Evaluating Speech-Based Smart Devices Using New Usability Heuristics

Zhuxiaona Wei
deeplearning.ai

James A. Landay
Stanford University

We developed a set of 17 usability heuristics for speech-based smart devices. An expert evaluation of three popular devices shows that these heuristics can be used to uncover existing usability problems as well as help design new interfaces.

A recent empirical study showed that in both English and Mandarin, speaking is almost three times faster than typing a short message.[1] Thanks to recent breakthroughs in speech and language technologies, speech user interfaces (SUIs) have improved rapidly, and voice-enabled devices are now common. Baidu's Deep Speech 2 system, for example, can recognize spoken words with human-level accuracy.[2]

Nevertheless, designing good SUIs remains challenging.[3] The state of an SUI is often opaque to users, leading to more user errors compared to graphical user interfaces (GUIs).[4] Unfortunately, simply transforming GUIs into speech interfaces does not work well.[5] Although researchers have been working on SUI technology for three decades, much useful knowledge is in older papers and not easily accessible to designers. Moreover, the knowledge has not been updated to reflect recent improvements in speech-recognition accuracy. Consequently, those new to SUI design often feel lost.[6]

To help address these issues, we developed a new set of heuristics for designing and evaluating speech-based smart devices. To validate and improve these heuristics, we had a group of usability experts—half of whom specialized in SUIs—use them to empirically evaluate three state-of-the-art devices.

RELATED WORK

In the early 1990s, Jakob Nielsen developed a set of 10 usability heuristics for evaluating UIs (www.nngroup.com/articles/ten-usability-heuristics). Although these heuristics are most often applied to GUIs, he and his colleagues also used them to evaluate a telephone voice-response system.[7] However, the user input and system output options for the system were quite limited.

Published by the IEEE Computer Society
1536-1268/18/$33.00 ©2018 IEEE

Researchers have developed several SUI guidelines and best practices over the past 20 years.

In 1996, Alexander Rudnicky created 7 guidelines for SUIs integrated with visual applications.[4] However, today's devices are speech-first or even speech-only, and speech technologies have improved dramatically. The guidelines need to be updated to be used for today's smart devices.

In 2001, Laila Dybkjær and Niels Ole Bernsen created a usability testing guide for spoken-language dialogue systems.[8] However, we believe heuristic evaluation is more efficient than usability testing, especially since there are no good standards for SUI design yet. The heuristics can also be used for designing new systems.

In 2003, Bernhard Suhm created a database of SUI design problems and solutions to generate guidelines for telephone dialog system design.[3] He suggested using these guidelines for heuristic evaluation but did not validate the guidelines. Furthermore, unlike telephone systems, many of today's smart devices are not speech-only but also have a physical form with which users can interact, enabling a richer experience. There are likely different design problems due to these characteristics.

In their 2007 book *Wired for Speech*, Clifford Nass and Scott Brave presented valuable theoretical insights from years of research, many of which we incorporated into our new heuristics.[9] More recently, Cathy Pearl shared lessons from her career designing SUIs for mobile devices and interactive voice-response systems in *Designing Voice User Interfaces*.[10] Most of this knowledge is still applicable to today's smart devices, though it is hard to distill a set of manageable guidelines from her book.

In 2017, Google (https://developers.google.com/actions/design) and Amazon (https://developer.amazon.com/designing-for-voice) have each published a set of design guidelines for their own smart devices. However, to our knowledge there are no empirical evaluations of these guidelines.

In sum, no general guidelines have been developed specifically for evaluating state-of-the-art speech-based smart devices, nor have any empirical studies been done on these devices' usability. We believe both are critical for the research community to better understand existing problems and try to remedy them.

NEW SUI HEURISTICS

In adapting heuristic evaluation to SUIs, some researchers have modified Nielsen's 10 heuristics to be more applicable to the new interface style while others have extended them by adding SUI-specific heuristics. Drawing on the related work described above, we compiled a set of 17 new heuristics grouped into 5 categories: general (S1–S5); conversational style (S6–S8); guiding, teaching, and offering help (S9–S10); feedback and prompts (S11–S14); and errors (S15–S17). The heuristics are as follows:

S1: Give the agent a persona through language, sounds, and other styles.

S2: Make the system status clear.

S3: Speak the user's language.

S4: Start and stop conversations.

S5: Pay attention to what the user said and respect the user's context.

S6: Use spoken language characteristics.

S7: Make conversation a back-and-forth exchange.

S8: Adapt agent style to who users are, how they speak, and how they are feeling.

S9: Guide users through a conversation so they are not easily lost.

S10: Use responses to help users discover what is possible.

S11: Keep feedback and prompts short.

S12: Confirm input intelligently.

S13: Use speech-recognition system confidence to drive feedback style.

S14: Use multimodal feedback when available.

S15: Avoid cascading correction errors.

S16: Use normal language in communicating errors.

S17: Allow users to exit from errors or a mistaken conversation.

The list of heuristics along with detailed descriptions and examples can be found at http://hci.stanford.edu/publications/2018/speech-he/sui-heuristics.html.

EVALUATING THE NEW HEURISTICS

To validate and improve our heuristics, we had usability experts use them to empirically evaluate three state-of-the-art speech-based smart devices: Google Home, Amazon Echo, and Apple Siri (see Figure 1). Nielsen recommends using a minimum of 3–5 evaluators to identify most UI problems.[11] We felt that 8 evaluators should find most of the interface problems in the devices and could use the union of these problems as ground truth. Half of the participants in our study had an average of 11–20 years in SUI design; the rest were nonspeech usability experts, with an average of 11–20 years' experience, and each had completed more than 10 heuristic evaluations. Most of the evaluators were native American English speakers; one speech expert and one nonspeech expert were nonnative speakers. Seven of the evaluators were female. Only one of the participants did not currently use Apple Siri. Most of the speech experts owned at least one of the other two devices, which they used daily. Some of the nonspeech experts had tried but did not own an Amazon Echo. Each evaluation took 2.5–4 hours, and the evaluators received $100–$150 per hour as compensation based on their normal consulting rates. All 8 sessions were performed in a quiet meeting room at Baidu Research's office in Sunnyvale, California, to minimize noise and other distractions.

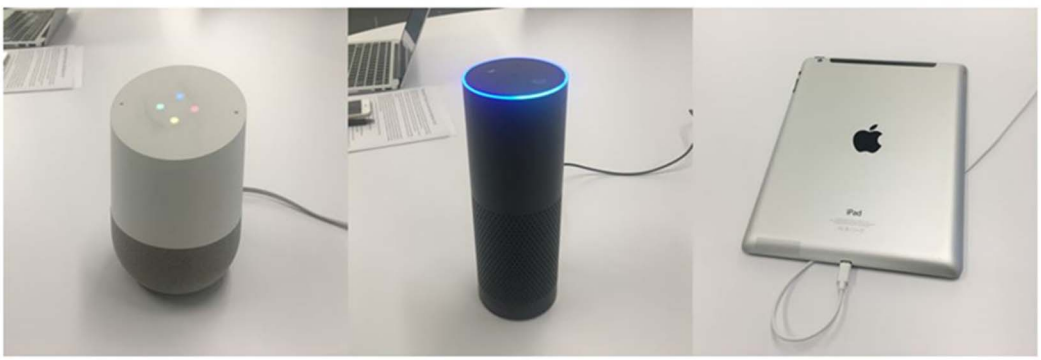

Figure 1. The three speech-based smart devices evaluated in our study: Google Home (left), Amazon Echo (middle), and Apple Siri (an iPad face down, right).

After a brief introduction, we presented the new heuristics to the evaluators, who were given time to read them and ask any questions. Next, we asked the evaluators to complete a set of 10 tasks on all three devices. They evaluated all 10 tasks in order on a single device at a time. Learning effects were eliminated by counterbalancing the order for evaluating the devices. All devices were reset after each session to ensure no language was learned from the prior interactions. Compiled from several market research reports, the tasks were the 10 most frequent real-life use cases for smart speakers: general questions, music, weather, local business, shopping, radio/news, messaging, calendar, to-do/reminders, and timers/alarms. As the devices have different functionality, not all 10 tasks are supported by each device. The evaluators rated each task on a scale of 1 (very difficult) to 7 (very easy). After completing each task, the evaluators documented the usability problems they found, along with the heuristic violated by each problem and

a severity rating. Severity was recorded using Nielsen's scale: 1—cosmetic problem, 2—minor problem, 3—major problem, and 4—usability catastrophe. After evaluating each device, the participants filled out a standard Subjective Usability Scale (SUS).[12] They then proceeded to the next device and repeated the procedure. Finally, we conducted a follow-up interview with the evaluators about their overall experience with the three devices and, more importantly, how the heuristics might be improved.

EVALUATION RESULTS

The evaluators initially found 388 problems. We analyzed their problem descriptions to identify identical ones—we considered problems that had similar descriptions and the same violated heuristic as the same problem. Some evaluators occasionally listed several heuristics for a single problem. In these cases, we chose the heuristic we judged to be closest to the problem description. After this process, we were left with 279 unique problems. We averaged the severity ratings for each problem. We considered problems with severity ratings equal to or greater than 2.5 as high severity-problems.

Problems Found

Table 1 shows the average difficulty level of each task on each device as rated by the evaluators. We report this for context only—our goal was not to compare the usability of the devices, which are designed to support different tasks.

Table 1. Average difficulty level of each task, from 1 (very difficult) to 7 (very easy).

Task	Google Home	Amazon Echo	Apple Siri
1. General Questions	3.5	2.9	3.0
2. Music	3.1	2.1	2.4
3. Weather	5.8	6.1	5.4
4. Local Business	5.3	4.1	4.4
5. Shopping	3.9	3.5	2.6
6. Radio/News	5.0	4.8	3.1
7. Messaging	2.0*	4.9	5.1
8. Calendar	5.0	6.3	5.1
9. To-Do/reminders	2.8*	5.6	5.0
10. Timers/alarms	5.0	5.6	4.4

*Task not explicitly supported by the device.

Figure 2 summarizes the number of high-severity and low-severity problems found by speech and nonspeech experts for each device. The total number of problems found for Google Home and Amazon Echo were similar—84 and 83, respectively. The evaluators found 33 percent more problems (112) with Apple Siri.

The four speech experts found 70 percent of the total number of problems, which is significantly higher than the 45 percent of problems found by the four nonspeech experts: $t(18) = 4.152$, $p < .001$. Surprisingly, only 15 percent of the problems were found by more than one evaluator; the overlapping percentage of problems found on Google Home was especially small (7 percent). There was greater overlap finding problems among nonspeech experts (28.4 percentage) than among speech experts (9.4 percent).

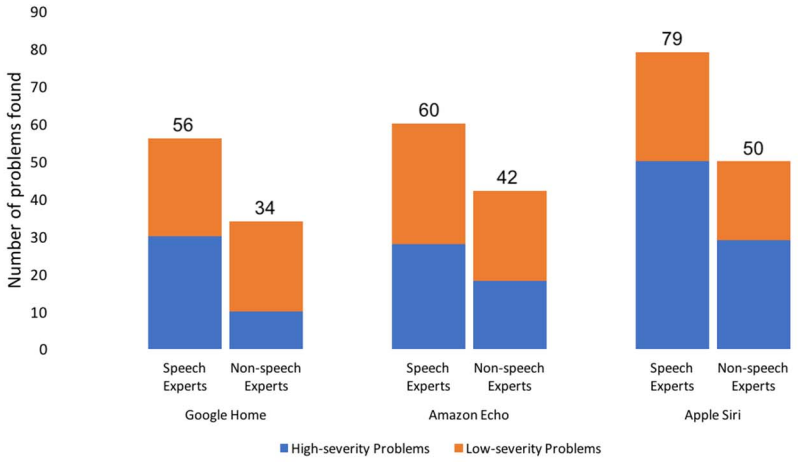

Figure 2. High-severity and low-severity problems found by speech experts and nonspeech experts on each device.

The evaluators found 141 high-severity problems, which is 50 percent of all the problems. The speech experts found 77 percent of these high-severity problems and the nonspeech experts found 40 percent, the same pattern observed in the entire set of problems. However, speech experts found even more of the high-severity problems with Google Home (83 percent) than nonspeech experts (28 percent).

Statistically, the number of total problems and the number of high-severity problems found by speech experts were both significantly higher than those found by nonspeech experts: $t(30) = 4.478$, $p < .001$, and $t(30) = 4.074$, $p < .001$, respectively. Apple Siri had significantly more high-severity problems than both Google Home and Amazon Echo: $F(2) = 3.133$, $p1 < .05$, $p2 < .05$.

We considered all the problems found by all of the evaluators as an estimate of the ground truth for the total number of problems existing in each SUI. Using the accumulated data, Figure 3 shows that 3 evaluators found 70 percent of the problems and 5 evaluators found 85 percent of the problems, which aligns with Jakob Nielsen and Thomas Landauer's mathematical model of the finding of usability problems.[13]

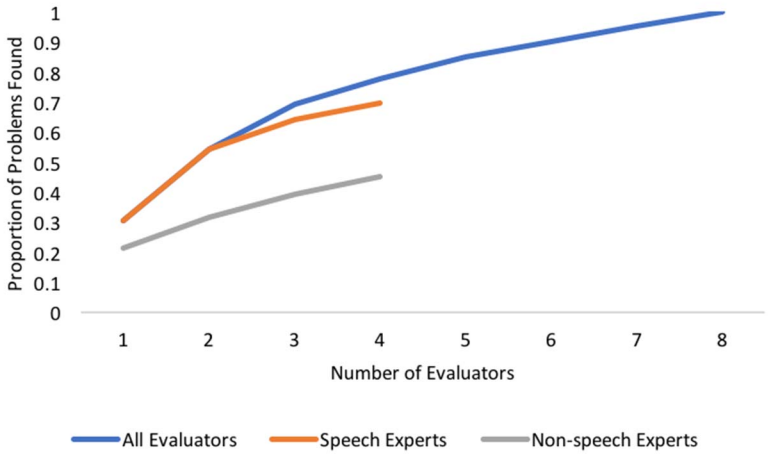

Figure 3. Average proportion of problems found as a function of the number of evaluators by all evaluators, speech experts, and nonspeech experts.

Heuristics Used

The evaluators used all 17 heuristics, with 2 heuristics, S5 and S12, accounting for over 26 percent and 5 heuristics accounting for less than 11 percent of the total problems found (see Figure 4). S5 accounted for most high-severity problems (18 percent), with S9 the second most common in that category (10.6 percent).

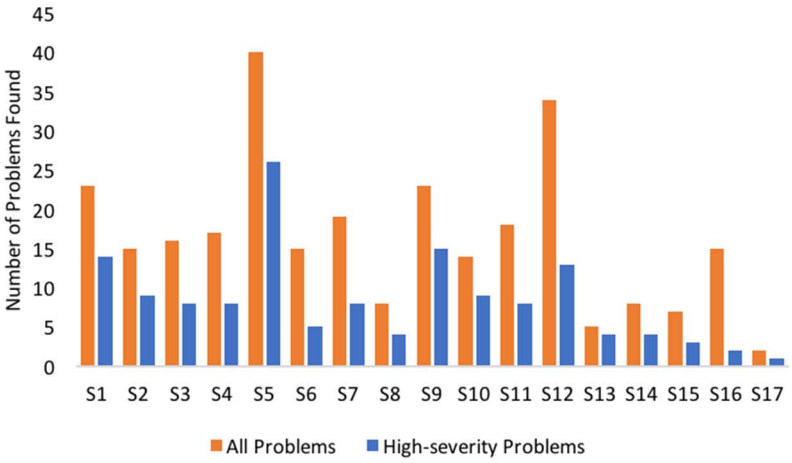

Figure 4. Number of all problems and high-severity problems identified using each heuristic.

S5, S12, S9, and S1 were the 4 heuristics most frequently used to find both all problems and high-severity problems. S7 was also frequently used to identify all problems, but not high-severity problems. S17 was only used twice to find all problems. S8, S13, S14, and S15 were each used to identify less than 3 percent of problems. Interestingly, S16 was used to find 5.4 percent of all problems but only 1.4 percent of high-severity problems. In general, the heuristic violations seem well-distributed.

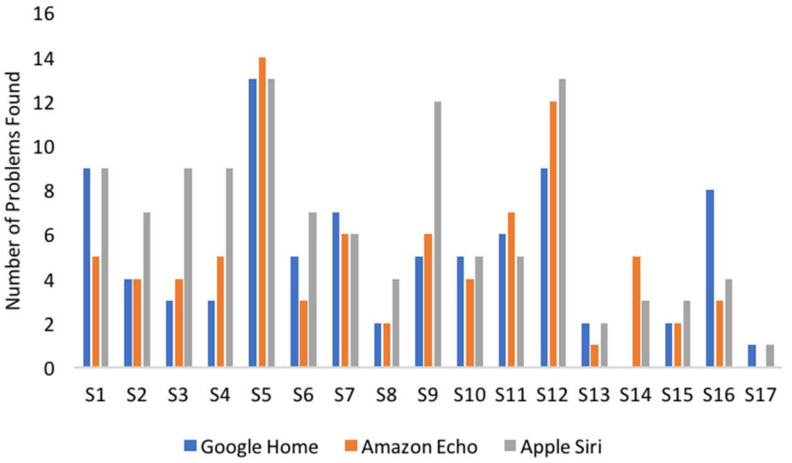

Figure 5. Number of problems found using each heuristic in Google Home, Amazon Echo, and Apple Siri.

Figure 5 shows the number of problems found using each heuristic for the three devices. S1, S5, S7, S11, S12, and S16 were used to find more than 5 problems in Google Home. A similar set were used to find more than 5 problems in Amazon Echo: S5, S7, S9, S11, and S12. A more diverse set of heuristics—S1, S2, S3, S4, S5, S6, S7, S9, and S12—was used to find more than 5 problems on Apple Siri. The evaluators found more problems with Apple Siri, which is different from the others in being screen-based.

Key Problems

We performed a frequency analysis of the 279 unique problems to group together the common types of problems that violated the most frequently used heuristics. We describe and explain these in more detail below.

S5: Pay Attention to What the User Said and Respect the User's Context

In many instances, the device ignored what the user said or only got part of the user's input. For example, when an evaluator requested information about books by Daniel Kahneman, the Amazon Echo typically responded with, "Audible lets you experience books in a whole new way. To try one, ask me to read *The Hobbit* or *The Great Gatsby*." Even when the evaluator tried this query multiple ways, the device continued promoting Audible. Annoyed that "it keeps giving me ads," one evaluator said she would "walk away" in real life. Amazon Echo correctly answered this query only a couple of times. Similarly, Apple Siri responded with "These books written by Daniel Kahneman are available on iBooks" or "Looking for books on iBooks." When the evaluator tried to refer to iBooks, the device said, "OK, here is iBooks," but as the iPad was face down the evaluator was not sure if it opened the iBook or not. Only Google Home correctly responded to this query, probably because Google is better at search. With Amazon Echo and Apple Siri, evaluators were unsure whether the system could not do something or they did not structure the question properly, so they kept trying.

Another problem was that the devices did not respond to follow-up questions, even in the same conversation. For example, in response to a prompt by an evaluator, Amazon Echo and Google Home would provide a list of restaurants. However, when the evaluator asked for the hours of the "first restaurant," the devices could not understand the request. Similar problems occurred with other questions that needed clarification. The evaluator usually had to wake up the device and restart the conversation. As one of the evaluators noted: "A lot of wake word speaking becomes tedious. In some ways, if certain queries result in follow-up questions, consider keeping the dialog open."

Finally, the devices did not always respect the users' context. When asked for the weather, for example, Google Home and Apple Siri obtained the evaluator's current location and then responded with the local weather. Amazon Echo, however, defaulted to Seattle. When the evaluator explicitly asked for the weather in "Sunnyvale" it gave the correct answer, but when asked the follow-up question "Will it rain on Friday?" it again told the evaluator the weather in Seattle.

S12: Confirm Input Intelligently

The devices sometimes lacked implicit confirmation. When asked to play a particular song, Apple Siri started playing the song without providing its name, leaving the evaluator unsure whether it was the correct song. Similarly, when prompted to set a reminder, Apple Siri responded with "OK, I will remind you" without confirming that she did so and when exactly she would execute the reminder. Likewise, Google Home assumed "2 o'clock" was "2 pm" and did not confirm this with the evaluator.

The devices also failed to explicitly confirm some critical actions. For example, when asked to cancel an alarm, Amazon Echo did not ask the evaluator which one and simply canceled the alarm the evaluator had just set.

S1: Give the Agent a Persona through Language, Sounds, and Other Styles

Most of the persona-related problems were found by one speech expert, who had a lot of experience designing personas for smart devices. Of all three devices she wrote, "The persona is not consistent; the inconsistencies themselves are distracting. For example, the visual light, the prompts, and the behavior do not have adequate coherence through time in order for me to perceive a coherent personality." The lack of persona makes it hard to distinguish the devices from the voice alone. Most of the evaluators said it was hard to evaluate the devices' personas because they are generic.

S9: Guide Users through a Conversation so They Are Not Easily Lost

The devices often failed to provide user guidance. For example, Amazon Echo continually promoted Audible without giving any other cues or responses. The evaluators became confused about what was wrong and even felt they were being ignored. Amazon Echo repeatedly replied that "I do not have that. Would you like to hear this?" One evaluator noted that instead it should communicate cues of what it can do—and certainly not guide the user to Audible. When Google Home cannot support something, it responds with, for example, "Sorry, I cannot 'send text' yet" or "Sorry, I cannot do that, I am still learning." One evaluator noted that the device should instead "tell me when it will be supported, or send a message to customer support, or notify me when it's supported."

S2: Make the System Status Clear

The evaluators sometimes had difficulty maintaining a conversation with a device. It was easy for the evaluators to ignore the LED feedback, especially when they were not directly looking at the device. There were times that Google Home and Amazon Echo cut off and stopped listening while an evaluator was still speaking. The devices contain sounds to indicate when a conversation is starting and stopping, but these sounds are turned off by default and must be activated in the app settings—a feature even we were unaware of until one of the evaluators requested we turn on the sound. More importantly, these devices either do not offer a physical exit mechanism or it is not obvious to users, as the evaluators had to speak loudly to stop the conversation or simply wait for it to stop.

The evaluators also criticized the devices' multimodal feedback. In the case of Apple Siri, the GUI was unusable because we placed the iPad face down, yet Apple Siri still referred to the GUI quite often even when it knew the device was face down. In addition, when asked about books by Daniel Kahneman it said "OK, here are some books" without reading out the list. Likewise, when asked for the best noodle restaurants nearby, it responded, "OK, here is a list of restaurants" without saying them. Google Home and Amazon Echo both have a companion app, and when they cannot do something such as change a setting they will respond with something akin to "please change your zip code/delivery address on your app." The evaluators observed that it would be preferable if the app automatically pulled up the required screen so that the user does not have to search for it.

S7: Make Conversation a Back-and-Forth Exchange

Just as the devices usually cannot answer follow-up questions, they do not ask if users want to learn more. As one evaluator noted, "After listing the noodle restaurants, it doesn't ask if you would like more information about those restaurants. User has to use the wake word again and start from scratch." This prevents the device from engaging in a real "conversation" and limits it to being a command-based voice response system.

The evaluators also commented that the devices do not take turns well when interacting with the user. They closed their microphone when the evaluator was still in the middle of a request and would prematurely respond. When reading a list of items, all three seemed to ignore the user's request even if it was "stop."

S10: Use Responses to Help Users Discover What Is Possible

Similar to problems that violated heuristic S5, all three devices lack discoverability of functionality. One evaluator said of Google Home: "Let me know what is available if something like local news isn't available. I had to use my expertise to get the news." Several evaluators noted that the system did not teach ways to ask for a result—the evaluators themselves had to guess and try multiple times. It should, one evaluator said, let the user know what is possible, rather than always say something is impossible. "The inability to do something is presented as a barrier to further engagement."

S11: Keep Feedback and Prompts Short

The evaluators noted that the devices' responses were not always clear or succinct, making it difficult for users to listen, understand, and remember. For example, when Google Home presented a list of books written by Daniel Kahneman, one evaluator said "it is hard to distinguish the title, unable to tell where one book title ended and the next title began." Also, when asked about the weather and restaurants, both Google Home and Amazon Echo responded with multiple items and kept reading them until the evaluator requested the device to stop: "As a user, I'd expect a quick overview and then be prompted if I need more details. That's not what it did." One speech expert noted that the system should not exceed listing three items, which aligns with a study showing that core verbal working-memory capacity is only three chunks.[14] This holds across list lengths and types.

Subjective Responses

We used the SUS—a simple, 10-item Likert scale for evaluating subjective assessments of usability[12]—to evaluate the study participants' perceived usability of all three devices. The average scores of the 8 evaluators (SUS scores range from 0 to 100) were 67.2 for Google Home, 65.0 for Amazon Echo, and 49.7 for Apple Siri. These scores are consistent with the total number of problems found on each device. Apple Siri's score is significantly worse than that of Google Home and Amazon Echo: $F(2) = 121.079$, $p1 < .001$, $p2 < .001$. Almost every evaluator currently used Apple Siri or had used it in the past but still found it the most undesirable. Also, although Google Home supported the smallest percentage of the tasks, all the evaluators agreed that it had the best user experience.

Heuristics Feedback

All of the evaluators said that the heuristics and accompanying examples helped them to evaluate the devices more thoroughly. The evaluators also provided good suggestions on how to improve the heuristics.

We initially had 20 heuristics, and it took our first evaluator, a nonspeech expert, about half an hour to read, ask questions about, and understand all of them. After this first evaluation, we decided to merge some of the heuristics to get the number down to 17 and added more explanations and examples to each one. For subsequent evaluators this made the heuristics easier to understand but also required more time to read and made them harder to memorize. In fact, 17 heuristics might still be too many. The evaluators read through all of the heuristics before undertaking each of the 10 tasks.

Most of the evaluators reported that the heuristics had a lot of overlap, sometimes making it unclear which one to use. For example, S2 and S14 both refer to multimodal feedback, in the former case to indicate system status and in the latter to deliver feedback or prompts. Also, S4 is about starting and stopping conversations and S17 is about exiting from a conversation, which is related. S17 usage was very low (0.7 percent), leading us to consider eliminating it or merging it with another heuristic. Likewise, S3, S8, and S16 all touch on language consistency. Ambiguity about the proper heuristic to use is a common complaint about Nielsen's heuristics as well. It is less important than finding the problem, but there might be a better way to structure and categorize the heuristics.

The evaluators pointed out that some of our heuristics are not applicable to today's smart devices. For example, S8—"adapt agent style to who users are, how they speak, and how they are feeling"—is too advanced for the devices in our study. Should we evaluate devices based on the ideal user experience or their current technical capability? Also, most evaluators found it hard to apply S1, the heuristic about giving the agent a persona, without some standard for what constitutes a good persona.

The speech experts had more comments on the scope of the heuristics given their experience in SUI design. For example, speech-based smart devices are starting to support multi-speaker identification, yet we did not include anything in the heuristics about this topic. Also, multimodal input/output and multi-device interaction might become more prevalent in the future. Our heuristics include some information concerning multimodal principles, but we do not touch on these problems deeply. One evaluator asked, "Are we testing one assistant on one device or one assistant across multiple devices?" Nowadays, the same assistant works on different platforms or devices—for example, Amazon Echo's Alexa is featured on mobile phones, Echo-family smart speakers, and other appliances. It is important to make sure that the user experience is consistent across platforms.

LESSONS LEARNED

Based on the results of our evaluation, here we discuss the problems shared by speech-based smart devices as well as problems unique to each device. We also discuss the usefulness of our heuristics and how they might be improved.

General Problems with Speech-Based Smart Devices

Even with usability and speech experts as participants, our study shows that users do not know exactly what speech-based smart devices can and cannot do. Although users have lower expectations communicating with these devices than with humans, they would like the interaction to be comparable. However, it is difficult to know a given machine's capabilities and how to adapt to its way of speaking. The evaluators in our study found 279 unique problems, and half of these were high-severity ones. Even accounting for current technical limitations, especially for natural language understanding, we believe that system designers could deliver a better user experience in at least four ways.

First, more effort should be put into error handling. Instead of constantly apologizing about what it cannot do, the device interface should guide users and help them to discover what is possible. This not only makes the user feel more confident using the system but also enables longer and richer interactions.

Second, these devices should provide more effective multimodal feedback to make the system status clearer. Users feel lost, angry, or even ignored if they do not know what is happening. All the evaluated devices lack both implicit and explicit confirmations. As indicated in heuristic S13, it is better to "use speech-recognition system confidence to drive feedback style." Designers should not assume that what users hear is correct—confirming a response shows respect for the user and can prevent errors.

Third, systems should leverage human conversational strategies, such as turn-taking and discourse markers. This will not only make interaction more natural but can also help prevent the system from cutting off a user or stopping too early. Discourse markers can also be used as a type of implicit confirmation.

Finally, designers should create a consistent persona. Computers are social actors, and voice is a social tool. As Nass and Brave point out, the key to meeting this goal is creating a consistent voice and emotional range.[9]

Problems Particular to Each Evaluated Device

As speech-based smart devices, Google Home, Amazon Echo, and Apple Siri have almost the same set of functionalities. However, they have different focuses that make them competitive in different areas.

As search is one of its core competencies, Google Home is better at answering general questions than the other devices but has less built-in functionality. At the time we conducted the evaluation (July/August 2017), Google Home could not support basic functions such as messaging, to-do lists, and reminders.

Amazon Echo offers more than 15,000 add-on "skills," and the experience using these is different from using the device's built-in functionality. For example, users must say the exact command to open a skill, and they must first exit from a skill to do something else. Consequently, we did not include these skills in our tasks. Amazon's core business is e-commerce, which it integrates into the Echo. The evaluators complained that the many promotions for Audible and Amazon Prime were annoying.

Our evaluation found that Apple Siri had 33 percent more problems than the other two devices. Some of these issues might be due to the fact that the system is designed to be screen-based and we set it up without the screen. As such, the user experience will likely be different from that of the Siri-powered HomePod (https://www.apple.com/homepod), which was released after our study. This is being marketed as a speaker first and foremost. Unlike Google Home and Amazon Echo, HomePod is completely closed.

The New Speech Heuristics

In developing our new heuristics to evaluate speech-based smart devices, we had three objectives. First, we wanted to provide thorough coverage of the usability problems that can occur when interacting with such devices. Second, we wanted the heuristics to be easy to understand for nonspeech experts so that they too can evaluate existing devices or design new devices with few problems. Third, we wanted designers to identify real problems using the heuristics so we could see how well they worked.

Our evaluation suggests that the heuristics generally meet these objectives. Importantly, the nonspeech experts were able to find many problems without any prior experience with SUIs or any of the devices.

Three of the nonspeech experts found more problems (60, 42, and 39) than two of the speech experts (37 and 24 problems). The other two speech experts found the most problems (86 and 73), which accounted for 55 percent of the total number of problems found. We attribute these results to the fact that the two less successful speech experts relied primarily on their own experience and less on the heuristics, while the two more successful speech experts not only used their experience but also adhered to the heuristics, which helped them find many of the problems that their counterparts did not.

The evaluators clearly understood all 17 heuristics. An inspection of the problem descriptions shows that they matched the area of coverage described by the specified heuristic. Although there is little overlap (15 percent) between the problems found by the speech experts and the nonspeech experts, we are unsure if all the problems are real or whether there are false positives lurking in the set. An end-user study is needed to identify those issues that would affect the user experience. A thorough evaluation of the heuristics will also require additional testing of more devices by more evaluators. However, we believe the results described here provide enough evidence for us to conclude that the heuristics are well suited to uncovering important usability problems in the smart device context.

Because the heuristics were based on prior research that focused on telephone- or workstation-based speech applications as well as the existing design guidelines from Google and Amazon, we believe that they can be used to evaluate most speech-only smart devices. However, whether the heuristics can also be used to evaluate speech-plus-screen devices is unclear. Multimodal I/O will be different from speech-only devices, so we think additional research is required.

There are some clear places we can improve the new heuristics. For example, we could merge heuristics related to multimodal feedback, exiting from a conversation, and language consistency. Moreover, some of the heuristics, such as S8, are more prescriptive and might be too far out for usage today.

CONCLUSION

Our 17 heuristics for evaluating speech-based smart devices are easy for both nonspeech and speech experts to understand and to help identify real usability problems. The heuristics have good coverage of the design space and can also serve as a set of early design principles. Our evaluation of three popular devices by 8 usability specialists serves as an initial validation of the usefulness of the new heuristics, which we will continue to refine with more testing. We hope this research will help and inspire designers to create more effective and user-friendly speech-based smart devices, and inspire researchers to conduct more studies on this finally-ready-for-prime-time interaction modality.

ACKNOWLEDGMENTS

We thank all the experts who participated in this study and gave us invaluable feedback. We also thank former Baidu Research colleagues' support for this study.

REFERENCES

1. S. Ruan et al., "Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones," *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 4, 2016; doi.org/10.1145/3161187.
2. D. Amodei et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin," *Proc. 33rd Int'l Conf. Machine Learning* (ICML 16), 2016, pp. 173–182.
3. B. Suhm, "Towards Best Practices for Speech User Interface Design," *Proc. 8th European Conf. Speech Communication and Technology* (Interspeech 03), 2003, pp. 2217–2220.
4. A.I. Rudnicky, "Speech Interface Guidelines," 1996; http://www.speech.cs.cmu.edu/air/papers/SpInGuidelines/SpInGuidelines.html.
5. N. Yankelovich, G.-A. Levow, and M. Marx, "Designing SpeechActs: Issues in Speech User Interfaces," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI 95), 1995, pp. 369–376.
6. N. Yankelovich and J. Lai, "Designing Speech User Interfaces," *CHI 98 Conf. Summary on Human Factors in Computing Systems* (CHI 98), 1998, pp. 131–132.
7. J. Nielsen, "Finding Usability Problems through Heuristic Evaluation," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI 92), 1992, pp. 373–380.
8. L. Dybkjær and N.O. Bernsen, "Usability Evaluation in Spoken Language Dialogue Systems," *Proc. Workshop Evaluation for Language and Dialogue Systems* (ELDS 01), 2001; doi.org/10.3115/1118053.1118055.
9. C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*, The MIT Press, 2007.
10. C. Pearl, *Designing Voice User Interfaces: Principles of Conversational Experiences*, O'Reilly Media, 2016.
11. J. Nielsen and R. Molich, "Heuristic Evaluation of User Interfaces," *Proc. SIGCHI Conf. Human Factors in Computing Systems* (CHI 90), 1990, pp. 249–256.
12. J. Brooke, "SUS—A Quick and Dirty Usability Scale: Usability Evaluation in Industry," 1996; http://dag.idi.ntnu.no/IT3402_2009/sus_background.pdf.
13. J. Nielsen and T.K. Landauer, "A Mathematical Model of the Finding of Usability Problems," *Proc. INTERACT 93 and CHI 93 Conf. Human Factors in Computing Systems* (CHI 93), 1993, pp. 206–213.

14. Z. Chen and N. Cowan, “Core Verbal Working-Memory Capacity: The Limit in Words Retained without Covert Articulation,” *Q. J. Experimental Psychology*, vol. 62, no. 7, 2009, pp. 1420–1429.

ABOUT THE AUTHORS

Zhuxiaona Wei (Nina) is a product manager at deeplearning.ai and was formerly a product designer at Baidu Research, where this work was conducted. Her research focuses are SUI/CUI design and AI-powered products. Contact her at weizhuxiaona@gmail.com.

James A. Landay is a professor of computer science and the Anand Rajaraman and Venky Harinarayan Professor in the School of Engineering at Stanford University, specializing in human–computer interaction. This work was part of his consulting work with Baidu Research. Contact him at landay@stanford.edu.