Siri vs Google Now: A Comparative Evaluation

Harshad Golwalkar
RIT
Rochester NY
hg3583@g.rit.edu

Nidhi Palan
RIT
Rochester NY
nxp4195@g.rit.edu

Saish Vaniyamparambath
RIT
Rochester NY
sv3912@g.rit.edu

Shengjiao Wang
RIT
Rochester NY
sw4302@g.rit.edu

ABSTRACT
Intelligent personal assistants such as Siri, Google Now and Cortana has gained increasing interest in recent years. People can use their voice to access some of the functionalities on the mobile devices. However, they are often not as easy to use as they are advertised. Evaluating the performance of these assistant is challenging especially because of the variability of the tasks supported. We intended to find out which of Siri and Google Now perform tasks faster and with less errors. We conducted an experiment where participants were asked to perform tasks with Siri and Google now. We used time and number of failures taken to complete all tasks as two dependent variables to comparatively evaluate these two systems. Our results from the experiment show that there is no significant difference between Siri and Google Now in terms of time, but Google Now is less error-prone than Siri.

INTRODUCTION
Intelligent personal assistants (will be referred to as "intelligent assistants" for the rest of the paper) are becoming one of the most prevalent features on mobile devices. Smart phone users can perform many tasks using their voice: managing schedules, making phone calls, seeking information from a variety of online sources etc. There have been an increasing number of speech recognition based assistants from various tech companies such as Apple, Google, Amazon, etc. Although intelligent assistants look promising on commercials and demonstrations by their makers, our experience with them in reality is far from ideal. Driven by curiosity, we wanted to find out how well intelligent assistants fulfill their functions in daily lives.

However, it is challenging to evaluate intelligent assistants. Some studies compare system generated answers with human annotated answers [1]; some use correlation ratings of user experience and behavioral signals as major indicator of user experience [2]. Due to limitations of our knowledge and experience with the subject, we decided to use time taken to complete tasks and number of failures as measurements of evaluation. We chose two intelligent assistants to evaluate: Siri and Google Now because the devices were available, they have fairly larger market shares, and it was easier to recruit participants with experience using them. In this paper, there are two goals we try to learn: which of Siri and Google Now performs tasks faster; which of Siri and Google Now is more error-prone for users. To answer these questions, we set up two hypotheses to be tested by experiments.
Hypothesis 1:
H0: There is no difference in time taken to complete tasks between Siri and Google Now
Ha: There is a difference in terms of time used to complete tasks between Siri and Google Now

Hypothesis 2:
H0: There is no difference in the number of failures that occurred before tasks are completed between Siri and Google Now
Ha: There is a difference in the number of failures that occurred before tasks are completed between Siri and Google Now.

METHODS
To test the two hypotheses stated earlier, we conducted an experiment with 11 participants recruited from HCIN 600 class at RIT. Each participant performed four tasks (shown below) under two treatments: using Siri on an iPhone 6S plus and using Google Now on Nexus 5X. Before the experiment began, each participant was given detailed instructions on how to do the experiment and was told specifically that they should use voice only to complete the tasks.

Then after the participant agreed to the experiment, he/she read one task at a time on paper and completed it on one device. After all four tasks were completed, he/she switched to the other device. Two dependent variables (time and number of failures) were measured during the experiment.

One of experimenters measured the time with a smartphone and the number of failures the participant had. The time measured was the difference between when the participant tapped the microphone button on the phone and when the device produced the desired outcome. The number of failures was measured by counting the number of times the participant started over on the task. The second experimenter monitored the task and notified the participant to start over on the tasks when the assistant produced the

wrong result. After the experiment was completed, each participant received candy as reward.

The procedure involved each participant doing four verbal queries (tasks) in the same order as shown below. The following were the four tasks:

1. What's the weather like today?
2. Set an alarm for 7:30 pm today.
3. Text Andy "How's it going?".
4. Schedule a meeting at 2 pm on Tuesday for Research Methods.

Participants were instructed to use voice control exclusively and avoid using the screen. Participants received no instructions during the tasks but were notified whether the task was complete or not. A task is considered complete if the phone responds with exactly the desired outcome (for example, a text with the right content is sent to the right person).

An iPhone 6S Plus and a Nexus 5X were provided to participants to use because these two devices have similar specification (same sized screen, similar weights etc.). Devices were tested before each participant started using them.

We used within-subject design and each participant performed under two conditions: using Siri and using Google Now. The reasons for using within-subject design are two-fold. First, the sample was very small. Within-subject design allows us to apply two conditions (Siri and Google Now) on the same group of participants, thus only half of what between-subject design requires is needed. Second, there is no need to manage variance between groups. Given that participants recruited had very varied levels of proficiency in English, individual differences would be large and create noise if we chose between-subject design. Within-subject design avoids this problem.

The order of conditions for participants was randomized to counterbalance learning effect. Half of the participants used Siri first and then Google Now; the other half of the participants used Google Now first and then Siri. Because we choose within-subject design, counterbalancing must be considered to counteract the learning effect. Since participants will repeat identical tasks on both Siri and Google Now, they might become better at the tasks when they perform them the second time.

The four tasks were chosen for various reasons. Only straightforward and unambiguous tasks were considered so that participants wouldn't misunderstand the tasks. Also, enough details are provided in the task so that every participant would have identical results from the tasks. In this way, the completion of each task can be clearly defined and noted (for example, a text with the right content is sent to the right person). Tasks were designed to minimize variances as a result of the device. Since Siri can only be used on an iPhone and Google Now on an Android phone, the differences between these two devices might create unwanted variance (for example, the reminder app works differently on these platforms and we didn't ask the participants to set up a reminder). Therefore, all tasks chosen can be performed on both devices in the same manner. Also, all tasks were short to avoid fatigue effects. Under two conditions (Siri and Google Now), the participant completed 8 tasks in total. In our pilot study, it took from 3 to 5 minutes to complete all the tasks.

Data from one participant was removed because he/she failed to follow instructions by using the keyboard for one of the tasks. Therefore, we collected data from 10 participants in total (5 females and 5 males).

RESULTS

Time taken to complete four tasks on each assistant is summed for each participant; this was also done for the number of failures. Thus, each participant has four data points: total time on Siri, total time on Google Now, the total number of failures on Siri and total number of failures on Google Now. Histograms of time and number of failures for Siri and Google Now were made to determine the normality of the sample distribution. The histograms show that the sample was not normally distributed. The sample size (10) is too small to assume normality according to Central Limit Theorem. Therefore, we use a nonparametric test, Wilcoxon Signed-rank test to test our hypotheses. The study is a within-subject study where every participant performed all the tasks under two conditions.

	Siri	**Google Now**	**Difference**
Median Total Time (In Seconds)	82.205	64.855	17.35
Median Total Failures	2	1	1

Table 1 Medians of Total Time and Total Failures for Siri and Google Now

We conducted two tests. The first test indicated that there was no significant difference in the time needed to complete the task using Siri or Google Now ($Z = -1.5799$ and $p = 0.1141$) at significance level 0.05. The second test indicated that there was significant difference in the number of failures in the tasks completed by using Siri or Google

Now ($Z = -2.5205$ and $p = 0.014$) at significance level 0.05.

DISCUSSION

Based on the results in the previous section, we failed to reject h0 for hypothesis 1, which states that there is no difference in terms of time to complete tasks on Siri and Google Now. Our experiment showed that there was an insignificant difference in time taken to complete tasks between the two personal assistants. We can say that the time taken to complete the task on devices did not have a significant difference.

We reject h0 from hypothesis 2 since the p-value of results is less than the 0.05 significance level. Hypothesis 2 states that there is no difference in terms of the number of failures that occurred before tasks are completed between Siri and Google Now. Our data shows that Google Now (mdn = 1) had lower number of failures and was less error prone compared to Siri (mdn = 2).

In considering these results, we need to take into account a considerable number of limitations which might have created biases while studying the two personal assistants. The research has following limitations.

The task of recording the time by starting and stopping stopwatch was done manually by one of our team members. Since it was done manually, there might be a chance of it not being precise. To record accurate measurements on time, a native program that is able to track user interactions on the device is required. We used the number of times participants try until the successful completion of tasks as an estimator of error. People make errors that can be corrected without restarting the task such as changing the title of an event or editing the content of a text, and our experiment failed to account for them.

The two assistants were run on two different devices and the results might be influenced by the type of devices used in the experiment. The variance could have been reduced by asking the participants the type of phone they use and then conducting the test on another device which participant is not comfortable in using due to lack of practice.

The experiment was conducted in a lab setting which might make participants self-conscious. This would have either reduced the speed of performing a task due to nervousness or increase the speed due to the extra attention and effort they put in the tasks.

Another limitation of the study is the small sample size and the sample is primarily composed of non-native English speaking students at a university. Thus the sample does not sufficiently reflect the population. With more time and resources, we would have collected data from a larger and more diverse sample to get more reliable results.

The difficulty level of the tasks that we asked our participants to carry out was low due to time constraints. It is not certain that our findings will apply to more complex tasks. Also as tasks get more complex, there are more differences between devices that need to be accounted for. In addition, tasks were presented as a line of text in the experiment where participants simply needed to read the texts. However, the wording of tasks plays an important role in speech recognition and different wording of the same task might produce different results.

CONCLUSION

The paper compares two intelligent assistants, Siri and Google Now. We intended to find out which one performs better in terms of speed for completing tasks. We conducted experiments on participants who were asked to perform simple tasks using Siri and Google Now. Data from our experiment suggests that the two intelligent assistants perform equally in terms of time taken to perform although people make fewer errors on Google Now than Siri. As we discussed, there are many limitations to our study. These limitations can be addressed by future studies. Our findings can result from many factors and future studies further explore why Google Now is less error-prone than Siri.

REFERENCES

1. Jiepu Jiang et al. 2015. Automatic Online Evaluation of Intelligent Assistants. *Proceedings of the 24th International Conference on World Wide Web - WWW '15* (2015). DOI: http://dx.doi.org/10.1145/2736277.2741669

2. Ahmed Hassan Awadallah, Ranjitha Gurunath Kulkarni, Umut Ozertem, and Rosie Jones. 2015. Characterizing and Predicting Voice Query Reformulation. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15* (2015). DOI: http://dx.doi.org/10.1145/2806416.2806491

.