Understanding Differences between Heavy Users and Light Users in Difficulties with Voice User Interfaces

Hyunhoon Jung
Clova AI, NAVER Corp.
hyunhoon.j@navercorp.com

Hyeji Kim
Clova AI, NAVER Corp.
kim.hyeji@navercorp.com

Jung-Woo Ha
Clova AI, NAVER Corp.
jungwoo.ha@navercorp.com

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract
Voice user interfaces (VUIs) are growing in popularity. At this stage of VUIs adoption, distinctions between heavy users and light users are becoming emerging challenge. Some studies have focused on investigating how general users interact with VUIs; however few studies have focused solely on the differences in VUIs use between heavy and light users. In this paper, we conduct user study using our new restaurant reservation VUI, AiCall, to explore what kind of difficulties those two groups are facing and what are the differences between them. We found out that 1) heavy users could identify more diverse difficulty types than light users; 2) the types of difficulties that affect each group of users are different, and 3) in particular, the repetition of agent utterance was considered the most inconvenient by heavy users. Based on these findings, we discuss the VUI design and development considerations to satisfy both groups of users.

Author Keywords
Virtual Agent; Voice User Interface (VUI); Conversational Agent (CA); Conversational User interface (CUI)

CCS Concepts
•Human-centered computing → Natural language interfaces;

INTRODUCTION

As voice user interfaces (VUIs) become more popular, people can access such interfaces easily in various devices. The popularity of smart speakers by Amazon and Google enables people to be familiar with voice technology and VUIs. Some consumer reports[1] showed that the market related to voice technology and VUIs is entering an early majority phase globally. Under this circumstance, there are a number of heavy users who use VUIs in their daily lives very frequently, while there is still be a group of so called light users, who do not actively adopt VUIs widely in their lives. Understanding both groups of users is important in the early majority stage to improve VUIs from design to voice technology development by satisfying both groups.

	Rate	Recency
Heavy	>3 times	<1 week
Light	<2 times	>1 month

Table 1: Frequency and recency thresholds for the study. "Rate" column is based on monthly usage.

Although recent studies have focused on how users interact with VUIs from various perspectives [12, 5, 10], very few studies have investigated the differences in VUIs usage between the different groups of users in terms of VUIs adoption. Some studies have investigated interactions between users and VUIs, especially focusing on interaction failure. For instance, Myers et al. [11] analyzed the obstacles and users' tactic to overcome them while using VUIs. Beneteau et al. [1] focused on communication breakdown between families and VUIs. However, these studies did not cover comparing the use of VUIs by different user type in terms of interaction failure.

To understand and compare the difficulties that heavy users and light users face while using VUIs, we conducted user study by using our phone-based restaurant reservation VUI, AiCall. We recruited 60 participants, 30 heavy users and 30 of light users. In order to recruit the users suitable for our experiments, a pre-screening questionnaire was emailed to a potential participant pool. Participants were instructed to call four times a day with various types of task that AiCall could handle, and they were asked to answer a questionnaire. After that, we analyzed the self-reported qualitative data, especially focusing on difficulties, and we also analyzed quantitative data from users by using the IVR usability questionnaire, Speech User Interface Service Quality (SUISQ-MR) [9, 8, 4].

In the user study analysis, we found that heavy users could identify more types of difficulties than light users (heavy: 15, light: 11). Furthermore, the types of difficulties that affected each group of users were slightly different; heavy users reported that repetition of system utterance was especially inconvenient for them. Based on these results, we discuss design considerations and voice technology improvement strategies to make better VUIs to satisfy both heavy users and light users.

METHODOLOGY

We analyzed data from AiCall, our restaurant reservation voice agent that has not yet been released publicly.

AiCall

AiCall is a phone-based AI agent service, which helps human workers in a call center via VUI communication with customers. In aspects of functional technologies, AiCall is an integration of automatic speech recognition (ASR) [2], natural language understanding (NLU) [3], and speech synthesis [13]. Although AiCall can support diverse call center services, we performed user study on a restaurant reservation application. For convenience, AiCall service for restaurant reservation is referred to AiCall after this.

AiCall can handle four types of major restaurant reservation tasks as follows: 1) book a table, 2) change a reservation, 3) check reservation information, and 4) cancel a reserva-

[1] https://voicebot.ai/2018/10/19/phase-one-of-the-voice-assistant-era-is-over-long-live-phase-two/

Difficulty Types
D1: The agent repeated the same utterance too many times
D2: The agent didn't understand the exact meaning of what I said
D3: The system didn't seem to recognize what I said
D4: The voice of the agent was not natural
D5: The agent interrupted me while I was talking
D6: Sometimes it was hard to know what to say next
D7: The logical flow of the system was unnatural
D8: The system didn't seem to support multi-turn dialog
D9: It was unable to interrupt the agent
D10: The agent talked too much
D11: The amount of information was not sufficient
D12: The conversation flow was unnatural
D13: The agent spoke too fast
D14: The agent requested various information at once
D15: The agent responded slowly

Table 2: 15 difficulty types reported by both heavy and light users.

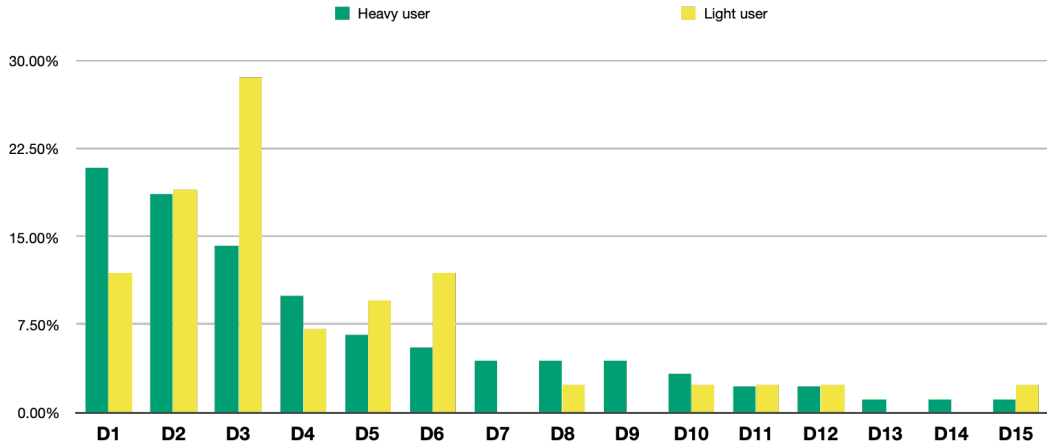

Figure 1: A comparison of difficulty types in heavy and light user groups. The percentage was calculated for each of the two users groups due to legal issue.

tion. On top of that, AiCall handles simple questions and answers about the target restaurant.

The persona designed for AiCall is a friendly, kind, and young restaurant service manager. When a user makes a phone call to AiCall, the voice agent leads the conversation with a greeting. Its conversation style is very similar to a kind human recipient in a casual restaurant. Given a request or question that AiCall cannot deal with, the request is escalated to the human service manager in the restaurant.

Participants

To explore the difficulties that both heavy users and light users experience, we recruited the equal numbers of participants for two groups for the user study. Because AiCall was not publicly released, we needed to recruit participants from among the employees in our company. Based on the definition of usage gauges in the earlier literature [7] (see Table 1), we developed a brief pre-screening questionnaire including demographic questions. After we collected the pre-screening questionnaire, we selected 30 heavy users (11 women and 19 men, mean age of 34.27, SD of 6.41) and 30 light users (10 women and 20 men, mean age of 32.73, SD of 5.71) for our user study.

Procedure and Analysis

To get natural interaction data from participants, a user study was conducted in the wild in one day. We sent an email to all participants to show how the tasks should be performed and how the questionnaire needed to be answered. Participants were asked to call AiCall four times to perform the four different main tasks. After finishing their final call, they needed to answer a questionnaire on Google Forms. The questionnaire included IVR usability evalua-

	Relative Values	
Type	*Heavy*	*Light*
D1	0.21	0.05
D2	0.19	0.09
D3	0.14	0.13
D4	0.10	0.03
D5	0.07	0.04
D6	0.05	0.05
D7	0.04	0.00
D8	0.04	0.01
D9	0.04	0.00
D10	0.03	0.01
D11	0.02	0.01
D12	0.02	0.01
D13	0.01	0.00
D14	0.01	0.00
D15	0.01	0.01
Total	**1.00**	**0.46**

Table 3: The reported values of each difficulty types from both heavy and light users. The results are relative values by normalizing with difficulty sum of heavy users due to legal issue.

tion questionnaire (SUISQ-MR) and open-ended questions about what kind of difficulties they suffered from while using AiCall. To further analyze interaction, their call logs were recorded and automatically transcribed by our in-house Call Session Viewer.

The collected qualitative data related to the difficulties reported from both participant groups were encoded and categorized by researchers based on grounded theory [6]. One researcher created the code scheme, and it was reviewed by another researcher. The code scheme agreed by all researchers was finally used for coding qualitative data.

RESULTS

From our user study, we collected qualitative data, which was self-reported by both groups, and quantitative data from the SUISQ-MR questionnaire. Through analyzing the qualitative data from user study, we categorized 15 major difficulties from both groups.

Difficulty Types

We categorized 15 difficulty types reported by both groups of participants (see Table 2).

The heavy user group identified more difficulty types

There were some difficulty types reported by the heavy user group that were not mentioned by the light user group, such as “The logical flow of the system was unnatural,” and “I was unable to interrupt the agent.” On the other hand, all difficulty types identified by light users were also reported by heavy users.

Difficulties affecting each group of users were different

The heavy user group reported that repeated system utterances were the problematic in using VUI. However, light user group reported this difficulty type was not the most problematic one (see Table 3). The ranking of difficulty types also differed between groups (see Figure 1). For instance, the top three items were the same for both groups but the rankings were different.

SUISQ Questionnaire Result

The Speech User Interface Service Quality (SUISQ) questionnaire is used to measure IVR usability [9]. We used 5-likert scale SUISQ Maximally Reduced (MR) version (SUISQ-MR) to reduce the burden of completing the questionnaire for participants. The questionnaire consists of 9 items and is divided into 4 factors: User goal orientation, Customer service behavior, Speech characteristics, and Verbosity.

Verbosity of the agent was problematic for heavy users

The SUISQ-MR results showed that verbosity of the agent affected heavy users more than light users (see Table 4). The mean scores of the verbosity from both groups of users were significantly different ($p < .005$). Three questions related to verbosity of the agent are as follows: 1) I felt like I had to wait too long for the system to stop talking so that I could respond, 2) The messages were repetitive, 3) The system was too talkative.

DISCUSSION

Our goal is to understand what types of difficulty both heavy users and light users face while using VUIs. In our user study, we found 15 types of difficulty and analyzed the related differences between heavy and light users. Based on these findings, we discuss implications for designing and developing more improved VUIs for both groups of users.

Heavy Users are Active and Wise

Based on difficulty type D1 (“The agent repeated the same utterances too many times”) and D9 (“It was unable to interrupt the agent”), we found that heavy users tended to control VUIs as they wanted. Specifically, type D1 and the

Table 4: Two-sample t-test for difference of means.

Factors	User Group	Difference of Means	SD	t	p
User Goal Orientation	Heavy Light	0.17	0.83 0.82	0.785	.436
Customer Service Behavior	Heavy Light	-0.03	0.70 0.76	-0.176	.861
Speech Characteristics	Heavy Light	0.19	0.85 1.10	0.723	.473
Verbosity	Heavy Light	0.58	0.81 0.68	3.000	.004**

result of the SUISQ questionnaire showed that they were not willing to patiently wait for repetitive messages because they could quickly understand that this repetition would reduce the efficiency of interaction. Participant 07 in the heavy user group reported that "I cannot stand repetitive messages: I don't think the confirmation message should be repeated. Just one time is enough. If it was designed that way, I thought I should be able to interrupt unnecessary confirmation messages from the agent." This report was also related to type D9. It is apparent that heavy users already knew how to use the VUIs, so they could identify the failure based on previous experience. At the same time, they also knew how to escape that failure. If their tactics of avoiding the failure do not work, they might perceive that failure as more problematic.

Light Users Still Need Some Guidance

From difficulty type D3 ("The system didn't seem to recognize what I said"), and type D6 ("Sometimes it was hard to know what to say next"), we found out that light users need a guide to use VUIs more confidently. Of all the difficulty types of light users, D3 (reported by 28.57 percent of users) stood out. Based on the result, we believed that light users might not be familiar with the known limitations of speech recognition of VUIs. Participant 22 of the light users reported that "The system did not understand whenever I said '28th, this week.' I tried another date this week. It worked perfectly, but I thought the system misrecognized '28th' again and again. My pronunciation could be problematic." The AiCall VUI used in our user study did not provide a paraphrase request feature when the system was not sure of what the user said. The absence of this type of feature could make users confused when speech recognition error happened. For the light user group, an explicit guide could help them use VUIs more confidently. Unlike type D3, D6 related directly the absence of conversation guide while using VUIs. These reports showed that the reason light users faced these types of difficulty is because VUIs did not explicitly guide them.

How to Satisfy Both User Groups

The reasons that lead the difficulties for the heavy users and the light users were different from each other. To improve performance, VUIs should satisfy these different

group of users' needs simultaneously. Accordingly, VUI design and development should tackle the users' problem more strategically. First of all, VUI designers should design not just each sequence in conversation but whole conversation space. Some utterances may be appropriate for specific dialog scenes, but these sets of utterances could cause repetition in course of a whole conversation. Secondly, on the other hand, designers should be cautious to reduce repetitive utterances. The excess reduction of semantically similar utterances could cause a lack of conversation guide for the users. This type of problem could affect the light user group directly. Thirdly, VUI development must be improved based on usage data. Because of the characteristics of VUIs (i.e., using human language for interaction), there could be unknown problems that are not considered before the test. Furthermore, as seen from the results of the user study, some problems needed to be solved urgently to satisfy both groups of users. These considerations for both VUI design and development could help to improve VUIs and also could increase both user groups' satisfaction with the interfaces.

Limitations and Future Work

There are some limitations to our study. First, our study is an in-company user study. Both user groups could represent heavy users and light users respectively; however, there may be difference from the same cohort from outside the company. Second, we conducted a user study in the limited context of usage in a single day. Both groups of users could report new types of difficulties after the long-term usage.

In the future work, we are going to analyze the conversation log of this user study. This analysis could help to understand what tactics each group of users take for escaping or resolving difficulties. Furthermore, after the public release of AiCall, we plan to conduct a user study outside the company with the agreement of real users. From the real users' data, we hope there could be other types of difficulties identified.

CONCLUSION

This study explores the difficulties in using VUIs for both heavy users and light users. We categorized difficulties into 15 types and compared them between different user groups. We discussed the characteristics of each group of users based on the reported difficulty types and considerations for VUI design and development. Our contribution to VUI design and development communities are as follows: 1) In the early majority phase of VUI adoption, we need to understand the difficulties of both heavy users and light users, 2) VUI designers should adopt the point of view of conversation space design to reduce the difficulties for both user types, and 3) VUI developers should actively exploit usage data to set the priorities of problem solving to satisfy both groups of users.

REFERENCES

[1] Erin Beneteau, Olivia K Richards, Mingrui Zhang, Julie A Kientz, Jason Yip, and Alexis Hiniker. 2019. Communication breakdowns between families and Alexa. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 243.

[2] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4960–4964.

[3] Qian Chen, Zhu Zhuo, and Wen Wang. 2019. Bert for joint intent classification and slot filling. *arXiv preprint*

arXiv:1902.10909 (2019).

[4] Leigh Clark, Phillip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, and Benjamin Cowan. 2018. The State of Speech in HCI: Trends, Themes and Challenges. *arXiv preprint arXiv:1810.06828* (2018).

[5] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Christine Murad, Cosmin Munteanu, Vincent Wade, and others. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. *arXiv preprint arXiv:1901.06525* (2019).

[6] Juliet Corbin and Anselm Strauss. 2014. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications.

[7] Randy Allen Harris. 2004. *Voice interaction design: crafting the new conversational speech systems*. Elsevier.

[8] A Baki Kocaballi, Liliana Laranjo, and Enrico Coiera. 2018. Measuring user experience in conversational interfaces: a comparison of six questionnaires. In *Proc. 32nd British Computer Society Human Computer Interaction Conference, Belfast, Northern Ireland*.

[9] James R Lewis. 2016. Standardized questionnaires for voice interaction design. *Voice Interaction Design* 1, 1 (2016).

[10] Ewa Luger and Abigail Sellen. 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*. ACM, 5286–5297.

[11] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 6.

[12] Martin Porcheron, Joel E Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. ACM, 640.

[13] Eunwoo Song, Kyungguen Byun, and Hong-Goo Kang. 2019. Excitnet vocoder: A neural excitation model for parametric speech synthesis systems. In *2019 27th European Signal Processing Conference (EUSIPCO)*. IEEE, 1–5.