How does your Alexa behave?: Evaluating Voice Applications by Design Guidelines Using an Automatic Voice Crawler

Xu Han
University of Colorado Boulder
Boulder, USA
xuha2442@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 still remains in the area of voice interaction design and evaluation. In our study, we deployed a voice crawler to collect responses from the 100 most popular Alexa skills within 10 different categories. We then evaluated these responses to assess their compliance to 8 selected design guidelines published by Amazon. Our findings show that design guidelines requiring basic commands support are the most followed ones while those related to personalized interaction are relatively less. There also exists variation in design guidelines compliance across different skill categories. Based on our findings and real skill examples, we offer suggestions for new guidelines to complement the existing ones and propose agendas for future HCI research to improve voice applications' user experiences.

CCS CONCEPTS

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

KEYWORDS

conversational agents; voice user interface design; user experience evaluation;

ACM Reference Format:
Xu Han and Tom Yeh. 2020. How does your Alexa behave?: Evaluating Voice Applications by Design Guidelines Using an Automatic Voice Crawler . In *IUI '20 Wokrshops, March 17, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

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 [17]. Amazon's Echo series devices make up 71.9% of the market, followed by Google's devices with 18.4% [17].

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 [8] and Google's Home Actions [12] 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 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 [3, 23] 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.

There is a huge body of literature in HCI that propose design guidelines to educate practitioners in the field who want to design and develop an application for a wide range of interactive technologies (i.e. web readability design [21], gesture user interface design [14]). However, most of these research efforts were concluded at the publication of these guidelines; few went further to understand whether these guidelines would be later on accepted and followed by designers and developers in the wild. In the example of Amazon, design guidelines for Echo, crafted by its own team of UX researchers, have been published for more than a couple of years [4]. Tens of thousands of developers have designed and published voice skills by following them. This situation provides a good opportunity to study the adoption pattern of design guidelines in the wild.

An example of a well-designed Alexa skill is *Would You Rather for Family*. This skill is an interactive Q&A game that exhibits several design features following Amazon's guidelines, including remembering where the last interaction ends, giving a personalized opening prompt to users, and speaking naturally. Deservedly, this skill has a high average rating – 4.9 out of 5 stars based on 3209 user reviews. In contrast, an example of a poorly-designed Alexa skill is *AccuWeather*. This skill's average rating is low – 2.2 out of 5 stars based on 182 user reviews. By interacting with this skill, we can tell that the skill's design violates several design guidelines, such as handling errors properly. These violations are also complained by some users in their reviews. By analyzing a large number of skills like these, we can gain insights into design guidelines' adoption pattern. We want to ask: **Among the design guidelines for voice applications, which are followed or violated more often by developers in the wild? (RQ1)**. Another phenomenon we observed is the high variance in user ratings across app categories. For instance, we found the average user rating of top 10 popular skills in the Games category is 4.5, comparing to 2.6 for those in

the Food & Drink category. Motivated by this phenomenon, a second research question can be opened: **Could this high degree of variability among categories be related to whether certain guidelines are followed or not followed? (RQ2)**.

To study these questions, we decided to limit the scope to Alexa skills in this paper. We selected a sample of 100 most popular Alexa skills from ten different categories and evaluated whether their designs follow the a selected subset of Amazon's official design guidelines. Note that our scoping decision does not imply an acknowledgement of Amazon's design guidelines as the gold standard nor an endorsement of Amazon's products. Rather, the decision is based on where we might be able to gather the most data, which platform has the largest number of developers in the wild, and which set of design guidelines are most likely read by these developers (which is unlikely an academic paper). To automate our data collection process, we deployed a voice skill crawler to collect responses from these skills under different commands input. We then analyzed the collected responses to determine whether or not certain guidelines are followed. Regarding the first research question, an example of key findings is that basic commands support are the most obeyed guidelines while personalized service-related guidelines are relatively less obeyed. Regarding the second research question, an example of key findings is that skills in the Games category on average obey the most guidelines while skills in the Entertainment category the least.

Furthermore, previous research (e.g. [11, 18]) has studied the general gulf between user expectations and real user experiences, which indicates a need of a comprehensive set of UX design guidelines for developers. Whilst UX design guidelines exist (e.g. Amazon and Google design guides), further revision iterations are still necessary. Thus, based on the findings on a large sample of skills in our evaluation process, we identified several aspects that current UX design guidelines do not cover and proposed additional design recommendations to fill this gap. In the remainder of the paper, we provide related work, a detailed description of our method, a comprehensive presentation of our findings regarding the two research questions, suggestions for how to improve the current design guidelines, and agendas for future HCI research.

2 RELATED WORK

2.1 Limitations on User Experience of VUIs

Recent years' advances in speech technology have led to voice user interfaces' (VUIs) improved accessibility and they have been studied in the HCI literature in a wide variety of application contexts (e.g. assistive services [28], education [9], health [25], entertainment [29] and Internet of Things (IoT) [22]). However, despite the benefits and convenience they have brought with us, VUIs still possess several limitations that would affect the user experience (UX). Some users may feel less in control since VUI provides no visual feedback [19] and the lack of VUI system transparency would result in users either feeling overwhelmed by the unknown potential, or led them to assume that the tasks they could accomplish were highly limited [18]. In some situations, voice interactions may evoke negative feelings in users [19]. In terms of subjective satisfaction, users may not feel comfortable talking with machines if the synthesized speech does not sound natural [30]. More specific to voice assistants (VAs) like Amazon's Alexa, several issues have been reported, such as concerns over users' privacy [20, 24], technical limitations of natural language processing [30] and restricted communication protocols [27]. Under these circumstances, user experience evaluations of VUI, specifically VAs, deserve further attention and studies.

2.2 User Experience Evaluation of VUIs

Based on the results of our literature survey, we noted several existing user experience evaluation methodologies that could be applied to VUIs or VAs. Traditional usability studies are very useful in gathering feedback and conducting evaluation analysis. In [18], researchers interviewed 14 users of VAs in an effort to understand the factors affecting everyday use. [7] deployed traditional lab-based usability studies using multiple fidelities like static mock, functional prototype and launched products for future design iterations. At the same time, longitudinal study is another effective methodology that can shed lights on real life scenarios and situations for using VAs[7].

Specifically for Alexa skills' user experience evaluation, although Alexa provides an overall platform for developers to check their skills before submitting to the review process, there is still no guarantee these skills follow the published voice design guide [2]. The user experience evaluation of skills still heavily relies on subjective data such as user ratings, reviews, feedback and reports. Thus, there is a need for a more systematic and objective approach to evaluating voice skills. Our study represents one possible approach by comparing across a large number of voice skills and examining their designs with respect to official design guidelines.

3 METHOD

In order to investigate the adoption and compliance pattern of current Alexa design guidelines, we first deployed a crawler system to collect responses from a sample of 100 Alexa skills and then manually labeled those collected responses to study whether or not they comply with the selected design guidelines. Here we elaborate our method in details.

3.1 Alexa Skills Selection

More than 30,000 Alexa voice skills [16] have been published by thousands of third-party developers. On Alexa's website, these skills are organized by categories. Because the variance on user average ratings across different categories is high, we are interested in studying these skills. In this case, we wanted to collect a representative sample for the purpose of our research. First, we identified the ten top categories with the most number of skills. The ten 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 and Finance (Shopping, Business & Finance), 9. Communication and Social and 10. Kids. We wrote a script to scrape Alexa's website to pick the top ten 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 the Music & Audio subcategory, two in the Novelty & Humor subcategory, and two in the Sports category. All in all, we selected a total of 100 skills for our study.

3.2 Alexa Skill Responses Crawler

The most common interaction flow of an Alexa skill is the "open-command-stop" flow. To begin interacting with a skill, A user first says "Alexa, open X", where X is a skill's invocation name. Then, the skill typically responds with an introduction or greeting message. After that, the user starts uttering specific commands to make use of the skill's functionality. The skill responds with its answers or follow-up questions. The conversation continues until the user says "Alexa, stop" to indicate their desire to quit. Sometimes, the skill responds with a goodbye message, but not always. In order to study our research questions, we needed to have conversations like this with each of the 100 skills in our sample, recorded how each skill responded, and analyzed whether its responses followed or violated certain design guidelines. Our initial attempt was fully manual. Given a skill, we spoke to it, listened to and wrote down its responses in an excel spreadsheet, and coded the responses with respect to their compliance with design guidelines. However, after about 20 skills, we found manual data collection time-consuming, difficult to scale to a large sample, and hard to replicate for other researchers. Thus, we were motivated to develop a method to automate certain parts of this process.

We present a crawler tool we developed to automatically converse with a given skill and record the skill's responses (the progress paper of this tool was presented in [15]) . The input to this tool is a list of skill names. The output is an excel spreadsheet containing each skill's responses (automatically recorded and transcribed) in various simulated conversation sessions. Researchers can then review and analyze the spreadsheet data for their own research questions, which in our case are what design guidelines are more frequently adopted (RQ1) and how such adoption varies across categories (RQ2).

Figure 1 provides a conceptual example of how our crawler simulates a voice conversation between users and Alexa devices.

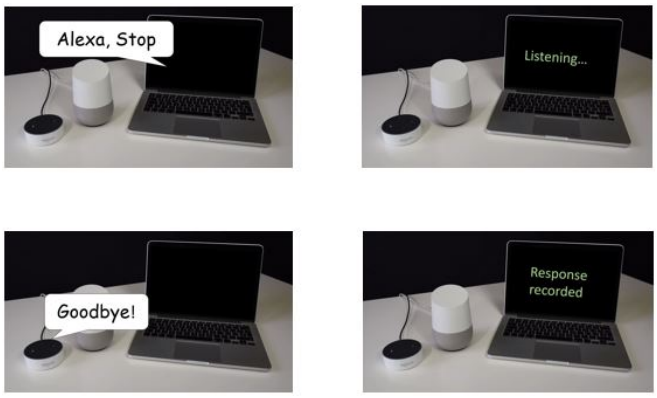

Figure 1: Working Process of the Voice Crawler

To simulate speaking a command to a skill, our tool uses Google's Text-to-Speech package for Python[1]. Then, it listens to and records the skill's responses. We used the Speech Recognition package for Python [2] to implement the listening ability. Given our sample of 100 skills, our crawler iterated through them, carried out a range of conversations with each skill, and listened to the skill's responses. This automatic data collection process is described pragmatically as follows:

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
```

3.3 Guideline-Specific Response Elicitation Design

Amazon's voice design guide [3] provides more than two dozens guidelines. In our study, we limited the scope to a sample of eight guidelines. They are denoted as G1 to G8 in the rest of the paper. For each design guideline, we needed to come up with an appropriate testing conversation flow that can be applied in the crawler in order to elicit responses we can evaluate, with respect to that guideline. The details are presented below.

Basic commands support (G1, G2, G3): Three design guidelines recommend voice skills should support users to start, get help on, and end an interaction. For Alexa skills, these translate into the ability to understand the basic commands of "open", "help", and "stop" respectively. Described in more details, when a user says "Alexa, open [skill's name]," the skill needs to remain open and wait for the user's responses (G1); when a user asks "Alexa, help," the skill is expected to provide informative instructions such as introducing its core functionality (G2); and when a users says "stop," the skill should end the conversation naturally and gracefully with few or no words (G3). In order to evaluate the compliance situation of these 100 skills with respect to G1, G2, G3, we 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). Based on the responses collected, we would get to know how many skills support basic commands and conduct our analysis.

Variety support (G4, G5): Two design guidelines recommend voice skills should provide varying responses to the "open" (G4) and "stop" (G5) commands so that the interaction can feel more natural and less robotic. To test whether a given skill complies with these guidelines, our crawler tool carried out an "open-help-[commands]-stop" dialogue where commands are what the specific

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

[2]The project website is: https://pypi.org/project/SpeechRecognition/

skill can support. This dialogue was repeated N times where N is default to three so that we can detect variations in the skill's responses to "open" and "stop" command, if any. Particularly, in the first run of this dialogue, the tested skills was first-time enabled. The second round was run after solving all the account linking, age verifying steps. The third round was run after the skill has been fully explored with a list of commands. The list of commands were automatically extracted from the skill's response to the "help" command in the second round. For example, the skill *"Examining the Scriptures Daily"* responded to "help" with the message *"You can say tell me my daily text for today or read me my daily text for last Monday. You can also say read me tomorrow's daily text."* We wrote a parser to extract three commands from this message: "tell me my daily text for today", "read me my daily text for last Monday" and "read me tomorrow's daily text." Then each of the extracted commands was applied in the crawler and given to the skill.

Error handling support (G6, G7): Problem handling is one of the most important aspects of user-aware design. We chose to evaluate two guidelines regarding error handling. The first guideline is when a skill receives no answer from a user regarding a question, it should deliver a re-prompt (G6). The second guideline is the re-prompt should be reworded or with more detailed instructional information(G7). To collect responses for a given skill regarding its error handling ability, our crawler first carried out a "open-help-stop" loop and repeated the "open-command-stop" loop three times to make sure the skill was fully explored (the commands were automatically extracted similar to as how we handled G4, G5). After that , we enabled this skill again and stopped giving further command to wait for how it would respond. Our crawler then repeated this process for all skills in our sample.

Memorizing support (G8): According to the design guide, users would appreciate it if a skill can remember their past interactions and provide more personalized services (G8). In order to test this, our crawler first fully explored a skill's capabilities (like G4 and G5), and then carried out an "open-help-stop" loop one more time to see whether the skill remembered its last interaction and personalized its responses accordingly.

3.4 Analyzing Responses by Design Guidelines

Given our sample of 100 skills and 8 design guidelines to test for, our crawler automatically collected more than 1000 responses (the dataset is included in the supplementary material). We manually analyzed the data as follows.

3.4.1 Data Correction. 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.

3.4.2 Data Coding. After data correction, two researchers independently coded each response's compliance with respect to design guidelines. Afterwards, two researchers compared their coding results and resolved their discrepancies.

For basic commands support, we examined collected responses to see whether the skill successfully executed the commands. For variety support, we compared the responses across repeated dialogues. If there are variations, we would code as following G4 or G5. For error handling support, we first determined if the skill supports G6 and then compared with previous messages to determine whether the re-prompt messages were reworded or not. Finally, for memorizing support, by comparing the last and the very first "open-help-stop" loop's responses, we judged whether the skill memorized previous interaction. The contents of the two responses were compared. If the second time's contents include any personalized information or previous interaction information while the first time doesn't, we would code as following G8.

3.4.3 Comparative Analysis. After we coded all the responses, we were finally able to address our research questions by comparing the results across guidelines (RQ1) and across categories (RQ2). For each guideline, we counted the number of skills that follow it and picked out both positive and negative examples for further investigation. By comparing across different guidelines, we were able to understand the guideline adoption pattern in the wild, that is, which guidelines are obeyed by more or fewer skills. By comparing across categories, we were able to examine whether category can be a factor associated with whether certain guidelines are followed (or violated).

4 FINDINGS

In this section, we present our findings regarding the current compliance situation of a sample of 100 skills with respect to eight voice design guidelines, in order to address the two research questions proposed earlier in the introduction. In the following Improving Design Guidelines section, we will discuss several illuminating real-world examples, both positive and negative, we discovered during the data collection process, which serve to motivate further design recommendations for voice skills.

As described before, we initially selected 100 most popular skills from 10 different categories as our sample. All the responses were collected in late 2018. However, after all the data was collected and cleaned, we needed to exclude six skills for which we failed to obtain meaningful results because of issues related to account linking or access permission. Hence, our findings presented below are based on 94 skills. [3]

4.1 Basic Commands Support

4.1.1 Open Command (G1). According to the design guide and Amazon's Alexa building requirements [1], every skill we tested is expected to support open command (G1). When a user invokes a skill without specific intents (e.g. "Alexa, open [skill name]"), the skill is supposed to remain open and wait for the user's responses. At the same time, a welcome message which could prompt the customer to continue interaction is also required. We found all 94 skills supported G1.

4.1.2 Help Command (G2). The "Help" command is used to help customers navigate a skill's core functionality. G2 states that every Alexa skill should implement the built-in "help" intent to provide better user experiences. We found only 81 out of 94 skills supported G2. This left thirteen skills not supporting G2, including eight in

[3] Our dataset can be accessed on request.

the audio/music/sound category like *4AFart* (a skill that plays fart sounds) , four one-shot [5] skills (skills that only involve single turn interactions) and one skill, *Escape the Room*, in the Game category.

What could be the reasons these skills do not support G2? One reason is that audio/music skills are meant for passive listening, as in the case of *NPR One* and *Thunderstorm Sounds*. Another reason is that some skills only involve one-shot interactions where a user asks a question or gives a command, the skill responds with an answer or confirmation, and the interaction is complete [6]. Since one-shot skills will end the interaction and exit automatically after answering open utterance, users do not have a chance to say more commands, including the help command. Fact skills (skills that randomly tell users a fact concerning a certain topic when invoked) like *Cat Facts* are good examples of these one-shot skills. Furthermore, some other skills provide instructive information through other ways rather than a help message, as in the case of *Escape the Room* from the Games category, which asks users to go to a website for reference in its opening message.

4.1.3 Stop Command (G3). G3 states that every skill should respond to a user's "Alexa, stop" command. After the stop command is heard, a skill should exit and optionally return a response that is appropriate for the skill's functionality, such as a goodbye message [6]. We found all 94 skills could successfully exit. Also, 74 of them gave a goodbye message. For those skills who did not provide goodbye messages, most of them are one-shot skills which exit automatically after an one-sentence response.

4.2 Variety Support

Compared to basic commands support, we found variety support is provided by much fewer number of voice skills in our sample. Below we present our findings for the two relevant design guidelines we studied.

4.2.1 Variety in open responses (G4). When a customer invokes a skill without specific intents ("Alexa, open [skill name]"), the skill should deliver an opening prompt. Skills are expected to provide several variations of opening prompts including one for first-time use, one for return and personalized prompts (G4). We found 34 out of 94 skills (36%) supported opening prompt variations. Furthermore, we observed they often served three use scenarios (with overlaps). 1. Some (n=8) were daily used skills or skills with regular updates; variety in opening prompts help keep users feel fresh and updated. 2. Some (n=16) were kills that remembered previous interactions; whenever users open the skill, its opening prompts will tell users where they left off last time. 3. Some (n=13) were skills with multiple states; the opening message will always inform users the current state.

For the first scenario (daily use), one good example is the *"Zyrtec"* skill which can report weather, pollen count and predominant allergens in a user-defined location. When this skill was opened the first time, its opening prompt was *"Hello! Let's get ahead of your allergies with today's Allergycast based on your location. Just follow these steps. One, Open your Alexa app on your phone ... [19 more words]"* For the second time, the skill said *"let's start with your city and state, then we can get ahead of those allergies by setting up your allergy test report. What's your city and state?"*. For the third time, after the location was set, the skill's opening prompt turned into *"Welcome to Zyrtec. Today in xxx, the pollen count is High, at 9.2 out of 12... [34 more words]"* Comparing these three opening prompts, we found that when the skill was first enabled or used, it provided instructions about setting up step by step and elaborated clearly about the location requirement. After the skill got the location permission, the opening prompt changed into daily report of pollen. The whole interaction was natural and personal for users. In contrast, a poor example is *Examining the Scriptures Daily*. We found the skill always responded with the same sentence: *"Which day do you like to hear"*. Although this opening prompt provided users with a cue to begin speaking and coached users on what to say next, the interaction could feel monotonous and less natural.

For the second scenario (remembering previous interactions), we found 26 of 94 skills (28%) could remember previous interactions but only 16 supported variations (17%). A good example is *7-min Workout* in the Health & Fitness category. This skill is used to play instructions and background music for people who work out. When the skill was firstly used, its opening message was *"Welcome to Seven Minute Workout. When you are ready, just say start workout."*. After a previous workout was interrupted, the skill's opening message changed into *"Welcome to Seven Minute Workout. To continue where you last left off, say ready. Otherwise, just say start workout."* In this situation, variety in opening prompts provides a personalized experience for users by allowing them to pick up where they left off.

For the third scenarios (multiple states), a good example was the popular *Magic Door* skill in the Games category. This skill always informs users the current game state in the beginning so that users can choose to resume or restart the game. A negative example is *Categories Game* skill, also in the Games category. The skill always says *"howdy. You're playing Categories Game! For instructions, say help me or, say start playing!"* in its welcome message and have to start the game all over again no matter how many times it has been played.

4.2.2 Variety in stop responses (G5). According to [26]'s work, "Alexa, stop" is the most frequently used command. In this case, variety in stop responses could greatly help users feel less like talking to a machine. We expected all the 74 skills which gave stop responses could add variety in them. However, based on our evaluation results, We found most skills had really short goodbye messages like *"OK"* and *"Goodbye."*. Only 19 of 74 skills (25%) varied its goodbye messages. One good example is again the *Zyrtec* skill. We found several variations such as *"Ok. If you need allergy info,* ***I am here for you. Unless you move me.*** *Then I am over there for you. If you need to stock up on Zyrtec, just say my name, then order Zyrtec."* , *"Ok, if you need allergen information,* ***I will be here for you****. Remember, if you need to stock up on Zyrtec, I can help. Just say my name, then order Zyrtec"* , and *"Ok. When you need allergen information,* ***I am here 24 7 365****. If you need to stock up on Zyrtec, just say my name, then order Zyrtec."* From these responses, we can see that although they expressed fairly the same meaning, the different wordings made the experiences felt less monotonous.

4.3 Error Handling Support

Error handling is an important part in any user interface design, voice skill design is no exception. In our study, we focused on a typical error handling scenario: when a customer responds to a skill prompt with silence. Under this situation, the skill is expected to deliver a re-prompt (G6) with rewording (G7) to disambiguate or elaborate on the kind of responses supported. Our findings are as follows.

4.3.1 Re-prompting (G6) with Rewording (G7). When doing the response coding, we manually determined that 82 skills (of 94 total) should have support for G6 and G7 (some skills are not expected to support G6 and G7, like "one-shot" skills and those which are meant for passive listening). For G6, We found a high percentage of skills supporting re-prompting—74 of 82 (90%).

For G7, however, we found a low percentage of skills—23 of 82 (28%)—that reword in their re-prompts. A good example is the *Amazon Story Time* skill. First, the skill greets users by saying *"Welcome back to Amazon Story time! Would you like to resume The Mouse and the Unicorn?"*. After the question, if it receives no responses from the user, it would re-prompt with *"you can say yes to resume or no to play the next story"*. In this example, the re-prompt offered more specific instruction for users to say yes or no. In contrast, a negative example is the *Bring* skill in the Shopping & Finance category, which simply stayed silent, without any instruction or hint to help users handle a potential error.

4.4 Memorizing Support (G8)

Just like conversing with a friend, users appreciate when Alexa remembers what happened previously and what was said, especially for frequent actions and static information. We found 27 of 94 (34%) skills that provided memorizing support.

One positive example is *Lemonade Stand* in the Kids category. It is a game where users can sell products and manage their income. The skill always remembers how the game ended last time. Each time a conversation began, this skill would say *"Today is your twelfth day selling lemonade. Currently, it's windy and cool with some clouds. The forecast is a very low chance it will be warm and partly cloudy. Your cost for lemonade is fifteen cents a cup. You have* ***three dollars and fifty cents****. How many cups do you want to sell? "* This message conveyed the key statistics to help users remember their progress. In contrast, the *5-min Plank Workout* skill in the Health & Fitness category did not say anything explicitly to users that it remembered what exercises users might have done in the previous session. It always asked users to start over again, which could be frustrating. As we examined further, we identified certain legitimate exceptions past interactions were not remembered. For example, a skill like *This Day in History* is designed to be relevant for that day where past interactions do not matter.

5 COMPARATIVE ANALYSIS

In previous sections, we presented our findings with respect to each of the eight guidelines (i.e., G1 to G8). In this section, we will compare our findings across both guidelines and skill categories. These comparisons address the two research questions in the introduction (i.e., RQ1 and RQ2).

5.1 Across Design Guidelines (Q1)

As shown in Table 1, we calculated the compliance rate for all 8 design guidelines and ranked them based on their rates. These results show that among the design guidelines we evaluated, some were more frequently violated than others.

Based on the ranking, open command support (G1) and stop command support (G3) were among those followed by the most number of skills. In contrast, Memorizing support (G8), rewording support (G7), and stop variation support(G5), were followed by the fewest skills. As we can see, both G8 and G5 are related to Alexa skills' personalized services. What could explain such differences in compliance rate across design guidelines? For guidelines related to personalized services, one possible explanation of their low adherence rate may be the difficulty in implementation, which involves user behavioral modeling, user data analysis and other techniques. Also, high-quality personalized services require users to provide more personal information. It is hard to strike a good balance between the quality of personalized services and users' concern about their privacy [24]. As for G7, the observed results told us that most of the skills only focused on providing re-prompts but did not take a further step to reword the repeated re-prompts to make a conversation more natural.

5.2 Across Voice Skill Categories (Q2)

In this part, we will make comparisons across different skill categories. As indicated in Table 2, we counted the number of design guidelines that a certain skill complied with and calculated the average number within one category. Based on the calculated results, we obtained a ranking where the Games category was the top-ranked one while the Health & Fitness and Entertainment categories had relatively low rankings. Moreover, we broke down the comparison into four types of design guidelines: basic commands support (G1, G2, G3), variety support (G4, G5), error handling support (G6, G7) and memorizing support (G8). We first calculated the percentage of skills that supported each design guidelines within each category and then computed the average compliance rate within each of the four types. The results are also shown in Table 2.

From the ranking shown in Table 2, we can see that the Games category had the highest compliance rate for variety support guidelines (G4, G5), memorizing support guidelines (G8) and close to the highest compliance rate for error handling support (G6, G7). Also can be seen is that the skills in the Game category were more likely to follow the rest of the guidelines. Game Skills are expected to involve more interactions with users and require more complicated user interface design, like remembering users' previous score and provide personalized game processes. When we looked at these 10 selected skills' user ratings on Amazon's website, they also achieved relatively high average user ratings (4.5/5), which matched with the comparative analysis results. The skills in the Kids category also held high ranking positions in our table. One explanation is that children are considered a sensitive population that tends to have a higher requirement for design quality.

Let us now turn attention to categories with relatively low compliance rates. Several interesting patterns emerged. For example, the Entertainment category had close to the lowest compliance rate across all four guideline types. At a quick glance, this finding was

Table 1: The Rate of Compliance for 8 Design Guidelines

Design Guidelines	Number of Skills That Actually Support	Number of Skills That Should Have Supported	Supporting Percentage
(G1) Open Command Support	94	94	100%
(G3) Stop Command Support	94	94	100%
(G6) Re-prompt When Alexa Receives No Responses	74	82	90.24%
(G2) Help Command Support	81	94	86.20%
(G4) Variety in Open Prompts	34	94	36.17%
(G8) Memorizing Support	27	94	34.04%
(G5) Variety in Stop Prompts	19	74	25.68%
(G7) Re-prompt With Rewording	23	82	28.00%

Table 2: The Average Support Rate in Four Design Features Across 10 Skill Categories

Category	Average Number of Guidelines That Skills Comply With	Average Support Rate for Basic Commands (**G1,G2,G3**)	Average Support Rate for Variety (**G4,G5**)	Average Support Rate for Error Handling (**G6,G7**)	Support Rate for Memorizing (**G8**)
Games	5.8	98.00%	55.00%	60.00%	60.00%
Kids	5.2	96.70%	40.00%	50.00%	50.00%
Food & Drink	5.1	100%	35.00%	45.00%	50.00%
Travel & Transportation	5.1	100%	20.00%	60.00%	50.00%
Communication & Social	4.8	95.80%	12.50%	62.50%	25.00%
Daily Activities	4.7	96.30%	28.00%	55.56%	22.20%
Shopping & Finance	4.6	96.30%	22.22%	44.44%	33.30%
Education & References	4.5	96.70%	25.00%	50.00%	10.00%
Health & Fitness	4.2	85.20%	27.80%	44.44%	22.20%
Entertainment	3.7	88.90%	16.70%	44.44%	11.10%

surprising because Entertainment and Games seemed similar yet occupied the two opposite ends in the ranking. Upon closer examinations, we realized skills in the Entertainment category tend to offer quick and instant "fun" such as telling a joke or a compliment, which do not need many user inputs and require less interaction design. Another example is the Communication and Social category that had a relatively low compliance rate with respect to variety support guidelines but a high compliance rate for error handling guidelines. One explanation could be that communication and social skills may involve users speaking longer and more intentional utterances and may be more prone to errors, which necessitates additional effort to handle errors. In conclusion, with respect to RQ2, we found evidence that design guideline compliance patterns do differ greatly across categories, which suggests associations between design guideline compliance and categories. However, we were unable to determine whether these associations are causal or correlational, which will require further studies.

6 IMPROVING DESIGN GUIDELINES

Based on our findings and real skill examples we encountered during the evaluation process, we derived a set of new design guidelines for voice skills to complement the existing ones. In order to make our ideas clearer, simulated user-Alexa dialogues are presented for some of the points.

6.1 Design Guidelines For One-Shot Skills

We found that a significant number of instances of guideline violations are associated with one-shot skills such as *Cat Facts* and *Damn Girl*. In our sample, fewer than 10% were one-shot skills but they accounted for a large number of guideline violations. Upon closer examinations, some of the violations could be excused (users do not have chances to go deeper). This observation may suggest that the official design guidelines need to be revised to consider the special needs of one-shot skills. Here we present two ideas for the revision informed by our findings.

6.1.1 Use Informative Invocation Name to Replace Help Command. Since a one-shot skill often exits automatically after responding to users' commands, users may not have a chance to interact with the skill deeper. Thus, it is advisable to carefully choose an invocation name that is informative to remind users of its core functionality. A good example is the *Rain Sounds* skill whose name clearly indicates

that this skill intends to play rain sounds for users. In contrast, a negative example is the *Damn Girl* skill, which carries a unusual name but gives users little information about what it does (in fact, it says a different compliment each time).

6.1.2 Personalize the Contents Based on User's Interactions with Other Skills. One-shot skills' interaction mode limits the collection of user inputs, which makes the process of personalizing their contents very difficult. In this case, a good way to solve the problem is to connect with other skills for more user inputs. For example, a one-shot skill aimed at providing basic facts about cats (such as *Cat Facts*) could make use of a user's previous inquiries about a cat's health, collected from other skills. With this personalized information, this one-shot skill could provide more relevant health facts about cats the next time the user opens it. But this approach must be implemented carefully to respect users' privacy preferences regarding sharing data across skills.

6.2 Design Guidelines For Personalized Skill Services

Our findings show that there is still a room for improvement in terms of providing personalized experiences for voice skill users. During the process of analyzing our data, we noted several real-world design examples that could inform new design guidelines for voice skill developers.

6.2.1 Change Interaction Mode for Repeat Users. Personalized service should not be limited to variety in responses, it should also be reflected through variations of the whole interaction mode. For example, through analyzing a user's interaction history, a skill could tell whether the user is a frequent user. If not, the skill could guide the user to explore its features in details. If yes, the skill could simplify or streamline the whole interaction flow to provide more personalized service. For example, repeat users could get what they want immediately or receive a list of recommended services based on interaction history. Here we present an ideal interaction mode variation example. First is the interaction mode for non-frequent users.
User: *Alexa, open Dishes Delivery.*
Alexa: *OK, what kind of dishes do you want?*
...(the skill acquire necessary information like dishes kind, price, personalized taste like dishes cooked with no peppers)
Alexa: *OK, got it. Your order is ready.*

Next are the good and bad examples of the interaction mode for frequent users.
User: *Alexa, open Dishes Delivery.*
Alexa: (Bad) *OK, what kind of dishes do you want?*
Alexa: (Good) *OK, welcome back. Do you still want "A" cooked with no peppers?*

6.2.2 Providing detailed information via other platforms. One limitation of a voice skill is the amount of information it can provide in a single utterance. Meanwhile, an overly long utterance in response to a user's question is highly discouraged. In this case, we found some voice skills take advantage of other platforms such as mobile, emails, and SMS to deliver extra information. A good example is the *Store Card* skill. When this skill needs to tell users information that is not suitable through voice interaction, such as an URL, instead of saying it aloud, it sends the information to a user's mobile app and explains to the user that *"we just made some improvements that you need to disable the skill and then enable it again. Please use the link we just sent to your app"*. This practice eliminates the need for users to listen and remember long text. Hence, we suggest that detailed information can be optionally provided via another platform.

6.3 Other Design Guidelines

Here we present several more design guidelines (not already covered by the official ones) informed by real world examples we observed, which reflected both good and bad design practices.

6.3.1 Give feedback to help users locate problems in their commands. Users might feel frustrated when their commands cannot be correctly processed by a voice skill several times in a row. Under this situation, if the skill could specifically tell users where the problems are in their input and give more specific instructions, it would more effectively help users adjust their input and receive the desired services from the skill. During the manual collection process, we found many skills just repeated the same generic sentence like *"Sorry, I didn't understand that. What would you like?"* when researchers gave commands that could not be understood. Those kind of responses do not provide any information about why Alexa cannot understand the user's command. We suggest an additional guideline that a skill should provide informative feedback such as telling users what it originally expected and why users' voice input did not match the expectation. Here is an example dialogue contrasting a good response with a bad response with respect to this guideline.
User: *Alexa, open Pizza Delivery.*
Alexa: *OK, what city do you live in?*
User: *My city is horse.*
Alexa: (Bad) *I didn't understand that. What city do you live in?*
Alexa: (Good) *(The skill's logic does not think 'horse' is a city name.) Sorry, "horse" is not a city name, can you say your city's name again?*

6.3.2 Let users know which skill they are currently interacting with. Sometimes users may mistakenly think they are interacting with a skill but in fact with another skill. They may say commands which are only meaningful for other skills but cannot be understood by this skill. In this case, a useful design guideline would be to remind users which skill they are interacting with when the skill fails to understand users. The *WebMD* skill is a good example following this guideline. Below is a sample dialogue that demonstartes WebMD's informative response.

(Suppose a user forgot to exit *WebMD* but thought he is interacting with a pizza delivery skill.)
User: *Alexa, order pizzas.*
Alexa: (Bad) *Sorry, I didn't understand. What would you like to know?*
Alexa: (Good) *Sorry, you are already speaking with the WebMD skill. You can ask things like "What is diabetes?" or "What are the side effects of Nexium?" What would you like to know?*

6.3.3 Recognize and acknowledge problems in users' input. During our evaluation and the process of reviewing users' reviews, we noticed that for some skills, even if users give incorrect input, those skills still continue with the wrong information and respond to users with irrelevant answers. For example, *Categories Game* is a skill that presents different categories and asks users to come up with a word that begins with a certain letter in each category. One of the reviews said that the skill sometimes does not seem to understand the words users actually spoke and continues the game regardless. Hence, we suggest a skill should improve the ability to recognize different types of errors and acknowledge those errors, rather than acting if there is no error.

6.3.4 Don't give advertisements or encouragements too often after "stop". We observed that some skills include advertisements or encourage users to give ratings in the goodbye message too often. For example, the skill *Big Sky* always asks users to write a review at termination. One reviewer complained *"A nice, helpful app, except that it frequently ends answers to queries with 'please consider writing a review for this skill...', etc. I end up spending more time stuck listening to it beg for a review than, say, finding out what the temperature is."* Whenever users give the stop command, they hope to stop the skill successfully rather than listen to other bunch of sentences. In this case, if the skill could reduce the frequency of or stop giving advertisements after user says "stop", it would provide better user experiences. Instead, skill developers should use other channels to do their advertising.

6.3.5 Don't include questions in goodbye messages. We observed that some voice skills include a question in its goodbye message in response to the "stop" command. For example, the *Alexa Prize Socialbots* responds to a stop command by uttering *"Thanks for chatting! Quick question. On a scale from 1 to 5 stars... how do you feel about speaking with this socialbot again?"* We found this practice problematic. As mentioned before, whenever users say "stop" to a skill, they indicate strongly that they do not wish to interact with the skill anymore. But asking a question in the goodbye message would require users to continue the interaction, which may negatively affect the user experiences. Hence, we suggest that voice skills hould not include questions in goodbye messages.

7 HCI RESEARCH AGENDAS

7.1 Improving User Experiences

Although current design guidelines and suggestions we offered before have already covered many design problems developers might meet, future research is still necessary to revise the guidelines to meet new needs.

First of all, research on understanding the design space of voice skills is critical. The design space of traditional voice user interfaces has been proposed [10], which includes three variables: grammars—possible things users can say in response to each prompt and which are understood by the system), dialog logic—actions taken by the system, and prompts—the recordings or synthesized speech played to the user during the dialog. However, the design space of the new generation of voice skills by third-party developers has not been adequately researched. Although voice skill design and voice user interface design share certain problems such as how to express effective information through a natural and conversational interaction without graphical assistance, voice skills still have their own characteristics including strong interaction objectives, shared interaction features across ecosystems and so on. We argue that design guidelines for traditional voice user interfaces only serve as a good starting point for understanding the design space of voice skills.

Next, we found that connecting with other platforms is important for a voice skill, especially when the skill has versions on other platforms, such as *Uber*, that have corresponding mobile versions. In order to improve user experiences, sharing information across different platforms is critical. For example, an Alexa skill could give users a concise message while detailed information could be sent to users' mobile application. Another benefit of connecting with other platforms is to allow users to receive consistent services. Hence, future studies are needed to understand how to best support a seamless cross-platform experience.

Finally, the design of personalized voice skills is also worth studying in the future. From our findings, we found personalized services do not have very high support rate among the current popular skills. This finding implies personalized design requires more attention and further revision. Studies show that in order to achieve a high degree of personalization, more personal or private information is often required from users. However, due to privacy concerns, users may want to disclose less personal information [11, 24]. In this case, how to mitigate users' concerns that their privacy might be invaded can be an inspiring topic to be explored in future research. Methodologies related to investigating VUI users' privacy concern have been adopted in various research works [11, 13], which can be deployed in the future.

7.2 Category-specific Design

One contribution of this paper is the finding that there exists a high degree of variation in design guideline compliance across different skill categories. Thus, the variation we found suggests each skill category has its own specific requirements and design challenges. This opens up several research questions for future, such as how should design guidelines and design space be adapted for different categories and even further, application scenarios? We can start from understanding which design guidelines are more important and needing more attention for each category. Also, we can study the variations in interaction flows across categories and understand the different challenges one may face in evaluating the skills in each category.

7.3 Evaluation Methodology

The evaluation methodology presented in this paper has several limitations. On the technical side, our crawler is a research prototype that covers only selected design guidelines. More research is still needed to support others. Also, the speed of our tool is limited by the natural speed of a human's voice (since our tool simulates a human's interaction with a voice skill). In terms of data collection, we only focused on popular skills and categories. We do not yet know whether our findings can be generalized to less popular skills. At the same time, only 8 design recommendations were evaluated

and we did not explore all the possible commands. Furthermore, the responses we collected were only one snapshot in time; we don't know whether tested skills have since updated their interaction models. In terms of the crawling algorithm, we cannot cover all the possible situations, like when we tested variety (G4,G5), the crawler only repeated the same commands for three times. The possibility of variety appeared in the fourth time or later was not eliminated. In terms of responses analysis, all the response labeling was conducted manually, which leaves room for improvement.

Correspondingly, there exist several possibilities for improving the evaluation methodologies in the future in order to better triangulate usability issues and design guideline violations. They include increasing the size and variety of the response data collection, integrating log data analysis (user's interaction history) [26] to enrich the commands and responses dataset, and automating the labeling and evaluation process.

8 CONCLUSIONS

With the popularity of customized voice services, evaluation on them is of more importance than ever before. In our paper, we conducted design evaluation of a sample of 100 most popular Alexa skills from ten different categories using a voice skill crawler. The entire evaluation was performed with respect to eight design guidelines. Our findings revealed how these selected skills followed the guidelines. Based on our findings and the real sample responses we encountered during the evaluation process, we made several suggestions for improving the design of voice skills and identified challenges as well as opportunities for future research.

REFERENCES

[1] Amazon Alexa. 2019. Choose the Invocation Name for a Custom Skill. https://developer.amazon.com/docs/custom-skills/choose-the-invocation-name-for-a-custom-skill.html

[2] Amazon Alexa. 2019. Test and Submit Your Skill for Certification. https://developer.amazon.com/docs//devconsole/test-and-submit-your-skill.html

[3] Amazon Alexa. 2019. Voice Design Guide. https://developer.amazon.com/designing-for-voice/

[4] Amazon Alexa. 2019. Voice Experiences | Alexa Design Guide. https://developer.amazon.com/en-US/docs/alexa/alexa-design/get-started.html. Accessed: 2019-12-1.

[5] Amazon Alexa. 2019. Voice Interface and User Experience Testing for a Custom Skill. https://developer.amazon.com/docs/custom-skills/voice-interface-and-user-experience-testing-for-a-custom-skill.html

[6] Amazon Alexa. 2019. Voice Interface and User Experience Testing for a Custom Skill | Custom Skills. https://developer.amazon.com/docs/custom-skills/voice-interface-and-user-experience-testing-for-a-custom-skill.html#46-one-shot-phrasing-for-sample-utterances. Accessed: 2019-12-1.

[7] 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

[8] 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

[9] Julia Cambre, Ying Liu, Rebecca E Taylor, and Chinmay Kulkarni. 2019. Vitro: Designing a Voice Assistant for the Scientific Lab Workplace. In *Proceedings of the 2019 on Designing Interactive Systems Conference (DIS '19)*. ACM, New York, NY, USA, 1531–1542.

[10] Michael H Cohen, Michael Harris Cohen, James P Giangola, and Jennifer Balogh. 2004. *Voice User Interface Design*. Addison-Wesley Professional.

[11] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. "What Can I Help You with?": Infrequent Users' Experiences of Intelligent Personal Assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '17)*. ACM, New York, NY, USA, 43:1–43:12.

[12] Jason Douglas. 2016. Start building Actions on Google. https://developers.googleblog.com/2016/12/start-building-actions-on-google.html

[13] Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. 2018. A Survey Investigating Usage of Virtual Personal Assistants. (July 2018). arXiv:cs.HC/1807.04606

[14] Bogdan-Florin Gheran, Jean Vanderdonckt, and Radu-Daniel Vatavu. 2018. Gestures for Smart Rings: Empirical Results, Insights, and Design Implications. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, NY, USA, 623–635.

[15] Xu Han and Tom Yeh. 2019. Evaluating Voice Applications by User-Aware Design Guidelines Using an Automatic Voice Crawler.. In *IUI Workshops*.

[16] 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/

[17] 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

[18] 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.

[19] Michal Luria, Guy Hoffman, and Oren Zuckerman. 2017. Comparing Social Robot, Screen and Voice Interfaces for Smart-Home Control. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 580–628.

[20] Emily McReynolds, Sarah Hubbard, Timothy Lau, Aditya Saraf, Maya Cakmak, and Franziska Roesner. 2017. Toys That Listen: A Study of Parents, Children, and Internet-Connected Toys. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 5197–5207.

[21] Aliaksei Miniukovich, Antonella De Angeli, Simone Sulpizio, and Paola Venuti. 2017. Design Guidelines for Web Readability. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 285–296.

[22] Chris Norval and Jatinder Singh. 2019. Explaining Automated Environments: Interrogating Scripts, Logs, and Provenance Using Voice-assistants. In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2019 ACM International Symposium on Wearable Computers (UbiComp/ISWC '19 Adjunct)*. ACM, New York, NY, USA, 332–335.

[23] Actions on Google. 2019. Conversation Design. https://designguidelines.withgoogle.com/conversation/

[24] Kambiz Saffarizadeh, Maheshwar Boodraj, and Tawfiq M Alashoor. 2017. Conversational Assistants: Investigating Privacy Concerns, Trust, and Self-Disclosure. In *ICIS 2017 Proceedings*. aisel.aisnet.org.

[25] Jamie Sanders and Aqueasha Martin-Hammond. 2019. Exploring Autonomy in the Design of an Intelligent Health Assistant for Older Adults. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion (IUI '19)*. ACM, New York, NY, USA, 95–96.

[26] A Sciuto, A Saini, J Forlizzi, and J I Hong. 2018. Hey Alexa, What's Up?: A Mixed-Methods Studies of In-Home Conversational Agent Usage. *Proceedings of the 2018 on* (2018).

[27] Alexandra Vtyurina. 2018. 5 Seconds After: Exploring User Actions with Voice Assistants in the Moments After a System Response. https://voiceux.files.wordpress.com/2018/03/vtyurina.pdf

[28] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W White. 2019. Bridging Screen Readers and Voice Assistants for Enhanced Eyes-Free Web Search. In *The World Wide Web Conference*. ACM, 3590–3594.

[29] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. (Dec. 2018). arXiv:cs.HC/1812.08989

[30] Hong Zou and Jutta Treviranus. 2015. ChartMaster: A Tool for Interacting with Stock Market Charts Using a Screen Reader. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)*. ACM, New York, NY, USA, 107–116.