Patterns of Gaze in Speech Agent Interaction

Razan Jaber
razan@dsv.su.se
Department of Computer and System Sciences
Stockholm University
Stockholm, Sweden

Donald McMillan
donald.mcmillan@dsv.su.se
Department of Computer and System Sciences
Stockholm University
Stockholm, Sweden

Jordi Solsona Belenguer
jordi@dsv.su.se
Department of Computer and System Sciences
Stockholm University
Stockholm, Sweden

Barry Brown
barry@dsv.su.se
Department of Computer and System Sciences
Stockholm University
Stockholm, Sweden

ABSTRACT

While gaze is an important part of human to human interaction, it has been neglected in the design of conversational agents. In this paper, we report on our experiments with adding gaze to a conventional speech agent system. Tama is a speech agent that makes use of users' gaze to initiate a query, rather than a wake word or phrase. In this paper, we analyse the patterns of detected gaze when interacting with the device. We use k-means clustering of the log data from ten users tested in a dual-participant discussion tasks. These patterns are verified and explained through close analysis of the video data of the trials. We present similarities of patterns between conditions both when querying the agent and listening to the answers. We also present the analysis of patterns detected when only in the gaze condition. Users can take advantage of their understanding of gaze in conversation to interact with a gaze-enabled agent but are also able to fluently adjust their use of gaze to interact with the technology successfully. Our results point to some patterns of interaction which can be used as a starting point to build gaze-awareness into voice-user interfaces.

CCS CONCEPTS

• **Human-centered computing → Interaction techniques**; *User studies*; *Interaction design.*

KEYWORDS

Smart Speaker, Voice Assistant, Gaze Interaction, Eye-Tracking

ACM Reference Format:
Razan Jaber, Donald McMillan, Jordi Solsona Belenguer, and Barry Brown. 2019. Patterns of Gaze in Speech Agent Interaction. In *1st International Conference on Conversational User Interfaces (CUI 2019), August 22–23, 2019, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3342775.3342791

CUI 2019, August 22–23, 2019, Dublin, Ireland

ACM ISBN 978-1-4503-7187-2/19/08…$15.00
https://doi.org/10.1145/3342775.3342791

1 INTRODUCTION

Voice assistants, such as Apple's Siri and Amazon's Alexa are available on all major smartphone and tablet operating systems, and are increasingly being deployed in homes in stand-alone smart-speaker devices. While these systems have had great success, they usually rely on a single modality of interaction — transcribed speech. In human to human conversation, speech is augmented with a range of other modalities — such as gaze, touch, prosody, and gesture. Yet these are mostly ignored in the current generation of speech systems. To address this, we have developed Tama, a gaze aware interactive speech agent. Tama uses gaze to trigger user's query, and can respond to users by establishing and holding mutual gaze.

Building on the Google Assistant service, Tama acts as a voice assistant. Tama has a tapered cylindrical shape, with a semi-transparent, retractable, dome-shaped head on top (see Figure 1). Inside the head, two full-colour LED eyes can move and 'look at' a user, establishing mutual gaze.

Tama is the first step towards a speech agent that takes advantage of multiple modalities to better fit with the complex, messy, human situations in which they are often used [48, 58]. There have been a number of sensors that can detect the direction of a user's gaze for system input [15] and evaluation [47] for some considerable time. However, using gaze for system interaction has traditionally been restricted to the domains of accessibility [32] and to support more natural, socially aware, human-robot interaction [1]. Yet gaze has potential to support a wider range of interactions with computational systems — such as in the work of Shell et al. [59].

For this paper we tested Tama's use in a realistic experimental task, recording users' interaction from 2 camera angles. Users were asked in pairs to agree on a close holiday destination that they both agreed on, and both had not visited. They were asked to use the Tama speech agent when appropriate.

The results in this paper focus on gaze patterns of interaction. For each interaction with Tama we generated a two 10-point vectors of the amount of gaze detected by the cameras, one while the query is being voiced, and the other while the answer is played through the speaker. These were clustered using k-means, uncovering **5** 'patterns of looking' while voicing queries, and **3** while listening to the answer which were then verified with close analysis of samples of the video data from the trial. Gaze detection is relatively complex and dependant on lighting, angle, distance, cameras, and

the training set employed in building the model. As such, these detected patterns should be taken as averages of orchestrated human interaction towards the system, able to be identified through current technology, and verified by close analysis of samples of the video data of the trial.

We show that our participants were detected looking at the assistant in similar ways before (using the wake-word in the control condition) and after gaze being introduced as an interaction mechanism. We also present patterns of gaze which exemplify types of interaction the participants were detected to be engaged in with the system. We end with a discussion on how anthropomorphic interactions can be detrimental to learning effective use and interaction with the system and the opportunities presented by gaze in conversational user interfaces.

2 BACKGROUND

2.1 Speech agents

Recently, there has been a wealth of research that investigates aspects of interaction with speech agents. Looking at Smart Speaker use in particular, Sciuto et al. [58] investigated the experience of households who are using conversational agents in their homes showing how families integrate conversational agent into their daily life. By analysing the logs of 75 Alexa users, they concluded their findings into four themes: how people initially use Alexa, the physical placement of the device, the daily patterns of conversational usage, and how children interact with the device.

Porcheron et al. [48] identify the characteristics of interaction with VUIs on mobile devices and how such interactions unfold in multi-party social settings. In further work [49] they examine of how 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. [40] conducted a between-subjects experiment to investigate the relationship between users' personality and a number of voice technologies for Smart Home environment, and how it affects users' social responses to these systems.

Luger and Sellen [37] highlighted the limited functionality of existing commercially-available voice interfaces and how it causes a gulf between their capabilities and the users' expectations. Meyer and Rakotonirainy [39] summarised the problems with commercialising speech agents as; installation, management, novelty of applications, quality of the user experience, and privacy — most of which remain challenges.

Although the only direct sensors on smart speakers are their microphones, microphones are perceived as one of the most privacy-invasive sensors next to video cameras [9]. This, as well as their placement in intimate spaces such as users' homes, results in smart speakers posing particular privacy challenges. Wake-word recognition is generally performed locally. To complement this, most smart speakers have a physical button to control the microphones. Companion apps and websites allow users to review and delete voice interactions with the device. That is said, privacy is one area where we hope Tama and gaze actuation can provide part of a solution.

2.2 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 [65, 68]. Goffman [20] observed that the direction of eye gaze plays a crucial role in the initiation and maintenance of social encounters. Kendon [31] conducted a detailed exploration of the function of gaze in face-to-face conversation. He 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 [31]. He observed how speakers gaze at their partner when they about to end their phrase, and how they averted gaze during hesitations.

Eye gaze can be used to signal both the end and the start of a speaking turn [27], express dissatisfaction or uncertainty [42], regulate turn taking [16], convey information, and regulate social intimacy [4, 31].

Vertegaal et al. [67] concluded that despite this variation, gaze is a predictor for turn taking, estimating an 88% chance that the person looked at is the person being listened to. Hearers gaze at speakers more than speakers gaze at hearers [3, 31, 42]. Speakers also tend to look away as they begin talking [3, 21, 31]. It has been suggested that gaze fills both a monitoring and regulating role [31].

Goodwin proposed [22] two rules of gaze, 1: a speaker should obtain the gaze of his recipient during the course of a turn of talk, and 2: A recipient should be gazing at the speaker when the speaker is gazing at the hearer. While gazing in conversation is driven by different contexts, for technology, we can use the understandings that gazing-toward and gazing-away from are related to turn-taking and regulation.

Many of the studies mentioned earlier are a result of research in American or British English in western societies [3, 4, 21, 31, 42]. These studies implicitly suggest that gaze behaviours are independent of attributes such as race, culture, and gender. However, some studies have explored such cultural and language differences concerning the use of gaze. For example, Rossano, et al. [53] investigated gaze behaviour in conversation using data from three different cultures and languages, with a focus on pervasive conversational practice when people gaze at each other during conversation. In their study, they have worked to understand whether gaze as an international practice has these universal properties across cultures as implied above. While they find similarities, they also show differences in the practical use of gaze between cultures and distinct cultural practices [52].

In general, Rossano, et al. [52, 53] find that across all cultures, gaze is tied to sequence initiation and sequence completion rather than directly to the turn-taking system. However, they observe practices that are culturally specific such as alternative "home positions" for the eyes, and the use of gaze to point to unseen subjects in some cultures more than others. Sacks [56] posits that most of the time during the daily interactions we wish to present our gaze as 'ordinary'. This suggests that there are norms associated with our use of the eyes during social interactions that are practices we deploy to sustain 'ordinary' gaze behaviour. These practices have to be learned and could be patterned somewhat differently in different cultures.

2.3 Gaze in interaction

Research has also looked at the mechanics of gaze interaction in different contexts [13, 34, 44], harnessing eye motion and gaze gestures for interaction [15], detecting patterns of gaze [14] and for enhancing human-robot engagement [8, 60]. Particularly gaze in interactions has been used for onscreen target pointing, to support selection tasks [69], selecting objects in large information space, as well as for zooming and panning [24, 61].

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 [7, 30]. Gaze has been extensively used to augment other input modalities [29], as well as to enhance the social interaction in using eye gaze for human-robot interaction with a focus on communicative social gaze within the interaction [1, 55]. Moreover, many studies have been conducted to investigate the use of gaze cues with conversational agents [11, 19, 26, 50, 60]. Exploration of the effect of gaze cues in turn taking in two- and multi-party discourse has been a popular area of research [6, 17, 33, 63, 67].

Most of these studies have focused on understanding how these cues might shape participant roles and how different forms of participation might affect the social outcome of human-agent conversations. Therefore, the complex roles that gaze play in human interaction have been the focus of many studies including the integration of vision and speech to show attention to conversational partners and objects in the surroundings [6, 64], for example, Szafir and Mutlu [64] built an embodied agent that monitored people attention and adapted its behaviour to improve the discourse and user engagement. Most of the studies with virtual agents and social robots have used eye gaze as a signal of capturing attention [28], demonstrating engagement [51], and increasing conversational fluidity with human users [10, 38].

Other research has focused on controlling the eye gaze of virtual embodied conversational agents [45]. Andrist et al. [2] presented a conversational gaze aversion robot able to generate and combine head motions to engaging in mutual gaze with users. Using gaze aversion can be used to demonstrate cognitive efforts, modulate intimacy, and mediate turn taking. Moreover, researchers in social robotics have found that the direction of gaze plays an important role in shaping conversational participant roles, which can be effective to shape people roles when designing virtual agents [5, 7, 41].

One of the most critical applications of gaze research is the design of social robots with appropriate gaze behaviour. Studies showed that virtual agents using gaze aversions are more successful at regulating the conversational flow than agents that do not perform gaze aversions [54]. Mutlu et al. showing conversation coordination through the establishment or breaking of eye contact [41].

3 DESIGNING TAMA

Interaction with Tama was based on the hypothesis that the smart speaker would be brought into a conversation much the same way as another human conversational partner [57], using the offer of mutual gaze and its reciprocation.

Tama is built like a tapered cylindrical 3D printed shell which shape is similar to Google home (Figure 1). On top, there is a semi-transparent dome-shaped head, the head can be retracted entirely inside the body for when mutual gaze interaction is not required. Inside the head, two full-colour LED eyes can move 360 by 40 degrees (pan and tilt) to perform mutual gaze at different heights and angles. The bottom of Tama has a LED light ring functionality for feedback when the head is not in use.

Figure 1: Tama. Closed head (left), and Mutual Gaze condition (right)

Two OMRON HVC-P2 cameras [43], which provide of the on-board face and gaze identification, were used to enable gaze interaction. These were connected to a Raspberry Pi 3 [18] along with an Arduino control board [36] which controlled the eye movements and a Respeaker directional 7-microphone array [62] to record and detect the direction of speech. For the voice assistant service, we used Google's Voice Assistant API.

3.1 Interaction Design

Interaction with Tama involved transitions between the following three states:

Idle: Tama is looking straight forward with the microphone off and eyes coloured yellow.

Activated: Tama has detected at least one user looking at, and starts the assistant and microphone. Tama will move the eyes towards the detected gaze and turn them green (Figure 1, right). If the two users are looking at Tama, the directional microphone is used to distinguish and prioritise the speaker. If there is an around 3 second period without detecting gaze, Tama will return to Idle and cancel the query.

Responding: Tama's eyes will turn pink, and as the answer is broadcast, it will look towards any gaze detected.

This interaction flow was designed to balance ease of intentional activation with minimisation of false activation. In a conversational scenario, the cancellation of interactions started with cursory or un-maintained glances towards the device stopped the device triggering when it was not supposed to, but at the cost of making it harder to complete queries when the conditions for gaze detection were not optimal.

4 EXPERIMENT

For the experiment, we recruited ten pairs of participants to perform three conditions. The trials lasted around 45 minutes, with the longest lasting 53 and the shortest 19 minutes. Of the *20* participants, *13* identified as male, *1* owned smart speakers, and a further *12*

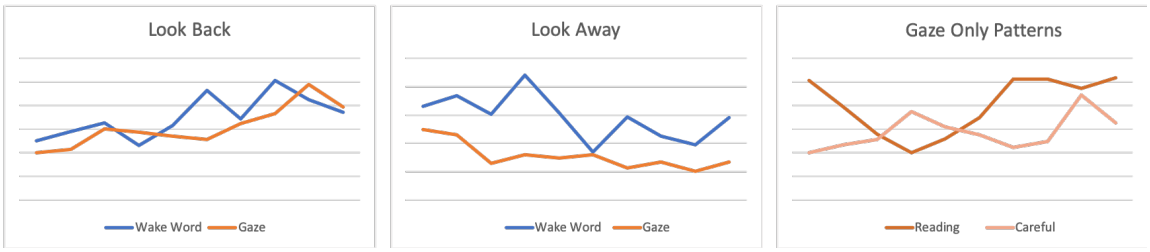

Figure 2: Gaze patterns during query: look back (left), Look away(middle), gaze only (right)

reported using voice assistants on other devices regularly. All but one of the participants were postgraduate students. Each participant received a $10 gift card.

As our system was designed to integrate into the conversation in a different way than the existing 'wake word' model, we were interested in how system use could be incorporated into an ongoing talk. Accordingly, we designed a task where two participants were to decide and agree on a holiday destination with a constraint that it should be the closest capital city from a given starting point that they both agreed on, and both had not visited. They were asked to use the Tama speech agent when appropriate.

For the trial, the participants faced each other across a table with Tama equidistant from them creating a triad. Each trial was recorded by two GoPro Hero 3 cameras, one behind and one facing the smart speaker. The trial was also streamed via webcam to the authors in a separate room. One author entered the room to explain each condition, leaving them alone to complete the task. Additionally, all events and data generated in the cameras and microphones was logged internally by Tama during the test for post analysis.

We had one control condition which each test started with (wake-word activation), and two test conditions which were balanced in order. For each condition, the starting city for the task was changed. In the mutual-gaze condition the device was activated with gaze and the head looked back, in the gaze-activation condition the head was retracted. 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.

5 RESULTS

With a system such as Tama the gaze patterns of users are crucial for understanding how the system works. In this paper we focus on our participants gaze patterns when using Tama in different ways. We analysed two datasets generated from the user trial. The first is the accumulation of logs from Tama during the trial. This collected data on the system's internal state and the interactions with the Google Assistant service. Here we focus on how the gaze cameras recorded where each participant was looking, and the direction-of-arrival of sound from the directional microphone. The output from the gaze cameras should not be seen as a 'ground truth' of participants' gaze towards the device in any moment, rather the overall pattern of gaze interaction that can be detected given the constraints of technology and setting.

This provides us a binary signal from the query owner and the query partner representing whether they are detected as looking at Tama or not throughout the trial. The log data was filtered to only include complete queries, i.e. ones that had both a successfully articulated query and an answer voiced by the speech agent. As a result, the dataset examined for patterns of gaze interaction contained **509** queries, with **221** in the wake-word condition, **135** in the gaze-activation, and **153** in the mutual-gaze conditions.

To deepen our understanding of what this means for the moment-by-moment interaction between our participants and the conversational agent we combined this with video analysis of the recordings of the trials. Both the video angles we recorded 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.

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. That is why we have transcribed those interactions as seen in Figures 3, 4, 5 and 8. We have added a 'glance track' under the transcript of what was spoken. The arrows indicate gaze – such as $P_1 - - > P_2$ for participant one looking at participant two, with mutual gaze indicated by double headed arrow e.g. $P_1 < -- > P_2$. 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 in these transcripts [25, appendix a][46]. The numbers in (brackets) indicate pauses, and we have broken these out into multiple (items) where the gaze changes (indicated on the gaze track). We have also added photographs of the interactions, indicated by an * in the transcript.

This gives us a dataset of 509 queries, and for each query we have second by second data on which participant was looking at Tama (or not), what was being said and Tama's answer.

To look for patterns in the gaze interaction we started by taking the data of gaze recorded by Tama's cameras. For each query we extracted the time period where the query was being voiced from

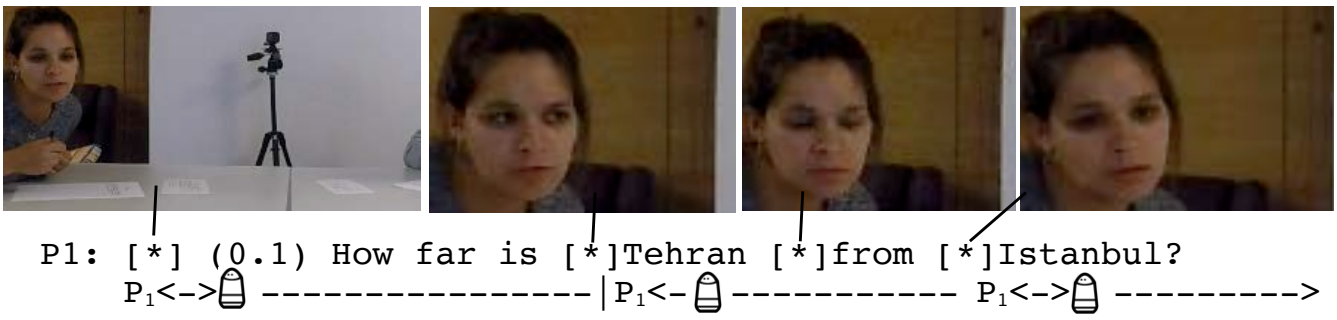

Figure 3: Transcript of Look-Back Interaction.

the time waiting and listening to the answer. For the gaze conditions we excluded the required activating gaze at the start of the query.

For each of our extracts we split them into ten equal time segments. For each of these segments we then measured the proportion of time that gaze was maintained - this gave us 10 measurements of how the gaze 'pattern' unfolded over the time. This ten dimension pattern is an approximation of the gaze behaviour for each query (or answer). Across the whole trial, queries had a mean length of 10.48 seconds, so this means each decile summarises gaze for around one second of interaction. Since the gaze cameras had some issues with accuracy this improves the accuracy of our gaze measurement as well as allowing us to compare patterns across different queries and trials. Looking at these resulting 10 dimension abstractions of gaze over time it became apparent that the patterns were consistent across both gaze conditions (i.e, the gaze-activation and the mutual-gaze conditions), with a chi-square showing no significant difference between detected gaze between the two (p=0.309), and combining them provided a stronger fit for clustering using K-means.

K-means clustering intends to partition objects into a fixed number, k, of clusters, which each object belonging to the cluster with the nearest mean (or, in this case, 10 dimension centroid) so that the within-cluster sum of squares is minimised. K-means tries to make the clusters as coherent, but as far apart from each other as possible. In each case, a range of possible values of k, or number of clusters, were tried and the results with the largest differentiation between clusters chosen as representative of the data. During this process each of the k centroids is initially a random data point from the dataset. Then the rest of the points in the dataset are added sequentially, adding each point to its closest cluster and recomputing the cluster's centroid. The process is then repeated starting with these new centroids. This continues until the centroids are stable. In these results, k-means was run with k from 2 to 6, and silhouette analysis [66] was used to determine the separation distance between the resulting clusters in order to choose the optimal value of k. This gives us measurements of different types of gaze patterns across the participant's queries - graphing the centroids gives us a way of looking at the different gaze patterns in our trials. For each cluster, the patterns of detected gaze assigned to it were inspected by three authors and interpreted through the lens of the close moment-by-moment understanding of the participants' interactions with Tama with a selection traced back to the transcribed query and its resulting video.

5.1 Common Query Patterns

One of the goals behind the development of Tama was to explore the possibility of activating a speech agent using gaze towards the devices, rather than — or in combination with — the wake word.

The patterns reported here are from the query owners, those who voiced the query, between activation with the wake-word or gaze, and the query being deemed complete by the Google Assistant service and sent to be processed for a reply. In each case k-means was run with k from 2 to 6, the result here presented as the clustering with the highest silhouette coefficient. In the wake-word condition 3 clusters with silhouette (0.504), and the gaze conditions 5 clusters with silhouette (0.167). One interesting thing to note here is that not only did Tama detected that participants looked at it during queries in the wake-word condition, but that the patterns of that looking shared very similar characteristics to that of the gaze conditions.

Figures 2 (left and centre) show two similar patterned centroids between the gaze and wake-word conditions.

The pattern shown in Figure 2 (centre) labelled 'look away', was detected in 17% of gaze queries, and 5% in the wake-word condition. In interactions that follow this pattern, users are detected looking at the device as they begin to voice their query, but this tails off as the query progresses. The second pattern detected during queries in both wake-word and gaze conditions was that shown in Figure 2 (left) labelled 'look back', where the participant was recognised looking towards Tama only as they drew to the end of their query. Such a pattern of gaze follows observations of interpersonal gaze explored by Goodwin [21], where in coming to the end of a turn, a conversational partner will use gaze as a means of speaker selection. An example of this can be seen in the transcript in Figure 3. Here our participant starts by establishing mutual gaze, then after a 0.1 second pause begins to articulate her query. As she reaches the object of the question, she turns her gaze to her left, then down, and then finally back to Tama on the final word of the query.

As noted above, the clustering algorithm settled on 3 clusters of gaze patterns for the wake-word conditions and 5 for the gaze conditions. The final pattern shared between both conditions was that of a consistent gaze (not included in figure 2) – be that at the

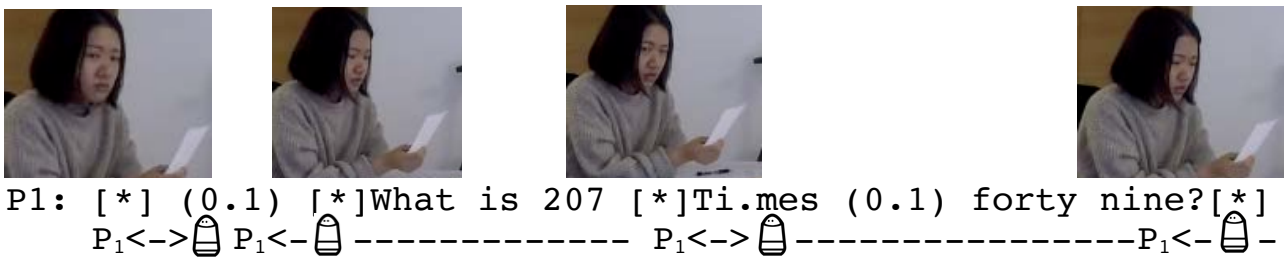

Figure 4: Transcript of reading interaction.

device, or not at the device. The k-means placed 80% of interactions with the wake-word, and 46% with the gaze condition in this category. In the gaze condition, the majority of these interactions involved the participants detected as looking at Tama throughout the query. This was a learned behaviour, as participants enacted gaze to ensure that the system continued to process their query and provided them with the answer that they sought. As noted above, the successful gaze queries clustered here were in almost 32% of cases coming after one or more unsuccessful attempts at interaction resulting in a repeat. One reason for these queries to be unsuccessful was the system being unable to detect gaze from a user for longer than the 3 second threshold and cancelling the query on the assumption that this was a spurious activation, either due to them looking away for too long or problems with the lighting or angle of the face. For users who experienced such interactional troubles, one method of reducing them was to fixate their gaze on the device throughout the query, resulting in a consistent gaze pattern as described in this cluster. For the wake-word condition, all of the inspected interactions with this pattern showed little to no gaze detected by the device.

5.2 Gaze Specific Query Patterns

The right hand graph in Figure 2 (right) is labelled 'reading', and 'careful' - both of 12%. The condition we labelled as 'reading' is representative of gaze being detected at the start of the query and the end, with it dropping off in the middle. While the example interactions explored in this cluster were contained examples where the participants referred to their notes on what cities they were discussing while voicing their query, it also provided examples of the participants glancing towards their conversational partner while constructing the query. In Figure 4, we can see our participant performing a test query for the device. Here the clip begins with her establishing mutual gaze with Tama, then before starting her query, she turns to focus on the prompt card, looking back towards Tama a few words later, then returning to look at the card while awaiting the answer. This kind of interaction was fraught with trouble for completing the query, as looking away while asking the question was taken by Tama to signal an accidental activation. Readers quickly learned to pepper their reading with glances towards the device. The pattern we have daubed 'careful' here is epitomised by gaze being detected in a staccato pattern, suggesting frequent looking away and back at Tama. Looking at the clips that this was representative of, however, we found that this was usually a result of problematic detection. In Figure 5, we can see our query owner, on the left, initiate gaze then continue to fixate on the device throughout the query until an answer was heard. For reasons of lighting and facial features, however, the gaze cameras detected this as intermittent gaze. In this trial, this participant had learned to continuously look at the device to be able to interact successfully.

5.3 Answers

In looking at the detected gaze while the Google Assistant articulated its response, we include the traces from both the owner of the query and the partner. In both the gaze (silhouette owner: 0.477, partner: 0.511) and wake-word conditions (silhouette owner:0.59, partner:0.647) there were three clusters detected.

Looking at the centroids for these clusters in Figures 6 and 7, we can see the same patterns emerging in both user-roles and in both conditions. While the most common pattern seemed to be to simply not look at the device while it was replying (79%, 81%, 86%, 86% for gaze-owner, gaze-partner, wake-owner, and wake-partner respectively), the answers showed small clusters of participants being detected looking at the device as it started to speak and tailing off (15%, 5%, 4%, 10% for gaze-owner, gaze-partner, wake-owner, and wake-partner respectively, Figure 7) as well as clusters of participants not detected looking at the device at the beginning, but detected as gazing at it more and more as the answer continued (3%, 16%, 10%, 4% for gaze-owner, gaze-partner, wake-owner, and wake-partner respectively, Figure 6).

An example of a 'look away' during the answer can be seen in the transcript in Figure 8, which also provides an example of the user looking towards the device while using the wake-word to activate it. Here the participant glances towards the device as he articulates 'Ok, Google' then looks down towards his task instructions. After a 3.8 second pause as the Google API processes this request, Tama started to articulate its answer – at which point the query owner looked towards the device (the partner did not). As the answer starts to seem suspect (with 15 items reported), the participants look at each other, bursting into laughter when the source of this information turns out to be TwistedSister.com, the home page of a heavy metal band and the 15 items tour locations.

6 DISCUSSION

As covered in the background section above, there are several methods for detecting the angle of gaze from users. However, as distance increases, angles become more acute to the camera, and lighting

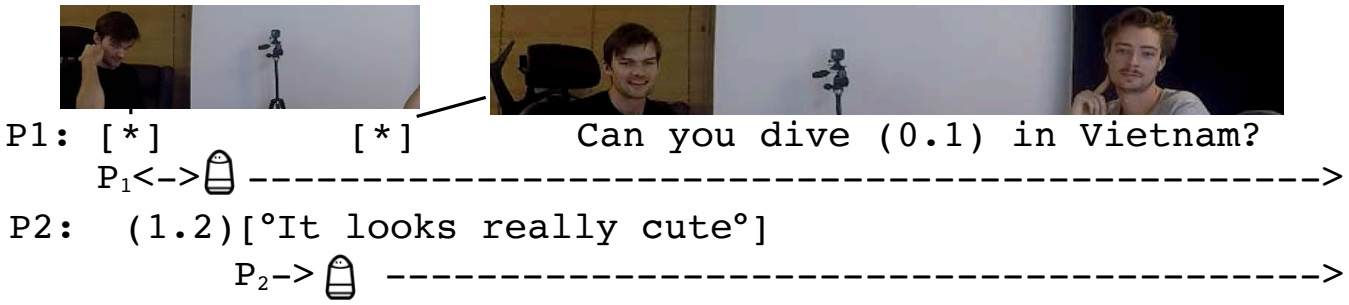

Figure 5: Transcript of careful interaction, the participants continuously look at Tama over the Query.

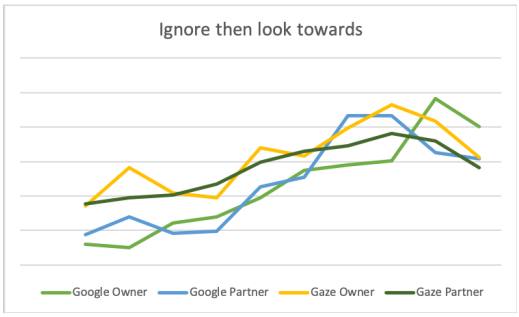

Figure 6: Look Towards Answer

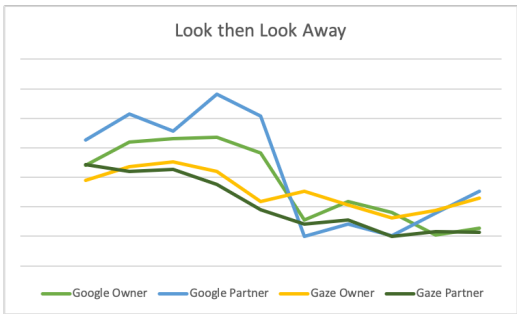

Figure 7: Look Away Answer

varies, all of these methods become more difficult. The trial setup in this paper represents controlled conditions, with users under 2m from the camera with bright light, and even then they were not perfectly accurate. In testing the cameras used here were able to work in more varied lighting conditions and at greater distances, however, such conditions caused more frequent losses of detected gaze and higher sensitivity to the angle of the head, glasses, or shadows. As with a lot of computer vision related research, increasing the speed, detail, and contrast of the image helps — yet the costs increase exponentially while the accuracy does not [23].

This presents considerable challenges in terms of accuracy. The most basic interaction with Tama replaces the wake-word ('Ok, Google') with a look towards the system while articulating a query. Wake-word technology has been well researched with models deployed by the main smart speaker producers proven to be robust, trained using hundreds of thousands of real-world examples [35], and highly accurate in most settings. Gaze interaction, on the other hand, is relatively complex and dependant on lighting, angle, distance, cameras, and the training set employed in building the model. As such, our gaze detection cameras, and therefore Tama's detection of intended interaction is less accurate than using a wake-word.

One alternative design would be to make use of infra-red detection of gaze, yet this has problems with natural light, as the sun is a problematic light source to compete with, reducing the real world application options. State of the art solutions achieve around 85% accuracy on real-world gaze detection tasks [12]. Given all these caveats on the accuracy of the cameras in any given 250-millisecond detection frame, the queries were successful around 72% of the time using gaze, and the patterns described here are supported by our close analysis of the videos. We see this as an example of designing interaction in a heuristic loop, where imperfect probabilistic algorithms detect, react to, and *influence* human behaviour.

Participants used a number of strategies, or modified their use of gaze and their behaviour towards the system, in light of these inaccuracies and limitations — as shown in the results, patterns of careful reading and looking were emergent in response to how the system detected the participants over the short interactions in the trial. Also, participants experimented with waving, weaving their head back and forth, taking off glasses, and covering their face so that their partner could have what they thought was 'uncontested' interaction with the system. This all points to a design challenge of surfacing not just that systems based on machine learning algorithms should expose that they have been unable to understand the users' interaction — in the case of Tama, by looking away or for the Google Speech agent by replying with a generic 'I can't help with that' message — but that they should support the user in repair. If Tama could provide more detailed, moment by moment output on the state of detection in a way that would not impact the interaction unless it was engaged with, then users could adjust their interaction with more certainty and accuracy to avoid problems.

For speech agents in general, simply providing more gradients of error would allow users to repair that which was causing the problem, rather than adjusting clear parts of the interaction. For example, differentiating between un-processable transcription results with a high noise ratio on the audio and those with a low

Figure 8: Wake-word query transcript. Participant looks at Tama while saying the wake-word, then looks back when it replies

noise ratio with answers along the lines of 'I couldn't hear.' and 'I didn't understand.' would allow the user to repair with increased volume or attempted rephrasing of the query. In our data, we saw a number of trial and error approaches to this, where the generic error message would first result in one or more louder attempts of the query, before the participants would attempt to adjust the phrasing to one that the system might be more adept at parsing.

Learning to modulate your speech, including word choice, pitch, and volume, in order to successfully and reliably interact with voice assistants is for many a non-trivial task. One problem identified with interacting with these devices [49] is that the systems' listen from the wake-word until they detect silence. This means that the utterance must be performed quickly, with no breaks between words or phrases — something which for many can be unduly challenging. By detecting if the user is still looking at the speech agent, such gaps in speech could be ignored if the current transcribed utterance does not conform to the rules of a fully formed query that the agent can act upon, allowing for much more variation in the speech that the system can support. This could also be put to use in noisy environments by using the directional microphones that most smart-speakers are built around to filter out audio from directions other than from where the user's gaze is coming from.

6.1 Opportunities of Designing with Gaze

The patterns identified in the previous section provide opportunities for improving interaction with our system, and with a broader range of voice interfaces.

The first area of opportunity centres around improving the accuracy of query detection, especially in challenging audio environments. By leveraging the pattern of 'look back' in query formulation, we can take advantage of a second channel of information signalling the approaching end of a query. This allows us to move beyond relying on the combination of grammatical inference and silence detection currently employed. Similarly, detecting that a user is 'reading' to the system gives the opportunity to increase the threshold of silence detection to allow them to complete this task.

An opportunity to improve the ongoing fit with the context of use is presented by the 'look away' condition while consuming an answer. Here the system could detect the lack of attention and engagement, and, for example, lower its output volume accordingly.

Another opportunity presented in the 'careful' condition. Understanding when and if users are experiencing trouble interacting with an interface allows designers to either modify the interface to suit (in this case, by increasing sensitivity for example) or to provide feedback to users to improve their interaction – such as recommendations on posture or lighting.

7 CONCLUSION

In this paper, we present an analysis of the patterns of gaze detected by our gaze-enabled voice assistant, Tama. This is based on the data from 10 pairs of users interacting with Tama during a discussion task and supported by video analysis of their interactions.

During conversations, gaze is one tool employed to establish and maintain interaction successfully. In this paper, we identify ways in which abstractions of this complex human behaviour can be used to drive interaction with a speech agent, and how some patterns of gaze are detected during interactions even when the agent does not respond to, or present, gaze. In the presentation of gaze, future work could include embedding these gaze patterns in the behaviour of our speech agent, this holds the potential for using these patterns to improve speech agent interaction. This could also be combined with a longer term, in-the-wild trial to understand how these patterns persist, or change over time, in real use contexts.

We also identified patterns of gaze which were detected during problematic interactions with the device, showing not only that these hold the potential to be detected to aid in the interaction but also that users are able to fluently adjust their use of gaze from mostly use with a conversational partner to explicit, articulated use for system interaction. In future work, there exists an exciting opportunity to explore the differences between both language dependent and culturally dependent patterns of gaze interaction with this system. This could provide more a nuanced understanding of gaze for the design of interaction.

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] 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

[3] Michael Argyle, Mansur Lalljee, and Mark Cook. 1968. The Effects of Visibility on Interaction in a Dyad. *Human relations* 21, 1 (1968), 3–17. https://doi.org/10.1177/001872676802100101

[4] 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

[5] 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

[6] 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

[7] 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

[8] Cynthia Breazeal. 2005. Socially Intelligent Robots. *Interactions* 12, 2 (March 2005), 19–22. https://doi.org/10.1145/1052438.1052455

[9] Joseph Bugeja, Andreas Jacobsson, and Paul Davidsson. 2016. On Privacy and Security Challenges in Smart Connected Homes. In *2016 European Intelligence and Security Informatics Conference (EISIC)*. 172–175. https://doi.org/10.1109/EISIC.2016.044

[10] 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

[11] 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/10.1145/192161.192272

[12] 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

[13] Brian D. Corneil and James K. Elsley. 2005. Countermanding Eye-Head Gaze Shifts in Humans: Marching Orders Are Delivered to the Head First. *Journal of Neurophysiology* 94, 1 (2005), 883–895. https://doi.org/10.1152/jn.01171.2004

[14] Mick Donegan, Jeffrey D. Morris, Fulvio Corno, Isabella Signorile, Adriano Chió, Valentina Pasian, Alessandro Vignola, Margret Buchholz, and Eva Holmqvist. 2009. Understanding Users and Their Needs. *Universal Access in the Information Society* 8, 4 (Nov. 2009), 259–275. https://doi.org/10.1007/s10209-009-0148-1

[15] Heiko Drewes and Albrecht Schmidt. 2007. Interacting with the Computer Using Gaze Gestures. In *Human-Computer Interaction – INTERACT 2007 (Lecture Notes in Computer Science)*, Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa (Eds.). Springer Berlin Heidelberg, 475–488. https://doi.org/10.1007/978-3-540-74800-7_43

[16] 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

[17] 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

[18] Raspberry Pi Foundation. Accessed on 2018-09-16 18:43:55. About the Raspberry Pi.

[19] 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

[20] 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

[21] 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

[22] 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.

[23] Yanming Guo, Yu Liu, Ard Oerlemans, Songyang Lao, Song Wu, and Michael S. Lew. 2016. Deep Learning for Visual Understanding: A Review. *Neurocomputing* 187 (April 2016), 27–48. https://doi.org/10.1016/j.neucom.2015.09.116

[24] Dan Witzner Hansen, Henrik H. T. Skovsgaard, John Paulin Hansen, and Emilie Møllenbach. 2008. Noise Tolerant Selection by Gaze-Controlled Pan and Zoom in 3D. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications - ETRA '08*. ACM Press, Savannah, Georgia, 205. https://doi.org/10.1145/1344471.1344521

[25] Christian Heath, Jon Hindmarsh, and Paul Luff. 2010. *Video in Qualitative Research*. SAGE Publications Ltd, Los Angeles.

[26] 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

[27] Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and Listening with the Eyes: Gaze Signaling during Dyadic Interactions. *PLOS ONE* 10, 8 (Aug. 2015), e0136905. https://doi.org/10.1371/journal.pone.0136905

[28] 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

[29] 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

[30] Michael Katzenmaier, Rainer Stiefelhagen, 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

[31] 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

[32] 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

[33] 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

[34] Michael F. Land and Mary Hayhoe. 2001. In What Ways Do Eye Movements Contribute to Everyday Activities? *Vision research* 41, 25-26 (2001), 3559–3565. https://doi.org/10.1016/s0042-6989(01)00102-x

[35] Bo Li, Tara N. Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, Haşim Sak, Golan Pundak, Kean Chin, Khe Chai Sim, Ron J. Weiss, Kevin W. Wilson, Ehsan Variani, Chanwoo Kim, Olivier Siohan, Mitchel Weintraub, Erik McDermott, Richard Rose, and Matt Shannon. 2017. Acoustic Modeling for Google Home. In *Interspeech 2017*. ISCA, 399–403. https://doi.org/10.21437/Interspeech.2017-234

[36] Arduino LLC. [n. d.]. Arduino. https://www.arduino.cc/.

[37] 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

[38] 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

[39] Sven Meyer and Andry Rakotonirainy. 2003. A Survey of Research on Context-Aware Homes. In *Proceedings of the Australasian Information Security Workshop Conference on ACSW Frontiers 2003 - Volume 21 (ACSW Frontiers '03)*. Australian Computer Society, Inc., Darlinghurst, Australia, Australia, 159–168.

[40] 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

[41] 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

[42] Gerhard Sigurd Nielsen. 1964. *Studies in Self-Confrontation.* Munksgaard.

[43] Omron. [n. d.]. B5T-007001 Human Vision Components (HVC-P2) | OMRON - Americas. https://www.components.omron.com/mobile/hvc_p2.

[44] Jeff Pelz, Mary Hayhoe, and Russ Loeber. 2001. The Coordination of Eye, Head, and Hand Movements in a Natural Task. *Experimental Brain Research* 139, 3 (2001), 266–277. https://doi.org/10.1007/s002210100745

[45] 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. https://doi.org/10.1007/978-3-540-74997-4_25

[46] 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

[47] Alex Poole and Linden J. Ball. 2006. Eye Tracking in HCI and Usability Research. In *Encyclopedia of Human Computer Interaction*. IGI Global, 211–219. https://doi.org/10.4018/978-1-59140-562-7.ch034

[48] 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

[49] 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

[50] 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. https://doi.org/10.1007/11550617_21

[51] Charles Rich, Brett Ponsler, Aaron Holroyd, and Candace 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

[52] Federico Rossano. 2012. Gaze in Conversation. In *The Handbook of Conversation Analysis*, Jack Sidnell and Tanya Stivers (Eds.). John Wiley & Sons, Ltd, Chichester, UK, 308–329. https://doi.org/10.1002/9781118325001.ch15

[53] Federico Rossano, Penelope Brown, and Stephen C. Levinson. 2009. Gaze, Questioning, and Culture. In *Conversation Analysis*, Jack Sidnell (Ed.). Cambridge University Press, Cambridge, 187–249. https://doi.org/10.1017/CBO9780511635670.008

[54] Kerstin Ruhland, Sean Andrist, Jeremy Badler, Christopher Peters, Norman Badler, Michael Gleicher, Bilge Mutlu, and Rachel Mcdonnell. 2014. Look Me in the Eyes: A Survey of Eye and Gaze Animation for Virtual Agents and Artificial Systems. *Eurographics State-of-the-Art Report* (April 2014), 69–91. https://doi.org/10.2312/egst.20141036

[55] Kerstin Ruhland, Christopher E. Peters, Sean Andrist, Jeremy B. Badler, Norm Badler, Michael Gleicher, Bilge Mutlu, and Rachel 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

[56] Harvey Sacks. 1984. Notes on Methodology. In *Structures of Social Action: Studies in Conversation Analysis*, John Heritage and J. Maxwell Atkinson (Eds.). Cambridge University Press, Cambridge, 2–27.

[57] Harvey Sacks, manuel A. Schegloff, and Gail Jefferson. 1974. A Simplest Systematics for the Organization of Turn Taking for Conversation. *Language* 50 (1974), 696–735. https://doi.org/10.2307/412243

[58] 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

[59] Jeffrey S. Shell, Roel Vertegaal, Daniel Cheng, Alexander W. Skaburskis, Changuk Sohn, A. James Stewart, Omar Aoudeh, and Connor Dickie. 2004. ECSGlasses and EyePliances: Using Attention to Open Sociable Windows of Interaction. In *Proceedings of the Eye Tracking Research & Applications Symposium on Eye Tracking Research & Applications - ETRA'2004*. ACM Press, San Antonio, Texas, 93–100. https://doi.org/10.1145/968363.968384

[60] 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. https://doi.org/10.1145/964456.964458

[61] Sophie Stellmach, Sebastian Stober, Andreas Nürnberger, and Raimund Dachselt. 2011. Designing Gaze-Supported Multimodal Interactions for the Exploration of Large Image Collections. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications - NGCA '11*. ACM Press, Karlskrona, Sweden, 1–8. https://doi.org/10.1145/1983302.1983303

[62] Seeed Studio. Accessed on 2018-09-16 18:44:30. ReSpeaker Mic Array - Far-Field w/ 7 PDM Microphones - IoT - Seeed Studio. https://www.seeedstudio.com/ReSpeaker-Mic-Array-Farfield-w-7-PDM-Microphones-p-2719.html.

[63] 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

[64] Daniel Szafir 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

[65] 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

[66] Jake VanderPlas. 2016. Python Data Science Handbook. (2016), 548.

[67] 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. https://doi.org/10.1145/365024.365119

[68] Roel Vertegaal and Harro Vons. 2000. Effects of Gaze on Multiparty Mediated Communication. *Graphics Interface* (2000), 95–102.

[69] Shumin Zhai, Carlos Morimoto, and Steven Ihde. 1999. Manual and Gaze Input Cascaded (MAGIC) Pointing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems the CHI Is the Limit - CHI '99*. ACM Press, Pittsburgh, Pennsylvania, United States, 246–253. https://doi.org/10.1145/302979.303053