# You Talkin' to Me? A Practical Attention-Aware Embodied Agent

Rahul R. Divekar[1,2](✉), Jeffrey O. Kephart[2], Xiangyang Mou[1], Lisha Chen[1], and Hui Su[1,2]

[1] Rensselaer Polytechnic Institute, Troy, NY, USA
{divekr,moux4,chenl21}@rpi.edu
[2] IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
{kephart,huisuibmres}@us.ibm.com

**Abstract.** Most present-day voice-based assistants require that users utter a wake-up word to signify that they are addressing the assistant. While this may be acceptable for one-shot requests such as "Turn on the lights", it becomes tiresome when one is engaged in an extended interaction with such an assistant. To support the goal of developing low-complexity, low-cost alternatives to a wake-up word, we present the results of two studies in which users engage with an assistant that infers whether it is being addressed from the user's head orientation. In the first experiment, we collected informal user feedback regarding a relatively simple application of head orientation as a substitute for a wake-up word. We discuss that feedback and how it influenced the design of a second prototype assistant designed to correct many of the issues identified in the first experiment. The most promising insight was that users were willing to adapt to the interface, leading us to hypothesize that it would be beneficial to provide visual feedback about the assistant's belief about the user's attentional state. In a second experiment conducted using the improved assistant, we collected more formal user feedback on likability and usability and used it to establish that, with high confidence, head orientation combined with visual feedback is preferable to the traditional wake-up word approach. We describe the visual feedback mechanisms and quantify their usefulness in the second experiment.

**Keywords:** Multimodal interaction · User interaction and experience · Natural language interfaces

## 1 Introduction

Voice-activated assistants typically require users to utter a wake-up word (such as "Hey Google" or "Alexa") to indicate that they are addressing the assistant. Most interactions with these devices consist of one or two commands (Bentley et al. 2018). The use of a wake-up word is generally acceptable for short atomic requests such as "Turn on the lights" or "Set a timer for 5 min", which require no

D. Lamas et al. (Eds.): INTERACT 2019, LNCS 11748, pp. 760–780, 2019.
https://doi.org/10.1007/978-3-030-29387-1_44

deliberation or discussion. However, assistants designed to assist humans with higher-level cognitive tasks are beginning to emerge, as demonstrated by Kephart et al. (2018) and Farrell et al. (2016). While by and large these agents still process one-shot requests, those requests tend to be issued in rapid succession, and may often be interleaved with discussions with other humans. Thus, from the users' perspective, communication with the agent is part of a broader conversation that they may be having with another human, making the overall dialogue multi-round. In this case, prefacing each command or request with a wake-up word is tedious and unnatural.

For assistants that support multi-round conversations with only one human, one previously-explored solution has been for the assistant to extend its period of attentiveness for a few seconds after its most recent response. However, this is not viable when multiple people are collaborating with one another and with the agent, as it becomes difficult for the assistant to distinguish requests from conversation among human collaborators. A confused agent may interrupt such side conversations with inappropriate and unwelcome chatter, such as "I'm sorry, Dave. I'm afraid I can't do that."

We seek to develop an alternative to the wake-up word that is sufficiently accurate without being unduly complex or expensive. As a starting point, we note that an approach that has been explored by the HCI/HRI community for robotic assistants is based upon real-time eye-gaze measurements (Wang and Ji 2017). A drawback of this approach is that it requires careful calibration, and moreover distance scales appropriate for multi-user scenarios require expensive Pan-Tilt-Zoom (PTZ) cameras. Fortunately, we note that other prior work has established that head orientation can adequately substitute for eye gaze in the context of gaming (Da Silva et al. 2008) and meeting analysis (Stiefelhagen and Zhu 2002)—suggesting that it may be acceptable in our scenario as well.

This paper describes our effort to ascertain whether the relatively inexpensive approach of using real-time head pose measurements as a proxy for user attention is a suitable alternative to using a wake-up word. After a review of the relevant literature in Sect. 2, Sect. 3 describes a first experiment in which we implemented a first prototype assistant that used a simple heuristic to determine whether the user was addressing the assistant. The assistant, which was based upon a previously-developed astrophysics assistant (Kephart et al. 2018) that helps users explore data about exoplanets (planets that orbit distant stars), was represented as an avatar displayed on a large TV screen, as depicted in Fig. 1. Informal feedback from this study involving several novice users indicated that the assistant was not sufficiently usable. Results from the pilot study also provided insights into how the assistant might be improved. Section 4 first describes how we translated lessons learned from the pilot study into technical enhancements to the prototype. Then, we report results from a controlled user study that we conducted with 8 university students (none of whom had participated in the pilot study). These results establish with reasonable confidence that the second (enhanced) prototype assistant is more usable and likable than a version of the assistant that is otherwise identical except that it uses a wake-up word.

We conclude in Sect. 5 with a summary and some thoughts about possible future extensions of this work.

## 2 Related Work

Our contribution is multi-disciplinary and hence the related work is discussed in three parts: addressee detection, HCI of Multi-modal Conversational UI and gaze detection.

### 2.1 Addressee Detection

Most prior work on addressee detection and turn-taking focus on using combinations of visual, acoustical and textual features from human participants in the interaction. Efforts have been strong towards fusing these multiple modalities, identifying importance of each modality and applying machine/deep learning algorithms at different stages e.g. modality fusion, attention detection, etc. as described further.

Ravuri and Stolcke (2015) have explored addressee detection but strictly with lexical and/or speech based modality. More recently, Norouzian et al. (2019) have explored sophisticated models for addressee detection based purely on acoustical cues. Frampton et al. (2009) have combined gaze and linguistic features to identify the addressee in conversations among groups of humans that involve ambiguous references like “you”.

Bakx et al. (2003) and Van Turnhout et al. (2005) have contributed to the addressee recognition problem by collecting data in a Wizard-of-Oz setting in which human subjects spoke to a human partner and a human-driven kiosk that posed as an intelligent machine. They conducted statistical analyses to relate manually annotated eye-gaze data to characteristics of the conversation and trained and evaluated a Naive Bayes classifier. They found that looking away from the machine strongly signified that the addressee was the human partner, but looking at the machine only weakly signified that the human was addressing the machine.

Baba et al. (2012) and Nakano et al. (2013) have experimented, analyzed and implemented conversations with the goal of addressee detection in human-human-agent settings. They find that the tone of voice while talking to agent is higher and a speech+head orientation signal in their SVM model has given them good results. The literature thus encourages our thoughts that head orientation is an important signal. Akhtiamov et al. (2017), Shriberg et al. (2013) have done similar work in addressee detection based on speech and textual features. Le Minh et al. (2018) have explored addressee detection using gaze in data that contained images and text using deep learning approaches. Tsai et al. (2015) have studied the effect of various multi-modal features in addressee detection in human-human-computer interaction and have concluded that voice based features are more important that visual features due to headpose being affected by situational attractors claiming that headpose, by itself, is not enough.

Akhtiamov and Palkov (2018) echo similar findings as their addressee detection accuracy is highest with acoustical+textual+visual features.

Further, in the HRI field, Katzenmaier (2004) have explored in depth how to identify the addressee in a human-human-robot conversation. As do we, they find that users may look at the agent even when they are actually addressing another human. In their parlance, people usually look at the subject to whom they are speaking except when there is another "situational attractor", which they define as "objects or situations in the environment that attract people's eye gaze when they are talking to each other". Work in this field seemed to be the theme of research in the HRI community in the early 2000's. Some examples of which are Sheikhi and Odobez (2015), Mutlu et al. (2012), Gu and Badler (2006). Their contributions have determined Visual Focus of Attention (VFOA) using eye-gaze and/or head pose, identifying the addressee based on combinations of VFOA and context using several models in an effort to make robots more humanistic.

Attention detection is closely related to turn taking in multi-party conversations. Andrist et al. (2016) have summarized the turn taking problem in HCI and further motivated this problem. Kendon (1967) have shed light on the non-verbal turn-yielding cues in human-human behavior such as body movement or gaze direction, while Gravano and Hirschberg (2009) have discussed turn-yielding cues such as speaking rate, intonation, etc. giving the research community heuristics. Bohus and Horvitz (2011) have used a kiosk scenario with a humanistic face to study and look more broadly in the turn-taking domain in a game-like context. They used hand-crafted turn-taking policies to enable their prototype and emphasize how a bad turn-taking system is a conversation breaker.

### 2.2 Interaction Design

Interaction mechanisms for the purpose of attention detection and turn taking have previously been explored in the HRI community. van Schendel and Cuijpers (2015) have demonstrated the positive effect of robots expressing turn-yielding cues such as stop arms, turn head, flash eyes. We encourage enthusiastic readers to see Admoni and Scassellati (2017) who have in detail reviewed the state-of-the-art in social eye-gaze in human-robot behavior and discussed its role in usability, conversation, attention and turn-taking.

Visual cues have been explored in the chatbot area and are seen deployed in commercial applications like Alexa Echo, Google Home, etc. They appear as a combination of colored and patterned flickering of lights to indicate when the bot is listening, talking, thinking, etc. The importance of visual feedback in this context has been motivated in a talk by the VP of Alexa and Echo (ZDNet 2018) and documentation for Alexa's attentional system (Amazon 2019) can be seen on their webpages. However, we have not found an academic discussion of the same.

A common focus appears to emerge from literature - improving accuracy of attention detection by fusing several conversational and visual cues. However, building a system to express and interpret all of the cues in real time is a hard problem in research and in computation. We want to build a system that is not too expensive to deploy, sufficiently accurate and easy to use. Hence we start with just using headpose as an estimation of users' gaze which activates the agent and focus on the user experience of the conversation. In doing this we have found it relatively easy to deploy interaction mechanisms that increase the usability of a system. We show that such an interaction is still preferred over the currently established paradigm of using a wake word in non-robotic conversational UI. We can only imagine that with future advances in computer science, considering other cues will be cheaper and lead to better accuracy which will even further increase the usability of our system.

### 2.3 Gaze Detection

Head pose estimation is one of the most popular topics in Computer Vision area. There are two major approaches applied to this specific task: the landmark-based approach and end-to-end approach.

Typically, the landmark-based approach entails three steps: first, find faces in a RGB image; then detect the facial landmarks as features i.e. contours of eyes, nose, mouth and face; and finally predict head orientation based on the landmarks. Researchers are pushing forward the frontiers for each step. For the first step, Lin and Tsai (2012) and Ranganatha and Gowramma (2017) both have focused on face detection and face tracking problems. They have use Haar-like features such as face edges and corners to find all the faces in a frame and apply different tracking algorithms for the faces in the coming frames to increase computation efficiency. Lin and Tsai (2012) used Kanade-Lucas-Tomasi (KLT) tracking algorithm and Ranganatha and Gowramma (2017) have used a combination of Continuously Adaptive Mean Shift (CAMShift) and Kalman filter. For the second step, Wu and Ji (2017) have conducted an elaborate survey about face landmark detection, grouping algorithms into 3 major categories according to the ways the facial appearance and shape information are utilized and, compared their performances. For the third step, Dementhon and Davis (1995) have described a method of pose estimation using a base of Orthography and Scaling Approximation (POS). A POS system finds translation and rotation matrices by solving a linear system. Dementhon and Davis (1995) loops over this procedure for a better pose estimation with faster computation and implementation speed and, because of its merits it eventually becomes a part of our approach. In fact, instead of taking these three steps completely apart, researchers are also interested in integrating them together to improve efficiency and performance. A unified framework has been proposed in Wu et al. (2017) with landmark detection, head pose estimation and facial deformation analysis taken into account simultaneously. It is shown to perform more robustly in cases where occlusion becomes a major issue for face detection. It is an intermediate method between independent methods and end-to-end methods. However, it is not adopted in our approach, because occlusion isn't an issue in our scenario and therefore its complexity doesn't contribute much to accomplishing our goal.

The success of deep learning and end-to-end model in various tasks and problems in Computer Vision area are encouraging researchers to utilize it in head pose estimation. Ahn et al. (2015) have proposed a deep neural network which took low-resolution RGB images for head pose estimation. They use regression in their architecture with the aid of GPUs. The model was able to provide continuous pose results in real time. De and Kautz (2017) have leveraged the merits of Recurrent Neural Networks, borrowing the information from preceding frames and bringing extra bonuses to applications in real-time and in video head pose estimation. To summarize, the overall advantage of an end-to-end model is that it does not rely on any explicit face features or independent face feature extractors, and hence outperforms landmark-based approaches in the cases where facial features can not be detected due to occlusion.

In addition to the effort of improving the model itself, researchers are experimenting with improvements to the model input for head pose estimation in specific scenarios. Borghi et al. (2017) and Venturelli et al. (2017) respectively have created a Generative Adversarial Network and a Convolutional Neural Network to predict head pose from depth images. Again, an end-to-end model was adopted, but instead of using RGB images, the authors took depth images as the input of the model. Its main advantage over the RGB-based approaches is that the depth sensors are not affected by the environmental illumination changes, and therefore the model can be more adaptive. However, end-to-end model is computationally intensive in nature and it runs counter to our goal of being low-cost and low-overhead.

## 3 Experiment 1

In this section, we begin with a technical description of our first prototype assistant. Then, we report on results and lessons learned from a pilot study of this assistant involving 10 novice users.

### 3.1 Technical Details

Figure 1 shows a typical setup in which one or more users sit across from a large display. The agent (embodied on the screen) uses the display as a canvas for showing requested information to the users. The existing exoplanets prototype already contained ASR and intent recognition capabilities made available on a pub-sub channel. We added to the existing prototype a head-orientation application that processed video signals from a webcam to infer the user's head orientation, plus capability that combined speech transcription with head orientation to determine whether the agent should assume that an utterance was directed to it.

Capturing voice and facial image data has implications of privacy. For voice transcription, we use a commercial module and rely on their privacy guidelines. Our scenario did not require users to mention any personally identifiable information when the microphones were on.

For face data which is much more sensitive, we used a designated machine to process it. None of this data was uploaded on to the internet. The machine received image data from the camera attached to it, computed the headpose coordinates and only sent these coordinates to the rest of the system.

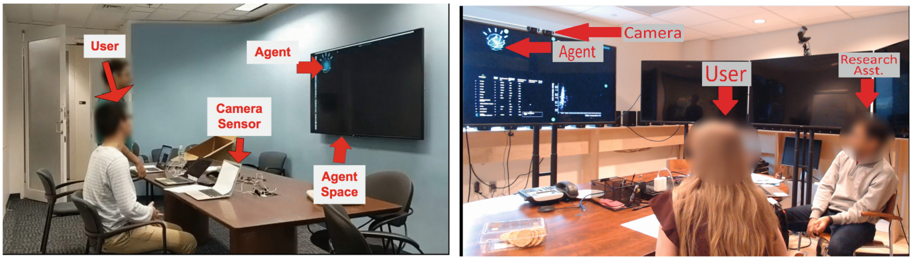
To build a low-cost, low-overhead head pose estimation system for real-time inference which, at the same time, can be adopted elsewhere easily with the smallest limitation, we decided to follow the landmark-based approach described in literature review. This is because end-to-end approaches are computationally expensive and need one or several powerful GPUs for real-time performance. Considering our indoor environment, the landmark-based approaches are already good enough to handle the cases of our interest. Enthusiastic readers are encouraged to see Zhao et al. (2018) where more details of this technique are elucidated. These calculations are performed at approximately 20 Hz and published on a head-orientation pub-sub channel.

An attention inference module subscribes to the speech transcription and head-orientation channels. It checks each head-orientation event to determine whether or not the head orientation falls within a defined Region of Interest (ROI) e.g. TV Screen in Fig. 1. If the current state is *non-attentive* and the head orientation falls within the ROI, a new *attention* window is started, which ends as soon as the head orientation falls outside the ROI or the user's face is no longer detected. The ROI can be configured as needed. Given the relatively low precision of headpose system which was traded-off for deploy-ability and preference to use markerless non-intrusive technologies, we selected the entire screen as the ROI. Results in further section will show that this worked. Similarly, a *transcription* window is started when speech is received and ends on detection of a pause. For each utterance, the overlap between the *transcription* and *attention* windows is computed and thresholded to determine whether or not the utterance was addressed to the assistant. If the transcription contains the attention word (e.g. Watson/Celia), it is assumed to be intended for the assistant regardless of head orientation.

### 3.2 Pilot Study

We conducted a pilot study designed to get periodic early feedback on our system's usability and the nature of any shortcomings it might exhibit. The feedback was based upon direct observation of user interactions with the system captured on video, as well as user responses to written survey questions and informal interviews conducted immediately following the each user's interaction with the system.

**Demographic.** A total of 10 people were recruited to interact with the system. All of our subjects had some background in technology which ranged from undergraduate/graduate university students to experienced research scientists. They all were aware of commercial conversational agents (e.g. Siri, Alexa, etc.) although some used them more infrequently than the others. The population contained native and non-native English speakers. 4 subjects who work in the same lab but had never seen this project before were also recruited as they



(a) Environment 1

(b) Environment 2

**Fig. 1.** Assistant and its environment

represent the demographic that is most acquainted with such type of a conversational agent and would give a critical assessment of the system. This group positively expressed that they would find such an attention awareness module in their demos/projects useful.

**Study Details.** Each subject was paired up with a partner who engaged the subject in a discussion with the AI assistant about exoplanets. Our aim in picking a specialized topic like exoplanets for conversation was so that previous experience and familiarity with the content of the conversation would not severely bias the findings. Due to inexperience with the domain, an instruction sheet encouraged the subjects to explore exoplanets and detailed the commands that the AI assistant would recognize. The task for the AI agent was to separate out the side conversation from commands given to it. The subjects mostly drove the system, except in rare circumstances. We used a combination of setups shown in Figs. 1a and b to conduct pilot testing. Both setups had the participants situated in a conference-room style seating. They both sat in chairs that faced the screen (agent space). A major difference between the two setups was location of the camera that helps determine headpose. Figure 1a used the built-in camera on a laptop to track the head orientation. This was a more fluid design and likely to be more typical of real-life meetings in which the head orientation of participants could be tracked using their own personal laptops. However, in this setup, calibration had to be done very carefully and slight changes in the laptop position would not work well. In our case, calibration would mean position of the camera, what pixels (as inferred by the headpose system) constituted as within the Region of Interest (ROI), what constituted as agent space, etc. Additionally, since the range of the camera vision is limited, subjects who were taller or shorter than the subjects for which this setup was calibrated had a tough time interacting with the system. In contrast, Fig. 1b has a web camera mounted on the display. This setup responded more reliably to variations in the physical aspects of the subjects given the field of vision of the camera.

The users were encouraged to think out loud. Their feedback was taken through a questionnaire and an informal interview. Since there existed variations in each iteration of the study, we used this study to find anecdotal patterns in any difficulties that arose in the course of the interaction. A sample hypothetical interaction can be seen in Table 1. Italicized parts are utterances that the system recognizes as commands and responds to. Determining whether to respond to them or not in these cases is the challenge for the system here.

**Table 1.** Sample interaction (H1 = Human1, H2 = Human2, AI: AI agent)

| Turn | Utterance |
|---|---|
| H1 to H2 | Let's start visualizing exoplanets by just plotting them? |
| H1 to AI | *Show me a plot of exoplanets* |
| AI | Okay (display of plot) |
| H2 to H1 | That plot doesn't tell me much. I wonder if the temperature of stars start to lower as they die down |
| H1 to H2 | Perhaps we can ask the system to *plot temperature against age* |
| H1 to AI | *Plot temperature against age* |
| AI | Done (changed axes) |
| H2 to H1 | Now that looks interesting. Looks like a huge cluster. |
| H1 to AI | *Change the x-axis to a log scale* |
| AI | Done (changed axes) |
| H2 to H1 | What are we looking at here? What is that outlier dot? You can ask the system to *tell us more about that star* |
| H1 to AI | *Tell me more about this star* |
| AI | Sure (Shows a table with more details) |

**Feedback and Discussion.** Two classes of problems emerged from the pilot study. The first arose from failures or inaccuracies in head-orientation measurements. The head-orientation system works based on facial landmarks which assumes that the full face can be seen by the camera. Head-orientation measurements failed when the system failed to detect a face, which occurred when the user turned their head beyond the angle of recognition or covered a portion of their face (e.g. while stroking their chin). It also failed when the user constantly moved their head too quickly presumably in confusion trying to get the systems attention. Moreover, mis-calibration resulting from individual differences in height or position sometimes caused inaccurate estimates of head orientation.

The second class of problem resulted from head orientation being an imperfect proxy for attention. We found that, while users looking at the display was a good first-order heuristic for determining whether an utterance was addressed to the system, there were several conditions under which intended commands were ignored, including the user reading from a page, looked away from the system trying to recollect a command or word, or looking at a human partner to seek help with completing a command. Moreover, there were situations where the system falsely interpreted an utterance as a command, such as when the user looked at a plot on the display while discussing it with their human partner leading to a repeated and displeasing "Sorry, I can't do that" response from the agent.

A more positive finding was that users were willing to adapt their behavior to accommodate deficiencies of the assistant; for example, when the system committed transcription errors they began to enunciate commands more clearly. Baba et al. (2012) have also found that their users spoke more slowly and loudly when talking to the agent as compared to talking to their partner giving us a hint that it can possibly be attributed to humans' willingness to accommodate for the agent. This prompted us to modify the UI to provide simple visual feedback regarding the assistant's mental model of the user's attention. Our hypothesis is that given enough feedback, users would be willing to slightly adapt their behavior and that would lead to a more pleasant/usable experience. The next sections focus on that.

# 4 Experiment 2

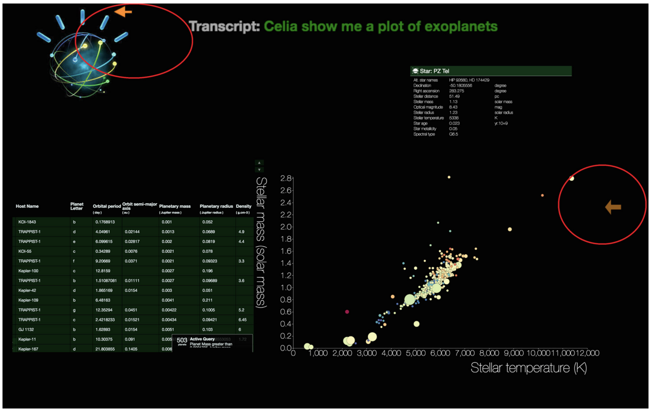
In this section, we first discuss how we translated lessons learned from the pilot study of Sect. 3 into an enhanced prototype assistant. Then, we present the results of a controlled user study of 8 novice users, none of whom had participated in the pilot study, from which we establish with reasonable certainty that the second prototype is more preferable than a comparable assistant that uses wake-up words to determine whether it should respond to the user.

## 4.1 Technical Details

This section discusses the observations and our technical approach to incorporating them into an improved version of the agent.

Some calibration issues were addressed by placing the camera further away from the participants i.e. mounted on the TV display itself. Others were alleviated by picking the whole TV display as a region of interest instead of the logo of the AI avatar on it. This helped because a low computing cost headpose tracking system is not designed to be as precise.

It was clear from the pilot study that the assistant needed to provide the user with some sort of feedback; the question was what *sort* of feedback. Ruhland et al. (2015) have summarized many benefits of multimodal output generation by assistants. However, these generations need an animated humanoid avatars which is not our case. Animated humanoid avatars are more expensive to build/run which may be justified if the avatars want to exhibit any social intelligence e.g. facial expressions which is not our focus. We want our agent to simplify interactions with itself in group discussions. For our embodiment of AI agent, we settled upon two approaches. First, we provided feedback on the assistant's understanding of the user's head orientation. If the assistant believes the user is looking in a direction other than the display, it uses an inward-pointing orange arrow at the edge closest to the user's inferred gaze to indicate the direction in



**Transcript:** Celia show me a plot of exoplanets

Star: PZ Tel

| | | |
|---|---|---|
| Alt. star names | HIP 92680, HD 174429 | |
| Declination | -50.1805556 | degree |
| Right ascension | 283.275 | degree |
| Stellar distance | 51.49 | pc |
| Stellar mass | 1.13 | solar mass |
| Optical magnitude | 8.43 | mag |
| Stellar radius | 1.23 | solar radius |
| Stellar temperature | 5338 | K |
| Star age | 0.023 | yr.10+9 |
| Star metallicity | 0.05 | |
| Spectral type | G6.5 | |

| Host Name | Planet Letter | Orbital period (day) | Orbit semi-major axis (au) | Planetary mass (Jupiter mass) | Planetary radius (Jupiter radius) | Density (g.cm-3) |
|---|---|---|---|---|---|---|
| KOI-1843 | b | 0.1768913 | | 0.001 | 0.052 | |
| TRAPPIST-1 | d | 4.04961 | 0.02144 | 0.0013 | 0.0689 | 4.9 |
| TRAPPIST-1 | e | 6.099615 | 0.02817 | 0.002 | 0.0819 | 4.4 |
| KOI-55 | c | 0.34289 | 0.0076 | 0.0021 | 0.078 | |
| TRAPPIST-1 | f | 9.20669 | 0.0371 | 0.0021 | 0.09323 | 3.3 |
| Kepler-100 | c | 12.8159 | | 0.0027 | 0.196 | |
| TRAPPIST-1 | b | 1.51087081 | 0.01111 | 0.0027 | 0.09689 | 3.6 |
| Kepler-42 | d | 1.865169 | 0.0154 | 0.003 | 0.051 | |
| Kepler-109 | b | 6.48163 | | 0.0041 | 0.211 | |
| TRAPPIST-1 | g | 12.35294 | 0.0451 | 0.00422 | 0.1005 | 5.2 |
| TRAPPIST-1 | c | 2.4218233 | 0.01521 | 0.00434 | 0.09421 | 6.45 |
| GJ 1132 | b | 1.62893 | 0.0154 | 0.0051 | 0.103 | 6 |
| Kepler-11 | b | 10.30375 | 0.091 | 0.005[illegible] | [illegible]8553053 | 1.72 |
| Kepler-167 | d | 21.803855 | 0.1405 | 0.006[illegible] | [illegible] | [illegible] |

503 planets **Active Query**
Planet Mass greater than

**Fig. 2.** Screenshot of the display showing inward pointing arrows and colored transcription (Color figure online)

which the user should move their gaze in order for the assistant to understand that the user is paying attention to it (Fig. 2). If the assistant cannot tell where the user is looking, it shows red dots in the center on all edges. When the assistant believes the user is paying attention to it, it shows green dots in the center on all four edges of the display analogous to a reciprocal gaze performed by the agent thereby letting the user know that it is listening.

Xu et al. (2016) have shown how humans were more coordinated and synchronized in their speech + gaze behavior when they successfully established mutual gaze with a robot. We think that users in our interaction paradigm will follow a similar pattern and the green dots (a proxy for system's gaze) will encourage the user to look at the system more while talking to it despite there being a "situational attractor" (Katzenmaier 2004) such as a human partner or the instruction sheet. We do not specifically aim to prove it in this paper, but use it to base our assumption that the feedback from the system makes it easy for the user to have a synchronized behaviour resulting in a usable system which is the focus.

As a second feedback element, we displayed a color-coded text transcription, such that utterances from which an actionable command was extracted while the user was looking at the display were colored green, as seen in Fig. 2. White represented "no attention", while red text represented "attention but no actionable command". This helped us eliminate unwelcome and long "Sorry I don't understand" response from the agent which takes seconds and substituted it with red text that takes just a fraction of a second and conveys the same meaning.

The effectiveness of these feedback elements is discussed through a controlled study described in the next section.

### 4.2 Controlled Study

**Demographic.** A total of 8 university students participated in the study. All of them were familiar with conversational chatbots and had interacted with them in various degrees through their phones and home devices. They had a technology background in the sense they were enrolled in STEM courses (mostly Computer Science) and were in varied levels in their formal education. We mention this as we see discussions on the effect of familiarity on the style and ease of interaction by Sciuto et al. (2018) and, by extension, perhaps the know-how of their internal working mechanism also affects it. We think that this would have minimal bias in our study because of the uniformity in the subject pool. Specifically, all subjects were at least slightly familiar with the concept of chatbots but not with a system such as ours.

**Experiment Design.** Each subject was paired with a research assistant who played the role of a conversation partner. A conference room style setting was used with a large display on which a camera was mounted, as seen in Fig. 1b The subjects were given a sheet listing the commands, and given an opportunity to study it for a few minutes prior to their interaction with the assistant. Once the interaction began, it consisted of interleaved conversation with the research assistant (who would explain and/or suggest specific commands) and the automated assistant, to whom the subject would issue commands.

Evaluating the quality of a conversational interface is a complicated task from the conversation intelligence perspective. Radziwill and Benton (2017) have proposed a good approach to evaluating chatbots, involving evaluation categories that include performance, humanity, affect, understanding social cues, etc. However, our goal here was not to evaluate the conversation or capability of the chatbot as a whole. Instead, we wanted to measure the usability of the headpose-based assistant, tease out the factors that contribute most greatly to its usability, and compare its usability to that of an otherwise identical assistant that requires a wake-up word.

In order to do this, we created two variants of the exoplanets assistant that were nearly identical, with the following exceptions:

1. *Condition A*. Users were required to use a wake-up word to signify that they were addressing the assistant.
2. *Condition B*. Users merely needed to look at the display to signify that they were addressing the assistant. The display included the visual feedback mechanisms (colored dots, live transcript and arrows) described in Sect. 4.1.

Each user interacted with both Condition A and Condition B. In order to reduce any bias that might result from the order in which they were exposed to these variants, half of the population were shown Condition A first while the other half were shown Condition B first. Following the interaction with each variant, we asked users questions from Table 2 and followed up with an interview.

Before the interaction began, the users were introduced to the system and were instructed about how they could interact with it. The interaction itself was moderated, such that the research assistant would help the users understand exoplanets by working their way through the commands in the sheet and understanding the output. The subjects were encouraged to think out loud. A typical dialogue looked similar to the hypothetical dialogue in Table 1.

**Results and Discussion.** In this section, we detail our findings from the experiment on the second prototype assistant and evaluate its usability and likability relative to that of an assistant that is identical in every aspect except that it uses a wake-up word.

Subjects typically spent about 20–30 min in the room, including time spent on logistics and explanations. All users combined, we recorded 67.53 min of total interaction with 31.27 min of Condition A (avg 3.9 ± 1.3 min) and 36.26 min of Condition B (avg. 4.53 ± 1.1 min). The time spent was a decision of the participant and research assistant, based on the number of types of commands issued and whether the participants felt they were ready to evaluate the system or not. This statistic is noted to give the readers an idea of how long a conversation lasted and does not imply likability or usability; these issues are discussed in later sections.

Table 2 lists seven questions that were addressed to the subjects. Q1 was addressed to the subjects after they had experienced both conditions. The other

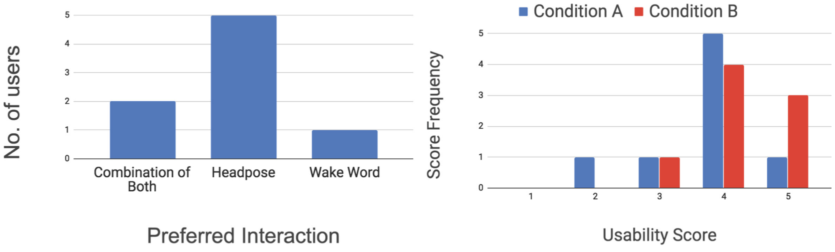
**Table 2.** Questions posed to users

| Question number | Question | Answer format |
|---|---|---|
| Q1 | Do you like to interact with the headpose more or without? (or similar) | Semi-structured interview |
| Q2 | How would you rate your overall experience? | Likert Scale of *Very Unusable - Very Usable* |
| Q3 | How easy was it to get [the bot] to know you are asking her to do something? | Likert Scale of *Very Difficult - Very Easy* |
| Q4 | How attentive was [the bot] to you? | Likert Scale of *Very Unattentive - Very Attentive* |
| Q5 | How helpful was it to see the transcription of what you were saying to the AI Agent? | Likert Scale of *Very Unhelpful - Very Helpful* |
| Q6 (Only Condition B) | How helpful was the green dot in giving you feedback about AI Agent's attentiveness? | Likert Scale of *Very Unhelpful - Very Helpful* |
| Q7 (Only Condition B) | How helpful were the arrows in knowing where to look to get AI Agent's attention? | Likert Scale of *Very Unusable - Very Usable* |

questions were addressed to subjects after they experienced each variant (questions Q6 and Q7 pertained to Condition B only). The third column of the table describes how answers were elicited. For those questions whose answers were numbers on a Likert scale, the possible answers were integers ranging from 1 (least favorable) to 5 (most favorable). The remainder of this subsection presents a comparative analysis of the usability and likability of the interaction under Conditions A and B, based upon an analysis of the answers to the questions in Table 2.

**Likability.** To assess likability, we explored question Q1 by conducting a semi-structured interview with the subjects right after they had experienced both conditions and given written feedback. We categorized their answers into "Head-pose" and "Wake Word" systems when their opinion strongly swayed in one or the other direction using thematic analysis of the comments. As is evident from Fig. 3a most users preferred the headpose-based system to the one requiring the wake-up word. We ran a Fisher's exact test on the preference indicated by the users towards the wake-up word based vs. headpose based systems. We chose this test because it is applicable in situations with small sample sizes, for the purpose of examining the significance of association between two kinds of classification. We used a 2 × 2 matrix with rows (Headpose based system, wakeup word based system) and columns (preferred, not preferred). Our findings that the head pose system is preferred were significant at $p < 0.1$ ($p = 0.08$).

Subjects who indicated that they would like a system that understood a combination of both wakeword and headpose based attentions expressed that there might be a case where they would not be able to directly look at the agent and would rather address it verbally. They were excluded from the test as their opinion did not strongly favor one or the other. It is worth mentioning that the Condition B version was able to do so and the users were not stopped from using the wakeword in their Condition B interactions. The subjects who indicated that they liked the wakeword system expressed concerns such as what would happen if the agent accidentally thought it needed to take action and how it would be more unwelcome than not taking any action as, the equivalent of "undo" does not typically exist in chatbots.



(a) Q1 - User Preferences

(b) Q2 - Usability of the System

**Fig. 3.** Preference and usability

**Usability and Perceived Discernment.** Question Q2 was aimed at assessing usability. Figure 3b shows the histogram of scores from the users on a Likert scale of Very Unusable (1) to Very Usable (5). We see that the new (Condition B) system is usable, which is our goal. On average, Condition B has better usability scores than Condition A: 4.25 ± 0.71 (Condition B) vs. 3.75 ± 0.89 (Condition A). However, applying a Wilcoxon-Mann-Whitney test to these results yields a p-value of 0.15, which is not quite enough to claim that the apparent usability advantages of the head-pose system are statistically significant.

To assess the assistant's perceived discernment—that is, the extent to which users perceived that the assistant correctly understood when it was and was not being addressed, and its attentiveness—we asked Questions Q3 and Q4 (see Table 2 for definitions and Likert scales). Likert scores for Conditions A and B were comparable in both cases: "somewhat easy" for Q3 (3.75 ± 1.04 for Condition B vs. 3.6 ± 0.74 for Condition A) and "attentive" for Q4 (4.25 ± 0.89 for Condition B vs. 4.13 ± 0.83 for Condition A). In other words, the perceived discernment of the two variants was essentially the same, and adequate.

**Usability and Likability Factors.** Here we analyze a variety of factors that contributed to the usability and likability of the headpose-based system.

**Color-coded Transcript.** For traditional HCI, users see feedback on their own input i.e. through text appearing as they type or the cursor responding as they move their mouse. Such feedback can also convey that the system is not frozen. However, standard feedback mechanisms do not exist for current voice-based systems. We believe that, for voice-based assistants in general, showing color-coded transcription would help users understand what the agent thought it heard (if anything), and thereby constitute a useful form of feedback. Figure 4, which summarizes the responses to Q5, supports such a belief in our case: the helpfulness score is 4.5 ± 0.53 for Condition B vs. 4.38 ± 0.74 for Condition A.

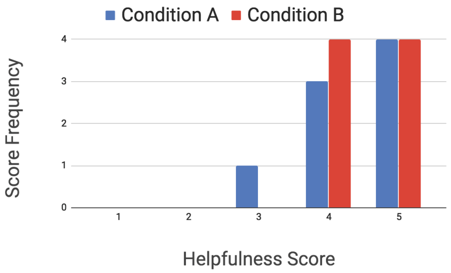
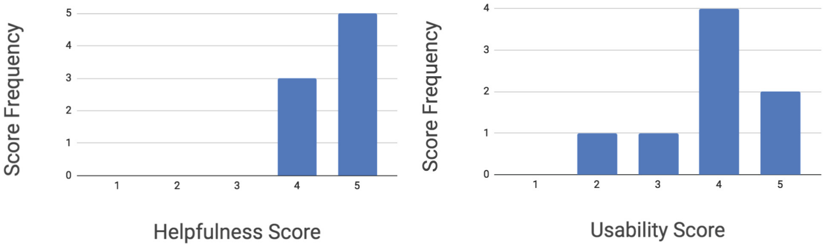


**Fig. 4.** Q5 - Helpfulness of displayed transcript

We note that displaying transcription has been shown to be helpful in the context of foreign language learning based upon conversation with AI agents (Divekar et al. 2018). However, their transcripts were not color-coded. To elicit whether the color-coded nature of the transcript was helpful, we asked another question—"Did you know when your utterance was recognized as a command vs. when it was not? How?". Using thematic analysis of responses, 6 of 7 (all except for one case of illegible data) could be strongly attributed to the color-coded nature of the transcript. Thus, we see that the color-coded nature of the transcript was noticed and the meaning it carried was well understood.

**Visual Attention Feedback.** In the Pilot Study section (Sect. 3.2), we theorized that it would be beneficial to provide visual feedback of the agent's understanding of the user's attentional state, and as described in Sect. 4.1 we added green dots and arrows for this purpose. In order to assess the helpfulness of these two feedback mechanisms, we asked the users Q6 and Q7 after they experienced Condition B. Figure 5 shows plots of the user's ratings on question Q6 and Q7 of helpfulness on a Likert scale of 1–5. As seen, the users found the green dot Very Helpful (avg. 4.63 ± 0.52) and the arrows Helpful (avg. 3.88 ± 0.99).



(a) Q6 - Helpfulness of green dot (b) Q7 - Helpfulness of arrows

**Fig. 5.** Helpfulness of visual feedback of attention

**ROI Selection.** We chose headpose estimation as a cheaper, more easily deployable, and less intrusive alternative to eye gaze estimation, but of course there is a cost: it is inherently less precise. Since we were unable to reliably detect whether users were looking at the avatar or not, we chose the entire display as the Region of Interest (ROI). Anecdotally, we noticed that users coordinated their speech and gaze, and waited for the green dot when they wanted to issue commands to the system. There were some instances when the agent mistook a human-human conversation as a command and interrupted out of turn. For example, there are several graphs and other objects of interest on the display that the user might sometimes want to look at while talking to their human partner. Such instances were rare and didn't seem to affect the usability/likability. We anticipate that advances in headpose recognition systems will result in improved accuracy, enabling the ROI to be reduced in area, which may further improve the likability of headposed-based assistants beyond what we have measured here.

**Natural, Learnable Interaction.** To gauge where people looked while giving commands, we manually annotated videos of 7 users[1] under Condition A, as

[1] Video data were missing for one subject.

this was the more natural case in which the user's head pose has no impact on the system's behavior. Annotations included the times at which each utterance started and ended and the times during which the subject was looking at the display. We found that users looked at the assistant's embodiment (the display) anyway, suggesting that this is indeed a natural interaction paradigm.

To help quantify this phenomenon, the overlap between their speech command and the time during which their headpose intersected with the display is shown in Fig. 6. Table 3 shows the percentage overlap between the users' speech and their head gaze oriented towards the display, in time, when intended to issue commands to the agent. Column 1 (Overall) shows the average percent intersection of all users. Column 2 and Column 3 show the average percent intersection of users who were exposed to Condition A first and Condition B first, respectively. Column 3 and Column 4 show the average percent overlap for long commands and short commands. Long commands are those which took more than 4 seconds to finish. This would happen in cases e.g. when the user would forget the command midway and would have to consult the command list for help. We observed that the overlap percent was significantly greater for short commands than it was for long commands, suggesting that utterance length might be a useful factor to include in follow-up experiments.

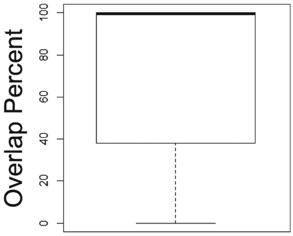


**Fig. 6.** Percentage overlap across users in Condition A

**Table 3.** Percentage of overlap between speech and head gaze (Condition A)

| | Overall | Condition A first | Condition B first | Long commands | Short commands |
|---|---|---|---|---|---|
| Avg | 70.06 ± 36.2 | 53.22 ± 38.47 | 76.96 ± 33.3 | 34.63 ± 37.16 | 78.57 ± 30.69 |
| N | 62 | 18 | 44 | 12 | 50 |

Comparing columns 2 and 3 of Table 3, it is apparent that subjects who were exposed to Condition B first (headpose based attention system) have a larger overlap percentage under Condition A (wake-up word system) than those who are first exposed to Condition A. To ascertain whether this observation was statistically significant, we applied a Wilcoxon Rank-Sum Test (which is applicable for non-normal distributions) to the data that underlie columns 2 and 3 of Table 3, finding that the overlap difference was significant with a p-value of 0.02 (thus significant at $p < 0.05$). The fact that users who first used the head pose system continued to exhibit a behavior that no longer had any impact may suggest that the behavior is readily learned, and so natural as to be almost subconscious.

## 5 Conclusions and Future Work

In this work, we have demonstrated that one can build a practical embodied agent that is capable of operating in environments where multiple humans are conversing with one another and interacting with the agent in an interleaved fashion. The agent is practical in the sense that (a) it does an adequate job of discerning when it is being addressed without imposing on the user the burden of using a wake-up word, and (b) it is relatively inexpensive to implement—requiring only a simple camera and headpose estimation software.

A key finding from a first informal pilot study was that head pose estimation did not work adequately by itself, but there were hints that users might adapt to some form of feedback indicating when the agent believed the user was looking at the system. Inspired by this finding, we enhanced the agent by providing such feedback in the form of dots and arrows, and ran a second experiment that allowed us to quantify its likability and usability relative to that of an alternate variant of the agent that required a wake-up word. We found that users adapted very readily (perhaps even subconsciously) to this form of feedback, thereby amplifying what would otherwise be a weaker signal. Analysis of user responses showed that the enhanced agent was both likeable and usable, and that its likability was greater than that of the wake-up word agent to a statistically significant degree. (There were indications that the usability was also greater for the headpose-based agent, but not quite at a statistically significant level).

Based upon these initial implementations and studies, we feel encouraged that head orientation can be used as a simple, low-cost basis for more natural interactions with cognitive assistants that engage in extended multi-modal dialogues with multiple people. We see multiple avenues for future efforts. In addition to pursuing improvements in the cost and accuracy of headpose estimation, it would be worthwhile to couple headpose estimation with other non-verbal clues regarding the addressee. We have identified the length of an utterance as one such factor; the Related Work section of this paper contains numerous other factors that prior authors have identified as being correlated with attention and are therefore good candidates for future study. An important question to be resolved is the tradeoff between the incremental accuracy (and concomitant likability and usability) provided by these additional factors versus their additional cost. Opening up the interaction to multiple agents who are aware of their human partners as well as the other agents is another exciting direction to pursue (for example, an early prototype of two shopkeeper agents negotiating with humans is reported by Divekar et al. 2019).

**Acknowledgment.** This work is supported by the Cognitive and Immersive Systems Laboratory (CISL), a collaboration between IBM Research and Rensselaer Polytechnic Institute, and also a center in IBM's AI Horizon Network. Part of this work was done while the first author was an intern at IBM T.J. Watson Research Center, Yorktown Heights, NY.

## References

Admoni, H., Scassellati, B.: Social eye gaze in human-robot interaction: a review. J. Hum. Robot Interact. **6**(1), 25–63 (2017)

Ahn, B., Park, J., Kweon, I.S.: Real-time head orientation from a monocular camera using deep neural network. In: Cremers, D., Reid, I., Saito, H., Yang, M.-H. (eds.) ACCV 2014. LNCS, vol. 9005, pp. 82–96. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-16811-1_6

Akhtiamov, O., Palkov, V.: Gaze, prosody and semantics: relevance of various multimodal signals to addressee detection in human-human-computer conversations. In: Karpov, A., Jokisch, O., Potapova, R. (eds.) SPECOM 2018. LNCS (LNAI), vol. 11096, pp. 1–10. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99579-3_1

Akhtiamov, O., Sidorov, M., Karpov, A.A., Minker, W.: Speech and text analysis for multimodal addressee detection in human-human-computer interaction. In: INTERSPEECH, pp. 2521–2525 (2017)

Amazon: Avs ux attention system (2019). https://developer.amazon.com/docs/alexa-voice-service/ux-design-attention.html. Accessed 24 Jan 2019

Andrist, S., Bohus, D., Mutlu, B., Schlangen, D.: Turn-taking and coordination in human-machine interaction. AI Mag. **37**(4), 5–6 (2016)

Baba, N., Huang, H.H., Nakano, Y.I.: Addressee identification for human-human-agent multiparty conversations in different proxemics. In: Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, p. 6. ACM (2012)

Bakx, I., Van Turnhout, K., Terken, J.M.: Facial orientation during multi-party interaction with information kiosks. In: INTERACT (2003)

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., Lottrjdge, D.: Understanding the long-term use of smart speaker assistants. Proc. ACM Interact Mobile Wearable Ubiquit. Technol. **2**(3), 91 (2018)

Bohus, D., Horvitz, E.: Multiparty turn taking in situated dialog: Study, lessons, and directions. In: Proceedings of the SIGDIAL 2011 Conference, pp. 98–109. Association for Computational Linguistics (2011)

Borghi, G., Fabbri, M., Vezzani, R., Calderara, S., Cucchiara, R.: Face-from-depth for head pose estimation on depth images. arXiv preprint arXiv:1712.05277 (2017)

Perreira Da Silva, M., Courboulay, V., Prigent, A., Estraillier, P.: Real-time face tracking for attention aware adaptive games. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) ICVS 2008. LNCS, vol. 5008, pp. 99–108. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-79547-6_10

De, J.G.X.Y.S., Kautz, M.J.: Dynamic facial analysis: from Bayesian filtering to recurrent neural network (2017)

Dementhon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. Int. J. Comput. Vis. **15**(1–2), 123–141 (1995)

Divekar, R.R., et al.: Interaction challenges in ai equipped environments built to teach foreign languages through dialogue and task-completion. In: Proceedings of the 2018 Designing Interactive Systems Conference, DIS 2018, pp. 597–609. ACM, New York (2018). ISBN 978-1-4503-5198-0, https://doi.org/10.1145/3196709.3196717

Divekar, R.R., Mou, X., Chen, L., de Bayser, M.G., Guerra, M.A., Su, H.: Embodied conversational AI agents in a multi-modal multi-agent competitive dialogue. In: IJCAI (2019)

Farrell, R.G., et al.: Symbiotic cognitive computing. AI Mag. **37**(3), 81–93 (2016)

Frampton, M., Fernández, R., Ehlen, P., Christoudias, M., Darrell, T., Peters, S.: Who is you?: combining linguistic and gaze features to resolve second-person references in dialogue. In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 273–281. Association for Computational Linguistics (2009)

Gravano, A., Hirschberg, J.: Turn-yielding cues in task-oriented dialogue. In: Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pp. 253–261. Association for Computational Linguistics (2009)

Gu, E., Badler, N.I.: Visual attention and eye gaze during multiparty conversations with distractions. In: Gratch, J., Young, M., Aylett, R., Ballin, D., Olivier, P. (eds.) IVA 2006. LNCS (LNAI), vol. 4133, pp. 193–204. Springer, Heidelberg (2006). https://doi.org/10.1007/11821830_16

Katzenmaier, M.: Identifying the addressee in human-human-robot interactions based on head pose and speech. Ph.D. thesis, Carnegie Mellon University, USA and University of Karlsruhe TH, Germany (2004)

Kendon, A.: Some functions of gaze-direction in social interaction. Acta Psychol. **26**, 22–63 (1967)

Kephart, J.O., Dibia, V.C., Ellis, J., Srivastava, B., Talamadupula, K., Dholakia, M.: A cognitive assistant for visualizing and analyzing exoplanets. In: Proc. AAAI 2018 (2018)

Le Minh, T., Shimizu, N., Miyazaki, T., Shinoda, K.: Deep learning based multi-modal addressee recognition in visual scenes with utterances. In: IJCAI 2018, pp. 1546–1553 (2018). https://doi.org/10.24963/ijcai.2018/214

Lin, G.S., Tsai, T.S.: A face tracking method using feature point tracking. In: 2012 International Conference on Information Security and Intelligence Control, ISIC, pp. 210–213. IEEE (2012)

Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J., Ishiguro, H.: Conversational gaze mechanisms for humanlike robots. ACM Transact. Interact. Intell. Syst. **1**(2), 1–33 (2012). https://doi.org/10.1145/2070719.2070725. ISSN 21606455, http://dl.acm.org/citation.cfm?doid=2070719.2070725

Nakano, Y.I., Baba, N., Huang, H.H., Hayashi, Y.: Implementation and evaluation of a multimodal addressee identification mechanism for multiparty conversation systems. In: Proceedings of the 15th ACM on International conference on multimodal interaction, pp. 35–42. ACM (2013)

Norouzian, A., Mazoure, B., Connolly, D., Willett, D.: Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed. arXiv preprint arXiv:1902.00570 (2019)

Radziwill, N.M., Benton, M.C.: Evaluating quality of chatbots and intelligent conversational agents. arXiv preprint arXiv:1704.04579 (2017)

Ranganatha, S., Gowramma, Y.: An integrated robust approach for fast face tracking in noisy real-world videos with visual constraints. In: 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 772–776. IEEE (2017)

Ravuri, S., Stolcke, A.: Recurrent neural network and LSTM models for lexical utterance classification. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)

Ruhland, K., et al.: A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. In: Computer Graphics Forum, vol. 34, pp. 299–326. Wiley (2015)

van Schendel, J.A., Cuijpers, R.H.: Turn-yielding cues in robot-human conversation. New Front. Hum. Robot Interact., p. 85 (2015). URL http://www.mahasalem.net/AISB2015/NF-HRI-2015-full_proceedings.pdf#page=86

Sciuto, A., Saini, A., Forlizzi, J., Hong, J.I.: Hey alexa, what's up?: A mixed-methods studies of in-home conversational agent usage. In: Proceedings of the 2018 on Designing Interactive Systems Conference 2018, pp. 857–868. ACM (2018)

Sheikhi, S., Odobez, J.M.: Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. Pattern Recogn. Lett. **66**, 81–90 (2015). https://doi.org/10.1016/j.patrec.2014.10.002. ISSN 01678655

Shriberg, E., Stolcke, A., Ravuri, S.V.: Addressee detection for dialog systems using temporal and spectral dimensions of speaking style. In: INTERSPEECH, pp. 2559–2563 (2013)

Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: CHI 2002 Extended Abstracts on Human Factors in Computing Systems, pp. 858–859. ACM (2002)

Tsai, T., Stolcke, A., Slaney, M.: A study of multimodal addressee detection in human-human-computer interaction. IEEE Transact. Multimedia **17**(9), 1550–1561 (2015)

Van Turnhout, K., Terken, J., Bakx, I., Eggen, B.: Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In: Proceedings of the 7th international conference on Multimodal interfaces, pp. 175–182. ACM (2005)

Venturelli, M., Borghi, G., Vezzani, R., Cucchiara, R.: From depth data to head pose estimation: a siamese approach. arXiv preprint arXiv:1703.03624 (2017)

Wang, K., Ji, Q.: Real time eye gaze tracking with 3D deformable eye-face model. In: Proceedings of IEEE CVPR, pp. 1003–1011 (2017)

Wu, Y., Gou, C., Ji, Q.: Simultaneous facial landmark detection, pose and deformation estimation under facial occlusion. arXiv preprint arXiv:1709.08130 (2017)

Wu, Y., Ji, Q.: Facial landmark detection: a literature survey. Int. J. Comput. Vis. **127**, 1–28 (2017)

Xu, T.L., Zhang, H., Yu, C.: See you see me: The role of eye contact in multimodal human-robot interaction. ACM Transact. Interact. Intell. Syst. (TIIS) **6**(1), 2 (2016)

ZDNet: How alexa developers are using visual elements for echo show (2018). https://www.youtube.com/watch?v=eZIouIY5p8Q

Zhao, R., Wang, K., Divekar, R., Rouhani, R., Su, H., Ji, Q.: An immersive system with multi-modal human-computer interaction. In: 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, pp. 517–524 (2018)