Too Slow to Be Useful?
On Incorporating Humans in the Loop of Smart Speakers

Shih-Hong Huang, Chieh-Yang Huang, Yuxin Deng,
Hua Shen, Szu-Chi Kuan, Ting-Hao 'Kenneth' Huang

College of Information Sciences and Technology, Pennsylvania State University
201 Old Main, University Park, PA 16802, USA
{szh277,chiehyang,ybd5063,huashen218,sbk5672,txh710}@psu.edu

Abstract

Real-time crowd-powered systems, such as Chorus/Evorus, and VizWiz, have shown how incorporating humans into automated systems can supplement where the automated solutions fall short. However, one bottleneck of applying such architectures to more scenarios is the longer latency of including humans in the loop of automated systems. For applications with hard constraints in turnaround times, human-operated components' longer latency and larger speed variation seem to be apparent deal breakers. This paper explicates and quantifies these limitations by using a human-powered text-based backend to hold conversations with users through a voice-only smart speaker. Smart speakers must respond to users' requests within seconds, so the workers behind the scenes only have a few seconds to compose answers. We measured the end-to-end system latency and the conversation quality with eight pairs of participants, showing the challenges and superiority of such systems.

Introduction

Real-time crowd-powered systems have achieved success in reducing the gaps between human agents and automated solutions. For example, Chorus (Huang et al. 2016) and Evorus (Huang, Chang, and Bigham 2018) used the crowd to hold sophisticated long conversations, VizWiz utilized crowd workers to answer visual questions quickly for blind people (Bigham et al. 2010), and Zensors (Laput et al. 2015) and Zensors++ (Guo et al. 2018) used the crowd to monitor running video feeds. However, despite much effort to speed up such systems, humans are, in many cases, still slower than computers. Many existing automated systems and their infrastructures were built with the assumption that all the internal components, when working properly, have short turnaround times. This reality makes realizing the vision of crowd-powered systems extra challenging. Taking modern smart speakers or voice-enabled devices for example, Amazon's Echo devices, Google Assistant, Apple Siri, and Samsung Bixby respond to users' requests within around 0.77 to 3.09 seconds (Koni et al. 2021). This range of turnaround time is too short for most crowd-powered systems. The average latency of a response from the deployed version of Chorus and Evorus was longer than 30 seconds; average response time per question for Vizwiz as 36 seconds and the latency of Zensors++ is 120 seconds.

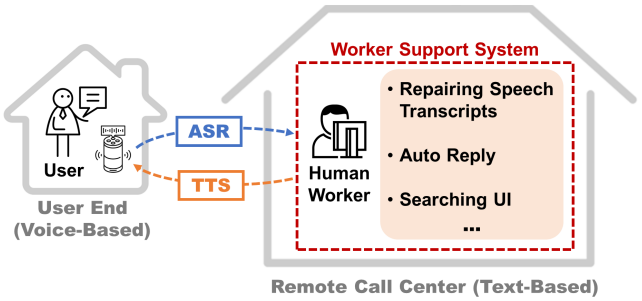

Figure 1: System overview of ECHOPAL.

This paper examines and explicates the challenges of incorporating a human-in-the-loop architecture into an already-existing and widely used technological infrastructure. In particular, we built **ECHOPAL**, a prototype system that allows a human worker to converse with the user synchronously via an Amazon's Echo device.[1] ECHOPAL works with the full set of Amazon Alexa's infrastructure, including the official Alexa Skills Kit and an Echo device. To incorporate human workers in the loop and to allow free conversation, we engineered a custom back end for ECHOPAL.

ECHOPAL System. Figure 1 overviews the system. 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 backend of ECHOPAL and presented to the human 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 uses Cleverbot (www.cleverbot.com) to generate suggested responses for the worker. We enforced a time constraint of 25 seconds for the worker to produce each response. The worker's response is then be sent back to the Echo device, where a built-in text-to-speech (TTS) function reads out the message to the user.

ECHOPAL inherits two real-world constraints of many

[1] ECHOPAL demo video: https://youtu.be/iMDsX52VWGY

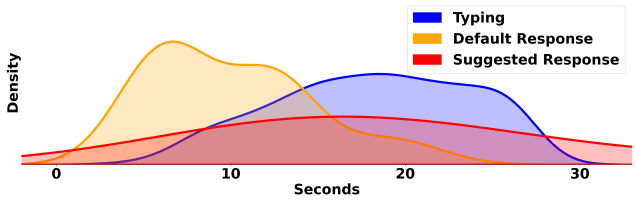

Figure 2: KDE distribution of the latency. From the distribution, we found that *(i)* default response gives systematically lower latency; *(ii)* typing normally takes around 20 seconds.

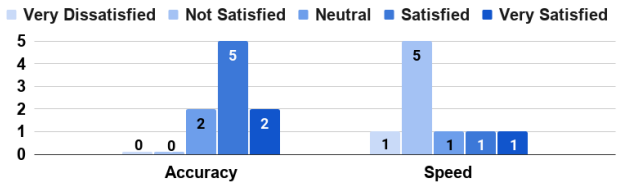

Figure 3: Users rating on "Response **accuracy** to your questions" and "Response **speed** to your questions". Most users agree that ECHOPAL gives accurate response but is slow.

commercial smart speakers that are particularly challenging for human-in-the-loop systems: *(i)* the extremely short response time, and *(ii)* the purely text-based back end. Alexa Skills Kit enforced a maximum processing time for each turn of speech received from the users; an Echo device also had a maximum listening time once it finished reading out the responses. Furthermore, most smart speakers do not allow API access to voice recordings to preserve user privacy. Namely, in ECHOPAL, the worker cannot hear the user's voice. The worker can only rely on the transcriptions that are automatically compiled, which could be imperfect or misleading, to communicate with users.

Experimental Results

Experimental Setup. An IRB-approved in-lab user study with 17 participants were conducted to evaluate ECHOPAL. The user study were conducted in pairs and involved two participants at the same time, one as the user (N=9) and the other as the worker (N=8).[2] User and worker were assigned to two different rooms so direct communication was prohibited. At the user station, the user first chatted with two Alexa Prize Socialbots (Gabriel et al. 2020) for five minutes each. It helped users get familiar with the interaction pattern of Echo devices and can be treated as baseline in the study. After interacting with Alexa Prize Socialbots, users were asked to chat with ECHOPAL for twenty minutes. Topics covered chit-chat and open-domain questions, but both sides were encouraged to chat freely without any kind of constraint on the topic. At the worker station, one author first walked the participant through how to use the interface and had them familiarize themselves with the system for ten minutes. This participant will then serve as the worker behind ECHOPAL for twenty minutes using all the provided functionalities. After the chat was over, users and workers needed to fill up a questionnaire as an assessment of their experience with ECHOPAL separately.

End-to-End System Latency. Eight pairs of participants produced a total of 350 turns of conversation. The latency is defined as the duration between the time when the system received the user message and when the system received the worker response. The average latency is **17.68 seconds** (SD = 6.29). Among the 350 worker responses, 12 (3.4%) were responses generated by Cleverbot with an average latency of 22.5 seconds (SD = 13.16); and 27 (7.7%) were the default responses sent by simply clicking pre-defined buttons such as "Yes, I agree.", with an average latency of 10.04 seconds (SD = 4.48). Figure 2 shows that the default response gives systematically lower latency, and typing on average takes around 20 seconds.

User Assessment of Conversation With ECHOPAL. In the post-study questionnaire. We asked users to rate their satisfaction levels on the *(i)* accuracy and *(ii)* speed using a 5-point Likert scale, from "Very Dissatisfied" (1) to "Very satisfied" (5). The average score of response quality is 4.0, but that of response speed is only 2.56 (Figure 3). Namely, **users were satisfied with the quality of the conversation, but strongly dissatisfied with the turnaround time of ECHOPAL.** We also asked the users to directly rate the overall quality of the system on a 5-point Likert scale, from "Low Quality" (1) to "High Quality" (5). The average score is 3.67. We also asked users to compare ECHOPAL to Alexa Socialbots with a 5-point Likert scale, from "Much Worse" (1) to "Much Better" (5). The average score is 3.56 with only one user rated ECHOPAL as "Worse".

Cut-offs of Conversations. A common problem that occurred in the user study is the cut-offs of conversations. Almost all users mentioned this problem in the survey. Cut-offs mostly happen in the middle of a sentence, often caused by a longer pause between two words, which the device will consider the speech to be over. Another kind of cut-off happen at the start of a speech. After Alexa finish reading out the message from the worker, it will enter the listening mode and start to transcribe. If the user does not start to talk fast enough, it will either try to transcribe any background noise picked up by the device or turn off the skill completely.

Conclusion and Future Work

This paper introduces ECHOPAL, a prototype system that allows a human worker to converse with the user synchronously via an Amazon's Echo device without access to the user's voice recordings. In our user study, many users expressed their frustration of the long latency of the system. We also observed that one of the main challenges for users is the cut-offs of ongoing conversations; and the main challenge for workers is the extremely short response time. Our work explores the possibilities and challenges of human-in-the-loop smart speakers, informing the designs of future systems facing various real-world constraints.

[2]Three participants signed up as a group and therefore work as a team of two users conversing with one worker.

Acknowledgements

We thank Ming-Ju Li, Tiffany Knearem, and Sooyeon Lee for their valuable help and feedback. We thank the participants who participated in our studies. We also thank the anonymous reviewers for their constructive feedback and comments.

References

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. ACM.

Gabriel, R.; Liu, Y.; Gottardi, A.; Eric, M.; Khatri, A.; Chadha, A.; Chen, Q.; Hedayatnia, B.; Rajan, P.; Binici, A.; et al. 2020. Further advances in open domain dialog systems in the third alexa prize socialbot grand challenge. *Alexa Prize Proceedings*, 3.

Guo, A.; Jain, A.; Ghose, S.; Laput, G.; Harrison, C.; and Bigham, J. P. 2018. Crowd-AI Camera Sensing in the Real World. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(3).

Huang, T.-H. K.; 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*, 295. ACM.

Huang, T.-H. K.; Lasecki, W. S.; Azaria, A.; and Bigham, J. P. 2016. " Is There Anything Else I Can Help You With?" Challenges in Deploying an On-Demand Crowd-Powered Conversational Agent. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

Koni, Y. J.; Al-Absi, M. A.; Saparmammedovich, S. A.; and Lee, H. J. 2021. AI-Based Voice Assistants Technology Comparison in Term of Conversational and Response Time. In Singh, M.; Kang, D.-K.; Lee, J.-H.; Tiwary, U. S.; Singh, D.; and Chung, W.-Y., eds., *Intelligent Human Computer Interaction*, 370–379. Cham: Springer International Publishing. ISBN 978-3-030-68452-5.

Laput, G.; Lasecki, W. S.; Wiese, J.; Xiao, R.; Bigham, J. P.; and Harrison, C. 2015. Zensors: Adaptive, rapidly deployable, human-intelligent sensor feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1935–1944. ACM.