5 Seconds After: Exploring User Actions with Voice Assistants in the Moments After a System Response

Alexandra Vtyurina
University of Waterloo
Waterloo, ON, Canada
avtyurin@uwaterloo.ca

Adam Fourney
Microsoft Research
Redmond, Washington
adamfo@microsoft.com

Edith Law
University of Waterloo
Waterloo, ON, Canada
edithlaw@uwaterloo.ca

Abstract

Voice-based conversational assistants are widely used for a variety of tasks, from setting up an alarm clock and checking the weather to booking a flight. However, the communication protocol that the majority of these assistants offer is very restrictive and often fails to take advantage of the full richness of human speech signals. In this paper, we explore human-agent voice interactions in an unrestricted environment, and discuss potential benefits and drawbacks from possible changes to standard communication protocols.

Author Keywords

conversational assistants; voice-only interaction.

ACM Classification Keywords

H.5.m. [Information Interfaces and Presentation (e.g. HCI)]: Miscellaneous

Introduction

Conversational assistants have a long history. Beginning with text-based chat-bots [9], and growing more sophisticated, they are now ubiquitous. Today, stationary devices, equipped with a smart assistant are often in people's houses ready to operate. In this paper, we discuss these home-based smart assistants, as they mostly rely on voice-only communication with people.

Voice-based interactions differ dramatically from the on-screen ones, and interaction models must be designed accordingly. The assistants are constantly moving towards better imitation of human speech. For example, they are able to make jokes and their voices are remarkably similar to those of people. However, their communication protocol is still very restrictive. The usual interaction is a three-step procedure: *<wake word> – <user query> – <system response>*, Figure 1. These triples need to be repeated for every user intent, bearing potential for user frustration.

User: *Hey Google*,
User: *What is the population of Montreal?*

Agent: *The population of Montreal was 1.741 million in 2014*.

Figure 1: Example exchange with a conversational assistant

Human-human conversations rarely follow the same paradigm. More often than not, people engage in multi-turn discussions with each other, even if the initial information need was satisfied in the first turn, e.g. it is socially accepted to thank the interlocutor after asking for a piece of information. For more complex information needs, a variety of signals is utilized to detect turn-taking behaviour [4]. In addition, a multitude of approaches are used by people to deliver certain signals, e.g. grounding, error handling [6, 5].

By restricting the communication protocol, voice-based assistants miss a big part of the human speech signal, that could be used to estimate and predict user engagement and satisfaction, enhance system answer ranking, and improve an overall experience.

In this paper we explore and analyze user behaviour when engaging with a conversational assistant, that does not impose any protocol limitations. We analyze the results of a Wizard of Oz study, where participants cooked a culinary recipe, using the help of a voice-based assistant. We target the following research questions:

- **RQ1**: Given an unrestricted communication protocol, do people provide feedback to the agent?
- **RQ2**: In what span of time do people usually provide their feedback?
- **RQ3**: What types of feedback are provided?

Data collection

In order to study people's behaviour with voice-based conversational assistants in an unrestricted environment, we conducted a high-fidelity Wizard of Oz study. The participants were asked to follow a culinary recipe[1] using a simulated conversational assistant. To keep the environment stable, the Wizard had a selection of candidate answers, from which they chose the best one, and that answer was read out loud to the participant using a text-to-speech software[2]. The Wizard also had a capability of constructing free-form answers, but this option was only used in about 3 per cent of the cases. The experiment is described in more detail in [8].

During the experiment the participants were instructed to say “Start cooking” to activate the system, and then talk to the system like they would with another person. The system did not require a wake-word to be used, was always active, and could answer participants' questions with high accuracy, due to the Wizard of Oz setup.

We invited 10 people (6 male, 4 female) to participate. Out of the 10 participants, 2 reported having used an intelligent assistant earlier that day, 5 – earlier that week, and 1 each – earlier that month, more than a month ago and never. During the experiment, the interactions were audio and video recorded for further analysis of collected data.

After the experiment, all user and system utterances were manually transcribed, and each user utterance was labelled with their respective category [8].

[1] http://allrecipes.com/recipe/17338/tasty-bbq-corn-on-the-cob/
[2] https://responsivevoice.org/

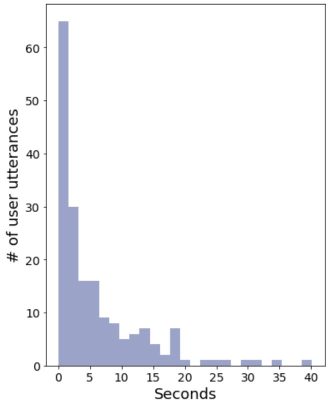

Figure 2: The distribution of time elapsed between an agent finishing speaking and a user starting their next utterance. We define this time as *user time delay*.

In order to explore the timing in user-agent interactions, we need to know the time of when each user and agent utterance started and ended. To achieve a high accuracy with timestamps, we used *webrtcvad*[3] Python library to identify the sections of audio recording, that contained speech. Since the Wizard was implemented using text-to-speech technology, our approach worked both for the system and the user utterances.

Results and analysis

In this section, we discuss the results obtained from the experiment. In particular, we investigate the distribution of users' follow up utterances in terms of time delay and type of utterance. In addition, we provide recommendations on how it can be used.

Timing of user follow up utterances

To investigate the timing of user follow up utterances, we considered all pairs of consecutive agent-user utterances, where a user utterance followed a system utterance. We defined *user time delay* as the time between the end of the system utterance and the beginning of the subsequent user utterance. We considered an overall distribution for all 10 participants. The histogram shown in Figure 2 illustrates that user follow up utterances follow the agent's utterance within a span of 5 seconds in the majority of the cases. The long tail shows, however, that users may also engage with the system after a much longer time period.

Based on this finding, we speculate, that provided an opportunity, users of voice-based assistants will engage in multi-turn conversations with their assistants. Moreover, the multi-turn exchange is likely to happen in the close temporal vicinity of the initial interaction.

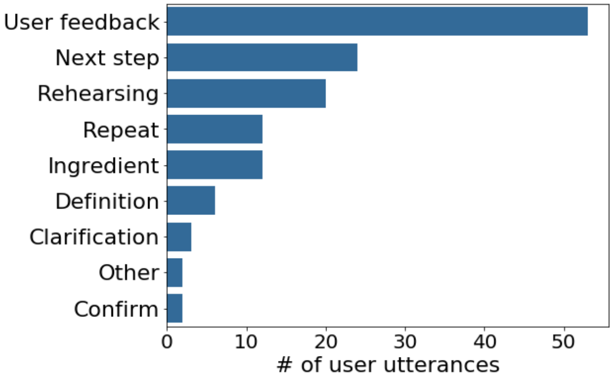

Figure 3: The distribution of types of user utterances that followed within 5 seconds of when the system stopped speaking.

Types of user follow up utterances

Next, we examine what types of user utterances follow an agent's utterance within the first 5 seconds after the system stops speaking. For that we use categories that were assigned to user utterances previously [8].

Figure 3 illustrates the summary of types of user utterances that occurred within 5 seconds of the moment when the agent stopped speaking. By far the most dominating category is *User feedback*. This category included examples of:

- **grounding feedback** [6] either by saying short positive utterances like *"Okay"*, *"Got it"*, or by repeating fully or partially the system's previous response – *"Black pepper. Okay."*.
- **intent feedback**. Being able to handle errors without starting the conversation anew is a crucial functionality for human-agent systems [2]. This category included such examples as *"No, I mean all the ingredients"*, where the participant wanted to make corrections to the agent's behaviour.

[3] https://pypi.python.org/pypi/webrtcvad

- **relevance feedback**. Often participants provided feedback about whether the agent's answer was satisfactory or not. For example, utterances like *"Okay, perfect."*, *"Oh, it's not very good"* occurred in response to the agent's utterances.

Other common intents that appeared in the 5 seconds window included asking for the next step in the recipe, asking questions related to the ingredients (e.g. quantity), asking for a definition of a term (e.g. *"What is an ear of corn?"*, *"How much is a pinch?"*), and other utterances that appeared anecdotally and did not constitute separate categories.

We defined *rehearsing behaviour* as utterances that were not addressed to the agent, but rather to the participant him/herself, to keep the object's name in memory, and repeat instructions while completing them.

Intents like *clarification* and *repeat*, occurring very quickly after the initial interaction may signal non-understanding of the information that is being transferred (*repeat*), or may be the user's implicit request to provide more explanation (*clarification*). For example, a user asking *"sorry, can you repeat that again?"* could indicate that they didn't hear the agent's utterance fully. And in the example interaction in Figure 4 the user indicates that he could benefit from an alternative way of describing the action.

Agent: *Step number 4: blend in the softened butter.*

User: *Blend?*

Figure 4: User expresses uncertainty about the agent's response

Relevance feedback
It is worth noting, that the category of *User feedback* constituted the majority of user utterances appearing in the 5 seconds frame. This signal at least partially could be considered as explicit relevance feedback. Relevance feedback is a subject well studied in the IR community. Collecting explicit relevance feedback, e.g explicitly indicating relevant documents, is both "expensive and limited in coverage"[1] and users often are not willing to use the tools for it [10]. Therefore other metrics were used, such as dwell time, scroll time, reformulation patterns, clickthrough behaviour, to approximate it [11].

Trippas et al. [7] report in their study of voice-only mediated search, that the searchers provided explicit relevance feedback, even when not asked to do it. In our data we also see that utterances of type *User feedback* dominate, and some of the utterances indeed do provide adjustments for previous agent's responses. While others could serve as a strong signal for estimating and predicting user satisfaction.

Future work

We have shown that in an unrestricted environment users often reply to the agent within a short amount of time, engaging in a multi-turn interaction. Currently existing human-agent communication protocols are restrictive and usually do not capture a response the user might provide. Exploring alternative communication protocols is needed to make solid conclusions about the benefits and drawbacks of enabling multi-turn interactions with voice-only conversational agents.

One obvious drawback of "keeping the mic open" after the system's response relates to questions of user privacy. While we acknowledge that this is a central topic to be investigated in the context of our protocol recommendation, we argue that there may be an acceptable time limit allowing user feedback to be captured while at the same time respecting users' privacy. Letting a user opt-in for a multi-turn conversations functionality could also be a part of a solution to this problem.

Moreover, in the experiment we described, the agent was powered by a Wizard, who could make correct judgments about whether the participant was talking to the agent or not. Identifying the addressee is a challenging issue, and even more so if the interaction happens in a multi-person environment, where the user may address the system, or another person, or the system may be addressed by several different users [3].

Conclusion

Voice-based stationary conversational devices are becoming ubiquitous. However, the communication protocol is rigid and restrictive. Currently existing conversational agents do not allow multi-turn interactions that were shown to carry a rich signal, that could be helpful for error handling, collecting relevance feedback, and evaluating user satisfaction. Further research is needed to explore the use of alternative communication protocols, as well as potential problems that might come with it.

REFERENCES

1. Steve Fox, Kuldeep Karnawat, Mark Mydland, Susan Dumais, and Thomas White. 2005. Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS)* 23, 2 (2005), 147–168.
2. Julia Kiseleva, Kyle Williams, Ahmed Hassan Awadallah, Aidan C Crook, Imed Zitouni, and Tasos Anastasakos. 2016. Predicting user satisfaction with intelligent assistants. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 45–54.
3. Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2016. “Do animals have accents?”: talking with agents in multi-party conversation. In *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing*.
4. Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* (1974), 696–735.
5. Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication* 45, 3 (2005), 325–341.
6. David R Traum and Peter A Heeman. 1996. Utterance units in spoken dialogue. In *Workshop on Dialogue Processing in Spoken Language Systems*. Springer, 125–140.
7. Johanne R Trippas, Damiano Spina, Lawrence Cavedon, Hideo Joho, and Mark Sanderson. 2018. Informing the Design of Spoken Conversational Search. In *Proceedings of 2018 Conference on Human Information Interaction Retrieval*. ACM.
8. Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM.
9. Joseph Weizenbaum. 1966. ELIZA – a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.
10. Ryen W White, Ian Ruthven, and Joemon M Jose. 2002. The use of implicit evidence for relevance feedback in web retrieval. In *European Conference on Information Retrieval*. Springer, 93–109.
11. Ryen W White, Ian Ruthven, and Joemon M Jose. 2005. A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 35–42.