Incorporating Human In The Loop For Voice Assistants

Shih-Hong Huang[1]
[1]The Pennsylvania State University
201 Old Main
University Park, Pennsylvania 16802
szh277@psu.edu

Abstract

Millions of voice-enabled smart devices equipped with voice assistants (VAs) have been sold. However one promised feature of VAs: Allowing users to converse with them naturally like human companions, still fail to deliver. Users have to interact with VAs in certain scripted ways to carry desired functionalities. Prior work suggested that it is possible to incorporate human-in-the-loop for text base conversational agents and provide quality responses. This proposal provided an overview on VAs and reviewed how people interact with them, then presents EchoPal, a prototype system allowing a human receptionist (worker) to converse with the user synchronously via connection between a designated web interface and an Amazon's Echo device. Pilot study is conducted to give insight on the possibility and usability of such a setup, advantages and trade offs of EchoPal are further discussed. Finally, the notion of asynchronous conversation is introduced considering the outcome of the pilot study and ThinkerPal is proposed with the aim to utilize asynchronous conversation to promote better VA using experience. Key characteristics of ThinkerPal are presented, possible research questions and directions regarding the system are further discussed.

Introduction

Voice interaction is one of the most utilized schemes by smart devices. Using voice as a medium to interact opens up the possibilities for hands-free using scenarios and further lower the barrier of accessing technologies and information. Amazon alone sold more than 100 million devices equipped with the company's Alexa voice assistant (VA) as of early 2019 (Bohn 2019). Utilizing conversations, on the other hand, allow people to elaborate themselves on the fly. By going back-and-forth on the topic, the needs of users can be better defined. Such pattern also encourages users to further explore the problem space iteratively. One goal of the VAs is to provide human-like conversational experience. However, users need to adapt to the structured conversational pattern provided by the system in order to have a smooth interaction with the VAs. A study based on the interactions between Alexa and users in a household scenario suggested that the design forced users to request responses from Alexa rather than engage in natural conversations (Porcheron et al. 2018). The request and response framework follows predefined structures, which is different from human conversations where rules are established as the conversation goes on (Porcheron et al. 2018). Moreover, it is hard for users to continue their conversations because it is not clear whether VAs have identity or personality (Luger and Sellen 2016), which further widens the gap between the expectation and user experiences with VAs.

The gap between the visioned human-like conversational experience and how people actually use them in the real world can be viewed as a "social-technical gap": The gap between what people want (*e.g.*, conversing naturally like talking to a human being) and what technologies can support (Ackerman 2000). The human-in-the-loop design may provide an option for better user experience with VAs. Prior work showed that it is possible for crowd-powered systems to be utilized in real time scenarios to enhance user experience (Bernstein et al. 2011). A great example is Evorus, a text-based crowd-artificial intelligence (AI) collaborative conversational agent, which provided quality responses with low latency compared to fully automated systems (Huang, Chang, and Bigham 2018). Another example is VizWiz, which utilized the crowd to successfully answer visual questions for blind people in near real-time (Bigham et al. 2010). However, there is one main concern about the use of crowd-powered architecture in a conversational setup, latency. It took nearly real time systems such as Evorus and VizWiz longer than 30 seconds to respond to the users. While response time of VAs is claimed to be around 2 seconds (Dgit 2018).

Limitations Of Automated VAs

Expectation for VAs is for users to converse with them as natural as with an actual human being. However, existing VAs on the market, such as Amazon Alexa, Apple Siri, and Google Assistant still focused on what VAs can "do" rather than how "human-like" they are. The utilitarian approach of VA design also echoed with how users interact with them on a daily basis. Aside from the request and response framework reported by Porcheron et al. (2018), Ammari et al. (2019) also investigated the use of Amazon Alexa and Google Home devices in domestic scenarios. Both qualitative interview sessions with VA users and qualitative ac-

tivity log analysis of VA usages were conducted. They concluded that music, hands-free search, and Internet of Things (IoT) control are the most used categories by the users. Clark et al. (2019) also pointed out the tendency of current interaction between users and VAs being limited and task oriented. The purposes of human conversations can be categorized roughly into two categories, transactional or social, and the two types can coexist in natural scenarios. Transactional conversation is goal driven and both sides of the know what objective should the conversation achieve. Social conversation is aimed to maintain or strengthen the bond between participants. It was suggested that current human-VA conversations are transactional and utilitarian while lacking in social features (Clark et al. 2019).

Human-In-The-Loop To Fill Social-Technical Gap

Adding humans into the loop of automated systems to compensate where machines fall short can be powerful and further fill the social-technical gap of VA usage. While traditional crowdsourcing setups do not support real time applications, Bernstein et al. (2011) provided a framework allowing workers in crowd-powered systems to be recruited for real time applications (within 2 seconds). Huang, Chang, and Bigham (2018) proposed a text-based crowd-Artificial Intelligence (AI) collaboration conversational agent - Evorus. Rather than waiting for fully automated systems to take over, Evorus provided a smooth transition toward automation overtime by utilizing collaboration between humans (dubbed as workers) and AIs. Workers were recruited real-time on MTurk and asked to provide responses via a chat room like interface. AIs in the system contribute by learning to choose chatbots that fit the scenario, utilizing previous responses, learning to vote on the responses (automatic voting) and providing responses at the same time. Evorus demonstrated a flexible framework that can be integrated with different services as the chatbot options are unlimited. The introduction of automated components did not make conversation quality worse while providing a more human-like conversational experience for users. Evorus showed that collaboration between workers and AIs for conversation purposes is possible and the participation of the crowd ensures quality of conversation from the start.

Pilot Study - EchoPal

A prototype system of integrating human in the loop for VAs, EchoPal, is designed and studied. EchoPal utilized Amazon Echo devices and Alexa to achieve its functionalities. When a user talks to the system, Echo records the audio and turns the speech into text through the built-in automatic speech recognition (ASR) system. The transcribed text is then sent to the back end of EchoPal and presented to the worker as a message in the worker interface, where the worker can see not only the transcribed message but also a set of possible alternative transcriptions generated by the system. Furthermore, EchoPal also automatically generates and presents a list of suggested responses to the worker. As each alternative transcription has its own suggested responses, the worker can click on the transcription to switch between them. Search support sites, such as Google Search or Google Weather, are shown on the right. A time constraint of responding within 25 seconds for the worker is set. Worker response is then be sent back to the Echo device, where a built-in text-to-speech (TTS) system reads out the message to the user. Users can trigger EchoPal by saying Alexa, open EchoPal and then start to talk. To leave EchoPal, saying cancel or stop will cause Alexa to close the application.

An in-lab user study consisting of 8 user and worker pairs showed that users in general considered the conversational quality to be better than Alexa socialbots. However the latency for Amazon Echo to respond to an user utterance was longer than 15 seconds, even under an optimal and simplest setup, where one user was conversing with one worker via an Amazon Echo synchronously. 17 participants were paired into 8 user/worker pairs for the pilot study. Users were asked to interact with existing Alexa bots first, then chat freely with EchoPal for 20 minutes. A total of 350 turns of conversation was produced. The latency is defined as the difference of the timestamps of the user message and the worker response in EchoPal, which is approximately the time from the user finishing talking to starting getting response. The average latency is **17.68 seconds** (SD = 6.29). It is also shown that the default response gives systematically lower latency, and typing on average takes around 20 seconds. On the other hand, suggested responses provided by Cleverbot was only selected 11 times, which did not tell much about their usefulness. Post study surveys from both users and workers were collected to reflect the user experience of participants.

From the users' perspective, EchoPal was considered to perform better than automated Alexa social bots in general. Direct quality rating of the system was also high in terms of Likert scale scoring. While users considered the response and content quality to be great, the latency of EchoPal was just too long. Aside from the overall experience, one main and common problem users face was the cut-off of conversations. When the user paused too long between two words during a speech, the device sometimes stopped listening and sent out the utterances even when users have not finished speaking yet.

For workers, the main challenge is the short response time. The 25 seconds respond time was set to accommodate the response time of voice-enabled devices. However, workers were not able to respond properly due to the short period. Even if workers managed to answer within the required time, brief responses were given under this circumstance, which demonstrated a clear trade-off between response quality and speed. Usefulness of worker-support features are also reported, with quick access to web search and default response buttons being the most useful designs. Echoing the users' positive feedback on the conversational quality, workers also considered human responses to be better than that automated agents. As human workers are better at guessing or even predicting user intents by making sense out of contextual information and prior conversations and can therefore produce better responses. Workers also provided ideas that can improve the system from a worker's perspective. Features such

as improved web searchers and embedded search engines are suggested to speed up the workflow.

Challenges And Goal

While EchoPal demonstrated the possibilities of human-in-the-loop VAs, the latency is the major concern regarding the usability. While most voice-enabled devices respond to users' requests in an average of 1.9 to 2.3 seconds (Dgit 2018), the average latency of a response from the deployed version of Chorus and Evorus is longer than 30 seconds. The average response time per question for Vizwiz is 36 seconds; and the average latency of Zensors++ is 120 seconds. These projects focused on recruiting workers quickly (Bernstein et al. 2011; Bigham et al. 2010) to make crowd-powered components fast enough to be useful in many areas, but not yet fast enough to fulfill the extremely short turnaround time embodied by voice-enabled devices.

This proposal aims to augment the capability, robustness, and usability of the already-popular voice-enabled devices by incorporating human intelligence into the loop more effectively. To tackle the main challenge set forth by such devices' extremely short response time, a novel direction will be explored: allowing **asynchronous conversations** between voice-enabled devices and the user, which will free such devices from the traditional synchronous interaction modality. This proposal imagines a human-in-the-loop asynchronous conversational partner, ***ThinkerPal***, which is designed to engage in "asynchronous" dialogue with the user. The user initiates a discussion with ThinkerPal by simply talking to their voice-enabled devices, but without expecting an immediate response. Behind the scenes, ThinkerPal employs both AI and human computation workflows to explore the question, collect relevant information, brainstorm and reason follow-up topics, and decide how to respond. When ready, the system reaches out to the user via multiple channels, including voice-enabled devices, emails, instant messaging applications, or even social media, to continue the conversation. The user can respond to the system via their preferred channel whenever they want, and ThinkerPal will use the response to develop the conversation further.

Asynchronous conversations generate many interesting research questions: How to create a workflow that can effectively respond to random open questions? When and why should the system reach out to the user? Which communication channels should be used for this question at this time? Can the system learn to judge when an asynchronous conversation is preferred over an instant, automated response? What are the benefits of using voice-enabled devices in such asynchronous conversations?

Research Objectives

Distinguish Complex And Simple Tasks

To better utilize asynchronous conversation, it is crucial to tell when the asynchronous portion should be engaged compared to conversations that can be done quickly, such as weather and time checking. One direction to approach the problem is to distinguish between simple and complex tasks and further tell which conversations take longer response time to complete. Being able to tell the complexity level of different conversations allows easier ones to be solved quick and fast while reserving the complex ones for longer processing time. When the system has the ability to detect the complexity level, it provides an entry point for engaging ThinkerPal. For example, "What is the weather now in New York?", "What time is it?" and "Play music by The Beatles." These operations can be completed by existing automated VAs without users waiting. Conversations like "Can you recommend a restaurant?" and "Organize trip to San Francisco." take more time to respond, these bigger topics often require back and forth discussion, further exploration or multiple steps in collaboration with the user instead of populating a straight forward answer. Indicators can be utilized to gauge the complexity of an conversation. For example, how well the question is defined, how difficult it is to find answers, if an objectively correct answer exist, and how frequently the answer changes. Prior studies have tried to predict some of these factors (Gurari and Grauman 2017) and predict the time it takes to accomplish a microtask (Saito et al. 2019). Zhang et al. (2021) provided another perspective by listing a series of tasks followed by the respective subtasks required to finish the main ones. The amount of subtasks can be considered as a indicator of the complexity of the main task. These technologies can be explored to better fit ThinkerPal and its usage.

Scenarios And Channels That Fit

While ThinkerPal can provide an alternative medium for interacting with VAs, under what circumstances it can be best utilized is still unclear. One advantage of VAs is their availability, different from reaching out to a real human being via phone call or social media, the popularity of voice enabled devices allow users to have easy access to VAs. On the other hand, the asynchronous aspect of the interaction allows the system to reach out to users instead of only users providing inputs. If granted access to different channels outside voice-enabled devices, such as text message, email, social media, and other messaging applications, ThinkerPal will be able to help users explore the problem space in a collaborative manner. Users can initiate conversation with their voice-enabled devices, for example, "Tell me more about the fair that is going on now around central Pennsylvania." ThinkerPal will first reply with the basic information about the ongoing event such as the location, time of operation, and ticket price. At the same time, detailed information such as description, web page, and reviews of the event will be organized and send to the user afterwards via email or messaging applications. Aside from only forwarding information, ThinkerPal can reach out to users asking about their preferences on the type of transportation or food. The design leads to interesting research questions such as: How and When should ThinkerPal reach out to the user? Which channel(s) should be used for each communication? Furthermore, with more time to "think" about the conversation, it is also possible for users to have more open-ended conversations with ThinkerPal, which is not available for current fully automated VAs.

References

Ackerman, M. S. 2000. The intellectual challenge of CSCW: The gap between social requirements and technical feasibility. *Human–Computer Interaction* 15(2-3): 179–203.

Ammari, T.; Kaye, J.; Tsai, J. Y.; and Bentley, F. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput. Hum. Interact.* 26(3): 17–1.

Bernstein, M. S.; Brandt, J.; Miller, R. C.; and Karger, D. R. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, 33–42.

Bigham, J. P.; Jayant, C.; Ji, H.; Little, G.; Miller, A.; Miller, R. C.; Miller, R.; Tatarowicz, A.; White, B.; White, S.; et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, 333–342.

Bohn, D. 2019. Amazon says 100 million Alexa devices have been sold - What's next? https://www.theverge.com/2019/1/4/18168565/amazon-alexa-devices-how-many-sold-number-100-million-dave-limp. Last checked on May 06, 2021.

Clark, L.; Pantidi, N.; Cooney, O.; Doyle, P.; Garaialde, D.; Edwards, J.; Spillane, B.; Gilmartin, E.; Murad, C.; Munteanu, C.; et al. 2019. What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Dgit. 2018. What's the best smart speaker? Apple HomePod vs Google Home vs Amazon Echo. https://dgit.com/apple-homepod-vs-google-home-vs-amazon-echo-53296/. Last checked on May 06, 2021.

Gurari, D.; and Grauman, K. 2017. CrowdVerge: Predicting If People Will Agree on the Answer to a Visual Question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, 3511–3522. New York, NY, USA: Association for Computing Machinery. ISBN 9781450346559. doi:10.1145/3025453.3025781. URL https://doi.org/10.1145/3025453.3025781.

Huang, T.-H.; Chang, J. C.; and Bigham, J. P. 2018. Evorus: A crowd-powered conversational assistant built to automate itself over time. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13.

Luger, E.; and Sellen, A. 2016. " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 5286–5297.

Porcheron, M.; Fischer, J. E.; Reeves, S.; and Sharples, S. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*, 1–12.

Saito, S.; Chiang, C.-W.; Savage, S.; Nakano, T.; Kobayashi, T.; and Bigham, J. 2019. Predicting the Working Time of Microtasks Based on Workers' Perception of Prediction Errors. *Human Computation* 6(1): 192–219.

Zhang, Y.; Jauhar, S. K.; Kiseleva, J.; White, R.; and Roth, D. 2021. Learning to Decompose and Organize Complex Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.