Evaluating Voice Applications by User-Aware Design Guidelines Using an Automatic Voice Crawler

Xu Han
University of Colorado Boulder
Boulder, USA
xu.han-1@colorado.edu

Tom Yeh
University of Colorado Boulder
Boulder, USA
tom.yeh@colorado.edu

ABSTRACT

Adaptive voice applications supported by conversational agents (CAs) are increasingly popular (i.e., Alexa Skills and Google Home Actions). However, much work remains in the area of voice interaction evaluation, especially in terms of user-awareness. In our study, we developed a voice skill crawler to collect responses from the 100 most popular Alexa skills within 10 categories. We then evaluated these responses to assess their compliance to three user-aware design guidelines published by Amazon. Our findings show that more than 50% of voice applications do not follow some of these guidelines and variation in guideline compliance across skill categories exists. As voice interaction continues to increase in consumer settings, our crawler can evaluate CA-based voice applications with high efficiency and scalability.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; *Interactive systems and tools*; **Systems and tools for interaction design**.

KEYWORDS

conversational agents; voice applications; user-awareness evaluation;

ACM Reference Format:
Xu Han and Tom Yeh. 2019. Evaluating Voice Applications by User-Aware Design Guidelines Using an Automatic Voice Crawler. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019* , 4 pages.

1 INTRODUCTION AND MOTIVATION

Voice-powered conversational agent (CA) devices have recently achieved significant commercial success. In the U.S.A., 47.3 million (19.7% of) households now own CA devices (March 2018), an increase from less than 1% two years ago [6]. Amazon's Echo series devices make up 71.9% of the market, followed by Google's devices with 18.4% [6].

One key characteristic that makes this new generation of CA devices adaptive is their API platform for third-party developers. Here, developers design and build voice applications and publish them on a marketplace with the potential to reach millions of users. Amazon's Alexa skills [3] and Google's Home Actions [4] are the two most popular examples. Yet, many third-party developers may not have prior experiences in designing and building voice applications, especially in terms of user-awareness. A well-designed user-aware voice application should adapt its interaction mode to different users and satisfy their individual needs. To help educate developers, Amazon and Google have published design guidelines [1, 8] to establish a set of design practices a voice application should try to comply with. These official design guidelines cover a variety of topics ranging from how to clearly communicate the purpose of a voice application to users to how to design a natural and adaptive interaction flow.

Despite the existence of official design guidelines, the quality of CA-based voice applications in terms of user-awareness varies widely. An example of a highly-rated (4.9 out of 5 stars based on 3209 user reviews) Alexa skill is *Would You Rather for Family*. This skill is an interactive Q&A game that exhibits several user-aware design features following Amazon's guidelines, including remembering where the last interaction ends and giving a personalized opening prompt to users. In contrast is the skill *AccuWeather*. This skill's average rating is low—2.2 out of 5 stars based on 182 user reviews. Within a user interaction, the skill's design violates several user-aware design guidelines, such as handling errors properly. Users also complain about these violations in their reviews.

At the same time, although several user experience evaluation methodologies have been adopted on CAs and voice applications, efficient ones are still lacking. Traditional usability studies are useful in gathering feedback and conducting evaluation analysis on CAs [7], longitudinal studies is another effective methodology for elucidating scenarios and situations involving the use of CAs [2]. In [7], researchers interviewed 14 users of CAs to understand the factors affecting everyday use. [2] deployed lab-based usability studies by developing CA prototypes of varying fidelity. However, by March 2018, more than 30,000 Alexa voice skills [5] have been published by thousands of third-party developers. The large number of voice applications require much more efficient evaluation methodologies, where traditional ones like usability studies and longitudinal studies cannot satisfy the needs.

The variability in user-aware design quality among third-party voice applications and the lack of efficient evaluation methods inspire these research questions:

(1) What is the current state of CA-based voice applications following (or violating) the user-aware design guidelines?
(2) How can we efficiently evaluate CA-based voice applications in terms of the user-aware design guidelines?

To study these questions, we focused on Alexa skills, and selected 100 most popular Alexa skills from ten categories. We developed a voice skill crawler which could efficiently collect responses from a wide variety of CA-based voice applications with differing input

IUI Workshops'19, March 20, 2019, Los Angeles, USA

commands. We then analyzed the collected responses to determine whether or not particular guidelines were followed. Regarding the first research question, we focused on the three design guidelines which are most relevant to user-awareness and assessed compliance. These three design guidelines are: A skill needs to memorize a user's previous interaction mode to provide more personalized service; a skill needs to adaptively re-prompt users to continue the interaction when it receives no input; and a skill needs to reword the re-prompt messages with more detailed information based on previous personified interaction. The key findings show that more than 50% of voice applications do not follow some of these guidelines and variation in guideline compliance across skill categories exists. Regarding the second research question, we argue that our voice skill crawler is a research prototype with great potential to do efficient user-awareness evaluation on CA-based voice applications.

2 VOICE CRAWLER DESIGN

We developed a crawler to automate the task of collecting a large sample of skills' responses to voice commands. This crawler follows the basic Alexa interaction mode of "open-command-stop" to initiate the skill and exit. As indicated in Figure 1, this crawler simulates the voice interaction between users and Alexa devices.

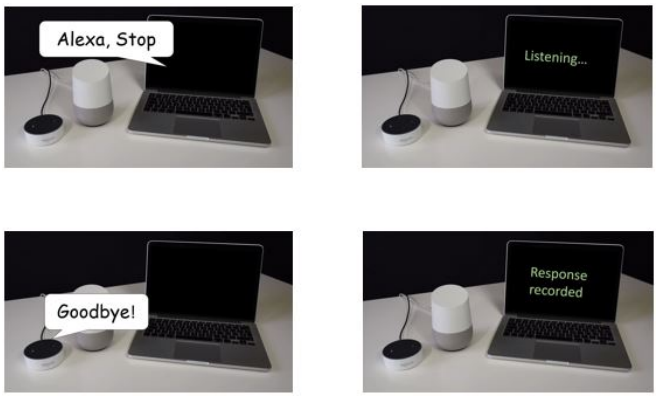

Figure 1: Working Process of Our Voice Crawler

The crawler uses a human voice generated by Google's Text-to-Speech package in Python[1] to speak a command to a skill. It then listens to the response using a speech recognition package [2]. Finally, it saves the responses in a file for further analysis. The crawler iterated through our sample of 100 skills to perform this sequence for each skill. The data collection process is described pragmatically as Algorithm 1.

Algorithm 1 Collect Responses to m Commands by n Skills

```
for skill in [s1,s2,...,sn] do
    speech ← TextToSpeech("Alexa, open {{skill's name}}");
    play speech
    for command in [c1,c2,...,cm] do
        speech ← TextToSpeech(command);
        play speech;
        audio ← listen;
        text ← SpeechToText(audio);
        save text;
    end for
end for
```

3 METHOD

3.1 Skills Selection

By March 2018, more than 30,000 Alexa voice skills [5] have been published and organized by categories on Alexa's website.
To collect a representative sample for the purpose of our research, we first identified the 10 top categories with the most number of skills. The top 10 categories (and their subcategories) are: 1. Daily Activities (News, Weather), 2. Entertainment (Movies & TV, Music & Audio, Novelty & Humor, Sports), 3. Education & Reference, 4.Health & Fitness, 5.Travel & Transportation, 6.Games, Trivia & Accessories, 7. Food & Drink, 8. Shopping & Finance (Shopping, Business & Finance), 9. Communication & Social and 10. Kids. We wrote a script to scrape Alexa's website for the top 10 skills for each category based on the number of reviews. For categories with subcategories, we tried to balance the number across the subcategories manually. For example, the ten skills we selected to represent the Entertainment category consist of three in the Movies & TV subcategory, three in and Music & Audio subcategory, two in the Novelty & Humor subcategory, and two in the Sports category. In all, we selected a total of 100 skills for evaluation.

3.2 Guideline-Specific Response Elicitation Design

Based on the Amazon's voice design guidelines [1], we chose 3 representative user-aware design guidelines to focus on. In this paper, we use G1 to denote the guideline wherein a skill needs to adaptively re-prompt users when it receives no input; G2 denotes the guideline wherein the re-prompt messages are supposed to be slightly reworded with more detailed information based on previous user interaction, and G3 denotes the guideline wherein a skill needs to memorize a user's previous interaction mode to provide more personalized service. For each design guideline, we derived appropriate testing commands in order to elicit responses that we could evaluate with respect to that guideline. The details are presented below.

Problem handling support (G1, G2): Problem handling is one of the most important aspects of user-aware design. One typical problem handling scenario we chose to evaluate was how the Alexa skill would react when it does not receive an answer from the user. In order to evaluate the compliance situation of these 100 skills with respect to G1 and G2, we first designed crawler loops by setting the basic commands as elicitation commands. Within one round of the crawler loop, the crawler will say "open", "help", "stop" commands and listen to the responses in turn (this crawler loop is denoted as "open-help-stop" loop in the rest of the paper). In our response collection process, we implemented an "open-help-stop" loop and then repeated the "open- elicitation command-stop" loop three times to make sure each skill was fully explored (using self-generated commands as elicitation). After that , we enabled this

[1]The project website is: https://gtts.readthedocs.io/en/latest/
[2]The project website is: https://pypi.org/project/SpeechRecognition/

skill again and stopped giving further command to wait for how it would respond. Our crawler then repeated this process for all skills in the sample.

Remember what was said (G3): According to the design guidelines, users will appreciate if skills remembered what was previously said by users to provide more personalized services. In order to test this, we first fully explored the skills (as how we handled G1 and G2), and then ran our "open-help-stop" loop one more time to see whether the skills remembered the last interaction and would change its responses accordingly.

3.3 Crawled Data Correction and Coding

After response data was collected, transcribed into text, and saved using our crawler, we manually analyzed the data as follows. First, we compared this dataset to a small pilot dataset of 20 skills we previously collected by hand in order to identify any discrepancy between machine and human transcribed responses. In doing so we were able to detect and correct problems brought by limitations of speech-to-text technology, such as typo and missing punctuation. After data correction, two researchers independently coded each response's compliance with respect to design guidelines. In terms of problem handling support, we first determined if the skill supports G1 and then compared with the previous welcome message to determine if the re-prompt messages were reworded and personalized. For G3, by comparing the last and the very first "open-help-stop" loop's responses, we judged whether the skill memorized previous interaction. Afterwards, two researchers compared their coding results and resolved their discrepancies.

4 FINDINGS AND ANALYSIS

Out of our sample of the 100 most popular skills from 10 categories, we did not retrieve responses from six skills due to account linking errors or access permission issues. Hence, our findings presented below are based on 94 skills.

4.1 User-Aware Design Guidelines' Compliance

Table 1 shows the compliance rate for each user-aware design guideline. For problem handling support (G1,G2), we manually determined that 82 skills (of 94 total) should have support for G1. (Some skills are not expected to support G1, like those which are meant for passive listening as well as "one-shot" skills[3]). Among these 82 skills, 74 (90.2%) of them supported G1. During the evaluation process, we encountered some skills which did not support re-prompting, such as *Scryb* and *Mastermind* from the Communication & Social categories. When they did not receive an input from the users, the program quit automatically. Among the 74 skills that supported re-prompting, only 23 of them supported G2 in their re-prompts. Some skills, such as *Bring* from Shopping & Finance category and *I'm Driving* from Travel & Transportation category, did not reword their responses and simply repeated their previous response when lacking user input.

In terms of personalizing based on previous interactions (G3), we found that only 32 (34.0%) skills memorized previous interactions and changed their interaction modes accordingly. Some example skills we encountered during the evaluation could help us identify the characteristics of skills which did not follow G3. An example of compliance is the skill *Lemonade Stand*. It is a CA-based game application. This skill consistently remembered where the game was previously interrupted and, when users reopened this skill, the skill would briefly summarize the current game status and prompt users to continue the game. On the contrary, skills such as *5-min Plank Workout* from Health & Fitness category and *Short Bedtime Story* from Entertainment category were not compliant. These skills restarted without communicating the previous interaction status. As we examined further, we identified certain legitimate exceptions past interactions were not remembered. In some skills, the contents are updated in a timely manner such that don't rely on user responses to function. For example, *This Day in History* is a skill that helps users learn more about historical events and is updated with new information daily). Similarly, some skills' operations are simple and don't require user input to update the experience (like *5-min Plank Workout* and *Short Bedtime Story*).

Table 1: The Rate of Compliance for User-Aware Design Guidelines

Design Guidelines	Number of Skills That Actually Support	Number of Skills That Should Have Supported	Supporting Percentage
(G1) Re-prompt When Receives No Response	74	82	90.20%
(G3) Remember What Was Said	27	94	34.04%
(G2) Re-prompt With Rewording	23	82	28.00%

4.2 Comparative Analysis Across Voice Skill Categories

As indicated in Table 2, we calculated a ranking based on each category's average number of Alexa skills which followed the user-aware design guides. We also calculated the percentage of skills that supported each design guideline within each category.

Table 2: The Average Support Rate in Three User-Aware Design Guideline Across 10 Skill Categories

Category	Average Number of Guidelines that A Skill Complys with	Support Rate for Re-prompt(G1)	Support Rate for Re-prompt with Rewording(G2)	Support Rate for Remember What Was Said (G3)
Games	1.8	100.00%	20%	60%
Travel & Transportation	1.7	80.00%	40.00%	50%
Kids	1.5	80%	20%	50%
Communication & Social	1.5	87.50%	37.50%	25%
Food & Drink	1.4	80.00%	10.00%	50%
Daily Activities	1.3	88.89%	22.20%	22.20%
Shopping & Finance	1.2	66.67%	22.20%	33.30%
Education & References	1.1	80.00%	20.00%	10%
Health & Fitness	1.1	55.56%	33.33%	22.20%
Entertainment	1	66.67%	22.20%	11.10%

The Games category has the highest compliance with personalized re-prompt (G1) and with remembering what was said by users (G3). Correspondingly, Games also achieved high user ratings from Amazon's Alexa skill webpage (the 10 selected skills' have an average rating of 4.5 out of 5).

[3]"one-shot" skills are those with which users can complete their tasks in a single utterance and do not have a chance to say more commands.

This result confirms that Games skills are expected to involve more interaction with users and require more complicated user interface designs, such as remembering users' previous score and provide personalized game dynamics.

As for categories with low G2 and G3 compliance, many of them were simple, straightforward applications that missed opportunities for personalized experiences. For example, skills in the Daily Activities category are meant to be used frequently on a daily basis; skills in the Entertainment category are meant to quickly entertain users in terms of music, jokes, etc.; skills in the Health & Fitness, Education & References category are meant to provide direct and accurate inquiry information. Although their interaction modes tend to be simple and straightforward, they can still benefit from personalized services. However, our results indicate that these categories do not perform very well in this aspect, which implies that more attention must be paid in the future developments.

5 DISCUSSION

5.1 Category-specific Evaluation

The finding that variation in user-aware design guideline compliance across skill categories exists help address our first research question. The variation suggests that each skill category has its own specific requirements and characteristics which need to be considered during evaluation. For example, Games category tend to involve more personalized design while categories which are meant to be used on frequent basis prefer more straightforward interaction. This opens up a further research question for future investigation: How should user-awareness evaluation be adapted for different categories and even further, application scenarios? We can start from studying which design guidelines are more integral to each category. Additionally, we can also study the category variations in terms of interaction flows and understand challenges that may arise.

5.2 Automating User-Aware Evaluation

To address our second research question of efficient evaluation, we find that the automatic voice application crawler we introduced in this paper could evolve to an automatic user-aware evaluation system in the future. On the technical side, our crawler focuses on automatic response data collection, which is the first step of an automatic evaluation system. Next, we would need to automate the data coding/labeling process. Innovative labeling algorithms could be developed based on our manual-labeling process; the study of common patterns contained in various collected responses could help conduct better topic analysis to increase the labeling accuracy.

6 CONCLUSIONS

With the popularity of CA-based voice applications, evaluating them with a focus on user-awareness is increasingly important. In this work, we contribute an automatic voice application crawler to evaluate the compliance of CA-based voice applications with user-aware design guidelines. Our findings revealed that only a small part of selected voice applications implemented particular user-aware design guidelines like G2,G3. This suggests a need for more robust evaluation tools, especially to support developers in assessing the usability of their own applications. Our findings also show the necessity of taking categories into consideration when doing user-aware evaluation. In sum, our research identified directions for automating the evaluation process. Based on our findings we showed that our CA-based voice application crawler should be a fundamental research prototype for user-aware evaluation tool design.

REFERENCES

[1] Amazon Alexa. 2018. Voice Design Guide. https://developer.amazon.com/designing-for-voice/
[2] Noor Ali-Hasan. 2018. Evaluating Smartphone Voice Assistants: A Review of UX Methods and Challenges. https://voiceux.files.wordpress.com/2018/03/ali-hasan.pdf
[3] Corey Badcock. 2015. First Alexa Third-Party Skills Now Available for Amazon Echo. https://developer.amazon.com/blogs/post/TxC2VHKFEIZ9SG/First-Alexa-Third-Party-Skills-NowAvailable-for-Amazon-Echo
[4] Jason Douglas. 2016. Start building Actions on Google. https://developers.googleblog.com/2016/12/start-building-actions-on-google.html
[5] Bret Kinsella. 2018. Amazon Alexa Skill Count Surpasses 30,000 in the U.S. https://voicebot.ai/2018/03/22/amazon-alexa-skill-count-surpasses-30000-u-s/
[6] Bret Kinsella and Ava Mutchler. 2018. Smart Speaker Consumer Adoption Report 2018. https://voicebot.ai/wp-content/uploads/2018/03/smart_speaker_consumer_adoption_report_2018.pdf
[7] 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*. dl.acm.org, 5286–5297.
[8] Actions on Google. 2018. Conversation Design. https://designguidelines.withgoogle.com/conversation/