Augmenting Speech Agent with Gaze for Enhancing Interaction

By Drawing from human-human Interaction

RAZAN JABER, DSV, Stockholm University, Sweden

Speech technologies are increasing in popularity by offering new interaction modalities for users. Despite the prevalence of these devices and the rapid improvement of the underlying technology, there has been a slower improvement in interaction with them. Spoken interaction design centers around the use of a wake-word to initiate interaction and the transcription of the users' spoken instructions to complete the task. However, in human-to-human conversation, speech is initiated by and supplemented with other modalities, such as gaze and gesture. My research focuses on the need to better understand how human-technology 'conversations' can be improved by borrowing from human-human interaction. Recent developments in gaze tracking present new opportunities for social computing. Therefore, Tama – a gaze-activated smart speaker, was designed to explore the use of gaze in conversational interaction. Tama uses gaze to indicate attention and intent to interact on behalf of the user and as feedback.

Additional Key Words and Phrases: Smart Speaker; Voice Assistant; Gaze Interaction; User Study.

1 INTRODUCTION

In developing speech technologies, designers usually rely on a single modality of interaction — transcribed speech. All of the speech commercial devices have the same wake-word based interaction paradigm — the speaker constantly listens for a small set of wake words or phrases (e.g., "Ok, Google", " Hey, Siri" or "Alexa"). However, in human-to-human conversation, speech is supplemented with a range of other modalities such as gaze, touch, prosody, and gesture to communicate information other than the spoken words [1]. However, these modalities are mostly ignored in the current generation of speech systems. This has motivated my research to introduce gaze modality to a speech agent to advance its interaction. In my research, I aim to incorporate abstractions of complex human behavior to enhance the interaction with speech agents, especially gaze and gesture in conversations. Human behaviors are contextually and socially dependent and require careful abstraction. In this research, human communication is thought of as a template to borrow from and guide the design process rather than a set of hard and fast rules to be followed to simulate how humans interact.

During the conversation, eye gaze plays a vital role in maintaining the flow, regulation of turns in talk [9], understanding the other person's interest and response, speaker selection [7], as well as a host of explicit communicative features such as deixis [3]. This all points to the profoundly ingrained role that the gaze plays in conversation and the opportunity it presents for interaction with voice-based systems. Tama (see Figure 1) was designed by taking advantage of how individuals use eye gaze to manage the conversation. Rather than being activated with a wake word (such as "Ok Google"), Tama detects a user's gaze, moving an articulated 'head' to achieve mutual gaze. It uses gaze to trigger and confirm that a command is being spoken. Building on the Google Voice Assistant service, Tama acts as a Intelligent Personal assistant (IPA). It affords being looked at, looking back, and the establishment of mutual gaze as part of speech interaction – opening this modality for both design and study. Drawing on conversation analysis provided particular resources for understanding how gaze unfolds and when and how Tama could behave and make use of gaze, especially the work of Goodwin [3] who documented in detail the different aspects of the use of gaze between listeners and speakers.

The aim of this research is not to give Tama behaviors that make it seem human or alive but to limit the interaction and expression to a *minimally anthropomorphic design* in order to reduce the expectations of nuance and complexity such interaction may engender in users. My research aims to perform that simplification and

IMX '22 - Doctoral Consortium, 2022, Portugal

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

projection in design, producing an interaction that is not human-like in its behavior but would take advantage of learned skills in human interactions.

2 RESEARCH QUESTIONS

My research investigates three interrelated questions:

RQ1: How could speech agents evolve by drawing from human-human interaction, conversational analysis, and sociology?

RQ2: How could the multimodal approach be used to overcome some of the practical limitations related to using speech as the only interaction method for speech interfaces?

RQ3 What is the role of non-speech modalities in interaction with speech agents?

3 WORK IN PROGRESS

My research draws on human-human interaction by taking advantage of research on gaze in conversation. My research has begun with the first phase of designing Tama to explore designing with gaze for multiparty conversationally situated interaction. At the current stage, which is the second phase of my research, I have been working on exploring the use of speech systems in different settings, such as cooking, to understand users' expectations and requirements in a natural environment.

Phase 1: Designing Tama – a gaze activated smart speaker

My Ph.D. research began with designing Tama to explore how a smart speaker could be brought into a conversation the same way as another human conversational partner [3, 10], using the offer of mutual gaze and its reciprocation. Tama is designed as a smart speaker device, somewhat like Amazon Alexa or Google Home, but responding not to a wake word (like 'Hey, Siri' or 'Ok, Google') but rather to the user's gaze on the device. To provide gaze feedback, it is designed as a retractable (Figure 1-left) spherical head containing two full-color LED eyes with 180 degrees of movement laterally and 60 degrees vertically. The device can then be looked at, look back, and establish mutual gaze as part of speech interaction. It uses the commercial Google's Voice Assistant API to process the queries and provide an appropriate voice response. When the two users engaged with gaze, Tama would change mutual gaze direction depending on the direction of arrival (DOA) of the voice detected using the directional microphone and lock it to the user speaking until both gaze and voice were detected from another direction.

A user study has been conducted in a semi-experimental multi-user setting to explore how the system could be incorporated into ongoing talk [8]. The main focus of the first user study was to examine the design of gaze as an input method by using gaze to trigger and confirm that a command was being spoken. This study has discussed how gaze interaction can benefit from ongoing co-present conversation. Also, to study and statistically analyze gaze behavior during a conversation with speech agents to suggest possible explanations about the role

of gaze in interaction with speech agents. This was the initial step to exploring the possibility of using gaze as an actuation with speech agents. In particular, this study showed that gaze interaction resulted in a better user experience when it included gaze output and input. The study suggested that gaze can be used to augment, or even replace, the wake-work in initiating interaction with speech agents. Analysis revealed interesting subtleties in how gaze is used to manage talk in interaction.

Further analysis was conducted using machine learning methods to categorize how and when users looked at the system in relation to the conversational interaction, exposing patterns of looking both similar to and distinct from human-human interaction [5]. This was done using the k-means clustering algorithm of the log data collected on the system's internal state and the interaction with the google assistant service during the experiment. Here, the focus was on how the gaze cameras recorded where each participant was looking and the direction of arrival of sound from the directional microphone. Interestingly, the challenge in designing with gaze, is to find out how gaze changes over time and the role of gaze in interaction. The results pointed to gaze patterns during query and listening to the answer, which can be used as a starting point to build gaze-awareness into voice-user interfaces. These patterns were verified and explained through close analysis of the video data of the trial. These findings could be incorporated into the design of conversational agents which would exhibit appropriate gaze behaviors during dialogues with human users.

After completing the first study, I have conducted a literature review focusing on systematically exploring the mobile conversation user interface research, which aimed at understanding the difference between developing a speech interface on a standalone smart speaker or an embodied robot and developing for a consumer mobile device [4]. The motivation was to explore how mobile devices offer specific challenges and opportunities for conversational user interfaces, which can inform the design of Conversational User Interfaces (CUIs) more broadly. This work presented the dominant themes distinctive to the form factor on which the CUIs were developed or studied and the four prevalent domains of application of CUIs on mobile devices. Moreover, it discussed the impacts of multi-modality on CUIs, how the form factor can influence the recovery from breakdowns in communication between humans and devices, and how the mobile device can aid in the contextualization of CUI interaction.

The main findings of the first phase suggest that the gaze can be used as an effective way of initiating interaction with a system and highlights the problems of using the gaze as an input method to establish and maintain a mutual gaze with the system. This has encouraged further work exploring speech agents' use in more real-world settings such as cooking.

Phase 2: Investigating the Use of Speech Systems in Everyday Tasks

The second phase of my research explores realistic user requirements for an ideal instructional video playback control while cooking in a wizard-of-oz setting [6]. Through investigating the use of an ideal voice-control system to follow-along cooking videos (as a popular example of non-linear instructional videos) in a home environment, I have highlighted the technical and social contexts, patterns of command formulation, and the challenges in request responding [11]. To better design VUIs, we need to understand the full range of machine behaviors users expect to leverage when approaching a speech interface. Not simply to implement all possible ways that a user may interact with a system, but rather to predict and mitigate potential breakdowns when these behaviors are not implemented or do not work as users expect. The findings highlighted four types of high-level interactions or factors: context, navigational command types, dimensions of content-based commands, and challenges for design.

Through the analysis of the issued commands and performed actions during this non-linear and complex task, I identified (1) patterns of command formulation, (2) challenges for design, and (3) how task and voice-based commands are interwoven in real-life. I have also discussed implications for the design and research of voice interactions for navigating instructional videos while performing complex tasks.

The findings of this work have led to further work to explore using Tama smart speaker platform (Figure 1) to compare 'cued-gaze' and spoken commands in advancing a list and when things go wrong. In this study, the

agent was designed to fail when providing the list of instructions to explore how the participants proceeded to recover from common failures. This showed that, for this use case, the cross-modality repair was more effective than a reformulation of speech. In the quantitative results, I showed that 'cued-gaze' is as effective a method of progressing lists as speech and also that it is (in certain circumstances) *more* effective when errors of interaction can be addressed by switching modalities to gaze.

Humans and machines perceive speech differently. In the design of conversational agents, we need to understand the range of behavior that users expect and the range of behavior they may attempt to employ when interacting with them. This will allow us to understand potential breakdowns, and provide support strategies within the interaction that match users' recovery procedures in challenging human-human interactions.

The following steps of this work are integrating more complex gaze use as both an input and output for conversational interaction with Tama, exploring the use of physical gestures and posture to enhance interaction, and exploring the use of the system in a real-world setting. The next step during my Ph.D. is investigating the usability of Tama for independent living in a specific domestic task such as cooking a meal at home. I am planning to explore older adults' expectations and preferences towards the use of speech agents in kitchen tasks. This work will focus on exploring the use of the system when providing speech-based instructions for three different activities: tool suggestion, recipe reminding, and orientation of action in the kitchen.

4 RESEARCH METHOD

Investigating the use of gaze as an actuation method for speech agents was challenging. Using methodological techniques has been key to collecting and analyzing rich data of interacting with the speech agent in a conversational setting. In taking a mixed-method approach, the data were analyzed quantitatively and qualitatively to understand what was happening when interacting with the device.

Video has been proven valuable for actual recording behavior and provides more complete (and visual) information than reported behavior. Video methods can be effective for research conducted in single rooms since the cameras can be set up in a fixed position, mainly focusing on the interaction with the device. I have used video recordings in my studies, which allowed me to capture simultaneous complex interactions when interacting with Tama, and has provided enough detail to analyze participants' interactions quantitatively and qualitatively. The resulting video data supported the analysis of details of the interaction, allowing precise understating of the verbal, visual, and physical interaction sequence. Video recordings have been viewed iteratively and codded collaboratively with other researchers to develop an understanding of the interaction with the device. In part of my work, inductive thematic analysis [2] has also been used to analyze transcribed video recordings and understand participants' overall experience when interacting with speech systems and interfaces.

In addition to video recording, computational logging has been used to automatically log the data generated from the cameras and the microphones for further analysis. Automatic logging involves instrumenting the device to provide quantitative data for every instance of interaction with the device. The logs from Tama during the trial provided data on the system's internal state and the interactions with the Google Assistant service. The gaze cameras recorded where each participant was looking and the direction of arrival of sound from the directional microphone. The quantitative data allowed me to gain insight into how participants gazed at the system and provided a resource to analyze the interaction with the system using machine learning methods.

5 CONCLUSION

Through this research, I hope to investigate the use of multiple modalities of interaction for speech systems. Augmenting speech interfaces with gaze awareness can effectively provide feedback and increase communicative engagement with users. Using gaze to initiate interaction is a valuable opportunity for interaction design. I aim to contribute new understandings of how speech agent interaction could be designed by adding more gaze

interaction and feedback, detecting gestures, and incorporating non-verbal information. Within the scope of my Ph.D. I aim to design more informative feedback using gaze and head movements to provide speech-based instructions for users when doing everyday tasks such as cooking.

6 ACKNOWLEDGEMENTS

Special thanks to Donald McMillan, Barry Brown, and Jordi Solsona Belenguer.

REFERENCES

[1] Nalini Ambady and Robert Rosenthal. 1998. Nonverbal communication. *Encyclopedia of mental health* 2 (1998), 775–782.

[2] V. Braun and V. Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. http://dx.doi.org/10.1191/1478088706qp063oa

[3] 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

[4] Razan Jaber and Donald McMillan. 2020. Conversational User Interfaces on Mobile Devices: Survey. In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3405755.3406130 tex.ids= jaberConversationalUserInterfaces2020a, jaber_conversational_2020-1.

[5] Razan Jaber, Donald McMillan, Jordi Solsona Belenguer, and Barry Brown. 2019. Patterns of Gaze in Speech Agent Interaction. In *Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19)*. ACM, New York, NY, USA, 16:1–16:10. https://doi.org/10.1145/3342775.3342791 tex.ids: jaber_patterns_2019-1, jaber_patterns_2019-2 event-place: Dublin, Ireland.

[6] John F. Kelley. 1983. An empirical methodology for writing user-friendly natural language computer applications. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 193–196. https://doi.org/10.1145/800045.801609

[7] Gene H. Lerner. 2003. Selecting next speaker: The context-sensitive operation of a context-free organization. *Language in Society* 32, 2 (April 2003), 177–201. https://doi.org/10.1017/S004740450332202X

[8] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. 2019. Designing with Gaze: Tama – a Gaze Activated Smart-Speaker. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 176:1–176:26. https://doi.org/10.1145/3359278

[9] David G. Novick, Brian Hansen, and Karen Ward. 1996. Coordinating turn-taking with gaze. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96*, Vol. 3. 1888–1891 vol.3. https://doi.org/10.1109/ICSLP.1996.608001

[10] 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

[11] Yaxi Zhao, Razan Jaber, Donald McMillan, and Cosmin Munteanu. 2022. "Rewind to the Jiggling Meat Part": Understanding Voice Control of Instructional Videos in Everyday Tasks. *(CHI '22)*. New York, NY, USA. https://doi.org/10.1145/3491102.3502036