distributed Smart Cameras*. 101–108. <https://doi.org/10.1109/ICDSC.2007.4357512>
- [34] Christian Heath, Jon Hindmarsh, and Paul Luff. 2010. *Video in Qualitative Research*. SAGE Publications Ltd, Los Angeles.
- [35] Christian Heath and P Luff. 2000. *Technology in Action*. Cambridge University Press.
- [36] Dirk Heylen, Ivo van Es, Anton Nijholt, and Betsy van Dijk. 2005. Controlling the Gaze of Conversational Agents. In *Advances in Natural Multimodal Dialogue Systems*, Jan C. J. van Kuppevelt, Laila Dybkjær, and Niels Ole Bernsen (Eds.). Springer Netherlands, Dordrecht, 245–262. https://doi.org/10.1007/1-4020-3933-6_11
- [37] Laurent Itti, Nitin Dhavale, and Frederic Pighin. 2003. Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention. In *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*, Vol. 5200. International Society for Optics and Photonics, 64–79. <https://doi.org/10.1117/12.512618>
- [38] Jonas Ivarsson and Christian Greiffenhagen. 2015. The Organization of Turn-Taking in Pool Skate Sessions. *Research on Language and Social Interaction* 48, 4 (2015), 406–429.
- [39] Robert J.K. Jacob, Audrey Girouard, Leanne M. Hirshfield, Michael S. Horn, Orit Shaer, Erin Treacy Solovey, and Jamie Zigelbaum. 2008. Reality-Based Interaction: A Framework for Post-WIMP Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 201–210. <https://doi.org/10.1145/1357054.1357089>
- [40] Robert J. K. Jacob. 1991. The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look at Is What You Get. *ACM Trans. Inf. Syst.* 9, 2 (April 1991), 152–169. <https://doi.org/10.1145/123078.128728>
- [41] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal Human–Computer Interaction: A Survey. *Computer Vision and Image Understanding* 108, 1 (Oct. 2007), 116–134. <https://doi.org/10.1016/j.cviu.2006.10.019>

- [42] Hiroko Kamide, Friederike Eyssel, and Tatsuo Arai. 2013. Psychological Anthropomorphism of Robots. In *Social Robotics (Lecture Notes in Computer Science)*, Guido Herrmann, Martin J. Pearson, Alexander Lenz, Paul Bremner, Adam Spiers, and Ute Leonards (Eds.). Springer International Publishing, 199–208.
- [43] Michael Katzenmaier, Rainer Stiefelhof, and Tanja Schultz. 2004. Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI '04)*. ACM, New York, NY, USA, 144–151. <https://doi.org/10.1145/1027933.1027959>
- [44] Adam Kendon. 1967. Some Functions of Gaze-Direction in Social Interaction. *Acta Psychologica* 26 (Jan. 1967), 22–63. [https://doi.org/10.1016/0001-6918\(67\)90005-4](https://doi.org/10.1016/0001-6918(67)90005-4)
- [45] Chris Kleinke. 1986. Gaze and Eye Contact. *Psychological Bulletin* 100, 1 (July 1986), 78–100.
- [46] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. 2005. A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist (Eds.). Springer Berlin Heidelberg, 329–343.
- [47] Spyros Kousidis and David Schlangen. 2015. The Power of a Glance: Evaluating Embodiment and Turn-Tracking Strategies of an Active Robotic Overhearer. In *2015 AAAI Spring Symposium Series*.
- [48] Manu Kumar, Andreas Paepcke, and Terry Winograd. 2007. EyePoint: Practical Pointing and Selection Using Gaze and Keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '07*. ACM Press, San Jose, California, USA, 421. <https://doi.org/10.1145/1240624.1240692>
- [49] Yoshinori Kuno, Kazuhisa Sadazuka, Michie Kawashima, Keiichi Yamazaki, Akiko Yamazaki, and Hideaki Kuzuoka. 2007. Museum Guide Robot Based on Sociological Interaction Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1191–1194. <https://doi.org/10.1145/1240624.1240804>
- [50] Gene H. Lerner. 2003. Selecting next Speaker: The Context-Sensitive Operation of a Context-Free Organization. *Language in Society* 32, 2 (April 2003), 177–201. <https://doi.org/10.1017/S004740450332202X>
- [51] Florent Levillain and Elisabetta Zibetti. 2017. Behavioral Objects: The Rise of the Evocative Machines. *J. Hum.-Robot Interact.* 6, 1 (May 2017), 4–24. <https://doi.org/10.5898/JHRI.6.1.Levillain>
- [52] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf Between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5286–5297. <https://doi.org/10.1145/2858036.2858288>
- [53] Nikolaos Mavridis. 2015. A Review of Verbal and Non-Verbal Human-Robot Interactive Communication. *Robotics and Autonomous Systems* 63 (Jan. 2015), 22–35. <https://doi.org/10.1016/j.robot.2014.09.031>
- [54] Yohan Moon, Ki Joon Kim, Dong-Hee Shin, Ki Joon Kim, and Dong-Hee Shin. 2016. Voices of the Internet of Things: An Exploration of Multiple Voice Effects in Smart Homes. In *Proceedings of the 4th International Conference on Distributed, Ambient, and Pervasive Interactions*, Vol. 9749. Springer International Publishing, Cham, 270–278. https://doi.org/10.1007/978-3-319-39862-4_25
- [55] Robert J Moore. 2013. Ethnomethodology and Conversation Analysis: Empirical Approaches to the Study of Digital Technology in Action. *The Sage handbook of digital technology research* (2013), 217–235.
- [56] Robert J. Moore and Raphael Arar. 2019. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*. Morgan & Claypool.
- [57] Jonathan Mumm and Bilge Mutlu. 2011. Human-Robot Proxemics: Physical and Psychological Distancing in Human-Robot Interaction. In *Proceedings of the 6th International Conference on Human-Robot Interaction (HRI '11)*. ACM, New York, NY, USA, 331–338. <https://doi.org/10.1145/1957656.1957786>
- [58] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. 2009. Footing in Human-Robot Conversations: How Robots Might Shape Participant Roles Using Gaze Cues. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction (HRI '09)*. ACM, New York, NY, USA, 61–68. <https://doi.org/10.1145/1514095.1514109>
- [59] D. G. Novick, B. Hansen, and K. Ward. 1996. Coordinating Turn-Taking with Gaze. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 3. 1888–1891 vol.3. <https://doi.org/10.1109/ICSLP.1996.608001>
- [60] William Odom, John Zimmerman, Jodi Forlizzi, Hajin Choi, Stephanie Meier, and Angela Park. 2012. Investigating the Presence, Form and Behavior of Virtual Possessions in the Context of a Teen Bedroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 327–336. <https://doi.org/10.1145/2207676.2207722>
- [61] Per Persson, Jarmo Laaksohaki, and Peter Lönnqvist. 2000. Anthropomorphism - a Multi-Layered Phenomenon.
- [62] Antoine Picot, Gérard Bailly, Frédéric Elisei, and Stephan Raidt. 2007. Scrutinizing Natural Scenes: Controlling the Gaze of an Embodied Conversational Agent. In *7th International Conference on Intelligent Virtual Agents, IVA '2007 (17-19 September 2007, Paris, France)*. Paris, France, 50–61.
- [63] Stefania Pizza, Barry Brown, Donald McMillan, and Airi Lampinen. 2016. Smartwatch in Vivo. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5456–5469.

<https://doi.org/10.1145/2858036.2858522>

- [64] Alex Poole and Linden J. Ball. 2006. Eye Tracking in HCI and Usability Research. In *Encyclopedia of Human Computer Interaction*. IGI Global, 211–219.
- [65] Martin Porcheron, Joel E. Fischer, Moira McGregor, Barry Brown, Ewa Luger, Heloisa Candello, and Kenton O’Hara. 2017. Talking with Conversational Agents in Collaborative Action. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW ’17 Companion)*. ACM, New York, NY, USA, 431–436. <https://doi.org/10.1145/3022198.3022666>
- [66] 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>
- [67] Matthias Rehm and Elisabeth André. 2005. Where Do They Look? Gaze Behaviors of Multiple Users Interacting with an Embodied Conversational Agent. In *Intelligent Virtual Agents (Lecture Notes in Computer Science)*, Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist (Eds.). Springer Berlin Heidelberg, 241–252.
- [68] C. Rich, B. Ponsler, A. Holroyd, and C. L. Sidner. 2010. Recognizing Engagement in Human-Robot Interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 375–382. <https://doi.org/10.1109/HRI.2010.5453163>
- [69] Viktor Richter, Birte Carlmeyer, Florian Lier, Sebastian Meyer zu Borgsen, David Schlangen, Franz Kummert, Sven Wachsmuth, and Britta Wrede. 2016. Are You Talking to Me?: Improving the Robustness of Dialogue Systems in a Multi Party HRI Scenario by Incorporating Gaze Direction and Lip Movement of Attendees. In *Proceedings of the Fourth International Conference on Human Agent Interaction (HAI ’16)*. ACM, New York, NY, USA, 43–50. <https://doi.org/10.1145/2974804.2974823>
- [70] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell. 2015. A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception. *Computer Graphics Forum* 34, 6 (Sept. 2015), 299–326. <https://doi.org/10.1111/cgf.12603>
- [71] H. Sacks, E. A. Schegloff, and G. Jefferson. 1974. A Simplest Systematics for the Organization of Turn Taking for Conversation. *Language* 50 (1974), 696–735.
- [72] Emanuel A. Schegloff. 1968. Sequencing in Conversational Openings1. *American Anthropologist* 70, 6 (Dec. 1968), 1075–1095. <https://doi.org/10.1525/aa.1968.70.6.02a00030>
- [73] 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>
- [74] Abigail J. Sellen. 1995. Remote Conversations: The Effects of Mediating Talk With Technology. *Human-Computer Interaction* 10, 4 (Dec. 1995), 401–444. https://doi.org/10.1207/s15327051hcci004_2
- [75] O. Shaer and E. Hornecker. 2010. *Tangible User Interfaces: Past, Present and Future Directions*. now.
- [76] Candace L Sidner, Cory D Kidd, Christopher Lee, and Neal Lesh. [n. d.]. Where to Look: A Study of Human-Robot Engagement. ([n. d.]), 7.
- [77] Gabriel Skantze, Martin Johansson, and Jonas Beskow. 2015. Exploring Turn-Taking Cues in Multi-Party Human-Robot Discussions About Objects. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI ’15)*. ACM, New York, NY, USA, 67–74. <https://doi.org/10.1145/2818346.2820749>
- [78] Masataka Suzuki, Ayano Izawa, Kazushi Takahashi, and Yoshihiko Yamazaki. 2008. The Coordination of Eye, Head, and Arm Movements during Rapid Gaze Orienting and Arm Pointing. *Experimental Brain Research* 184, 4 (Feb. 2008), 579–585. <https://doi.org/10.1007/s00221-007-1222-7>
- [79] Daniel Szafrin and Bilge Mutlu. 2012. Pay Attention!: Designing Adaptive Agents That Monitor and Improve User Engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI ’12)*. ACM, New York, NY, USA, 11–20. <https://doi.org/10.1145/2207676.2207679>
- [80] James W. Tankard. 1970. Effects of Eye Position on Person Perception. *Perceptual and Motor Skills* 31, 3 (Dec. 1970), 883–893. <https://doi.org/10.2466/pms.1970.31.3.883>
- [81] Roel Vertegaal, Robert Slagter, and Anton Nijholt. 2001. Eye Gaze Patterns in Conversations: There Is More to Conversational Agents Than Meets the Eyes. (2001), 8.
- [82] Roel Vertegaal and Harro Vons. 2000. Effects of Gaze on Multiparty Mediated Communication. *Graphics Interface* (2000), 95–102.
- [83] TIAN (LINGER) XU, HUI ZHANG, and CHEN YU. 2016. See You See Me: The Role of Eye Contact in Multimodal Human-Robot Interaction. *ACM transactions on interactive intelligent systems* 6, 1 (May 2016). <https://doi.org/10.1145/2882970>
- [84] Tomoko Yonezawa, Hirotake Yamazoe, Akira Utsumi, and Shinji Abe. 2007. Gaze-Communicative Behavior of Stuffed-Toy Robot with Joint Attention and Eye Contact Based on Ambient Gaze-Tracking. In *Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI ’07)*. ACM, New York, NY, USA, 140–145. <https://doi.org/10.1145/1322192>

1322218

- [85] Oren Zuckerman and Ayelet Gal-Oz. 2013. To TUI or Not to TUI: Evaluating Performance and Preference in Tangible vs. Graphical User Interfaces. *International Journal of Human-Computer Studies* 71, 7 (July 2013), 803–820. <https://doi.org/10.1016/j.ijhcs.2013.04.003>

Received April 2019; revised June 2019; accepted August 2019