How Good Are Mobile Virtual Assistants?

Tobias Mertz

Texttechnology
Goethe-University Frankfurt am Main

July 7, 2016

1 Abstract

Since Mobile Virtual Assistants are rising in popularity and come with most new smartphones out of the box and theoretical work in the field is hard to come by, a test is in order to establish the status quo of development. We did a manual test on six different Mobile Virtual Assistants in the categories Voice Recognition, Online Search, Phone Control and Natural Conversation and the results show that Siri is currently the best Mobile Virtual Assistant on the market with a success rate of 65.8% on average over all four categories.

2 Introduction

There have been many stories dealing with Artificial Intelligence (AI) throughout science fiction media for the last one hundred years with one of the oldest examples being Samuel Butler's 'Erewhon'[1] from 1872. This constant appearance of AI in science fiction is a sign for the fascination many people have with the idea of a 'living' machine. Since Apple's introduction of Siri on the iPhone 4S in 2011 more and more devices appear on the market, which supposedly replicate this idea in real life.

However with the release of new technologies always appears the problem of the distinction between marketing based texts and neutral reviews, which is why it is often times hard to find information about how well this new technology really works and how far advanced research in the field is.
Therefore one should try to keep an overview of the progress in research by determining the goal and the current state of the development. The goal is very easy to find in the aforementioned science fiction media[1]. But the current state is difficult to filter out of the promises made by most companies when trying to sell a product.

With the current lack of neutral scientific work in the field of Mobile Virtual Assistants you can only get this information by testing yourself.

3 Definition

Before we head into the preparations for the test itself we should first clarify, what conditions the test underlies. So we need to define our test subject exactly before we can continue.
There are many differing definitions for Mobile Virtual Assistants. Most of them are either very long and not to the point or do not provide the complete picture. The approach taken in this paper to define Mobile Virtual Assistants was to look for the definitions of the words, which make up the multitude of names for the subject, and to create a composite definition out of those. The most prominent of these names were 'Mobile Virtual Assistant', 'Personal Assistant'[2], 'Intelligent Personal Assistant' and 'Digital Assistant'[3]. The definitions of the individual component words can be found in the appendices (14).
Out of these components, we created the following definition:

Intelligent Mobile Virtual Personal Assistant:
A software that simulates a person, that helps a specific individual in particular kinds of work and is able to vary its action in response to varying situations and past experience as well as being able to be moved freely or easily.

The word 'digital' has been left out of the composite because of its significant overlap with the word 'virtual'. For the sake of simplicity, we will however continue to use the term 'Mobile Virtual Assistant' (MVA) synonymously.

4 Prior Assumptions

Going into a test of this kind, we need some sort of prior assumptions, that we can either verify or invalidate through our results. Our goal in this experiment was to be as neutral as possible, which lead us to our first approach on an initial hypothesis. *Current MVAs have a success rate of 50%.* This is the most neutral statement possible across all fields, because it is exactly in between the statements "everything works", which would correlate to 100%, and "nothing works", which correlates to 0%. If we talk about technological devices, the most neutral state is however if there is an exactly

[1] such as Marvel's "Iron Man", Microsoft and Bungie's "Halo" or Gene Roddenberry's "Star Trek"

50% chance of adoption for this particular device, meaning there is a 50% chance, that a consumer would find this device useful enough to spend the asking price on it. An MVA with a 50% success rate would not be considered useful by the largest part of potential consumers, which makes an expectation of a 50% success rate a negatively biased stance on the matter. There are however not many works out there to determine, which rate of success in a new technology would lead to an adoption rate of exactly 50%, which is why we did a small survey on this topic as well.

We asked 15 people, what largest rate of failure they would deem acceptable in a new device they were buying. The results of this survey showed that the average viewed 15% as the maximum acceptable rate of failure and therefore 85% as the minimum acceptable rate of success. We chose to use these 85% as initial hypothesis, since it represents the average value, which was viewed as enough. Assuming that demand by the consumer is normally distributed, this percentage would lead to a 50% adoption rate.

5 Candidates

Through this test we wanted to get an overview over the market as a whole. That means we needed to find suiting candidates for the test. The first names that come to mind are those of the largest mobile operating system manufacturers themselves, namely Google, Apple and Microsoft. But there are also a lot of lesser known alternatives out there that deserve to be mentioned. Since we personally only had an Android device available, an iPhone and Windowsphone test device had to be organized from other parties, which lead to a short time frame that could be utilized to test these devices. That is the reason, why the alternative MVAs in this test are Android apps only.

Since we already had three candidates in the market leaders, we chose three candidates as well for the lesser known alternatives. Those had to be on Android and preferably free. They could easily be found using the website `alternativeto.net` [4] and searching for 'Google Now'. The search results that were most promising and fit our preferences best were Speaktoit Assistant, Alicoid and Evi. This left us with the following candidates in the corresponding versions:

1. Evi (v. 1.2.41.0_1024610) [5]
2. Cortana (newest Alpha as of 6th of July 2015) [6] [7]
3. Speaktoit Assistant (v. 3.1.16) [8]
4. Siri (iOs v. 8.3) [9]
5. Google Now (v. 4.7.13.19.arm) [10]
6. Alicoid (v. 2.9.6t Free) [11]

6 Test-Queries

In addition to the selection of the right test participants, we needed to find a suitable test set of tasks, that represents as much of the variety of features MVAs provide as possible. The best overview of an MVA's features can be found on the website of the developer. We read through each of the candidates' online specifications as well as some other MVAs', that were not part of the test:

- Dragon Go! [12] [13]
- Sherpa [14]
- Hound [15] [16] [17]
- Viv [18] [19]

These MVAs were not included in the test because they were not available for free on Android in Germany at the time of testing.

While creating an overview of the features, we gathered, that these features could be categorized in one of five categories. These categories were the following:

6.1 Voice Recognition

In the Voice Recognition category we try to test how good MVAs recognize spoken words in different circumstances. When testing Voice Recognition, there are four variables, one can easily change to create different test cases. We chose to use the following categories for each of the variables.

- complexity in sentences: simple, colloquial / vulgar, scientific

- talking speed: slow, medium, fast
- language: German, English
- background noise: silence, music, train, crowd

To make the test as fair as possible, we decided to use recordings of music, train and crowd noises. The music was supposed to simulate a radio, that was playing in the background, so we needed music, that was ambient in its nature and not too energetic but it still needed some clearly noticeable breaks and beats to simulate change in between songs. The most fitting example of such music, we came up with was the soundtrack to "Frozen Synapse" by nervous_testpilot [20], which can also be found in this video [21]. For the train and crowd noises we also used Youtube videos containing looped recordings of these sounds. The videos can be found in [22] and [23]

For each complexity-category we used three example sentences. One of these sentences is particularly important and will serve as an example throughout this paper. This sentence, that is incidentally the most complex of all test-queries in this category, is: *"How does Outlier Detection using Gaussian Mixture Models work?"*. A full list of test-sentences can be found in the appendices at (14.3). As a whole this selection of categories and examples left us with a test set of 216 test cases.

6.2 Online Search

In the Online Search category we try to test features that supply information to the user. This does not only contain searching for information on the web, but also calculation of mathematical terms. We thought of seven different kinds of queries for this category and used three examples for each. An example for a query would be *"Who was Leonardo DaVinci?"* as a question with a complex answer. Since some MVAs might not have the same amount of features in all languages, we tested in German and English. This resulted in a combined number of test cases of 42 (14.4) in this category. Some of these queries should be highlighted individually, because they are somewhat special compared to the others. To test if the MVA can calculate a mathematical term, we tested three different queries, one of which being *"What is the square root of 45 plus four?"*. This query is a special case because we did not specify further, whether we wanted the MVA to calculate $\sqrt{45+4}$ or $\sqrt{45}$+4, but we were curious how the individual MVAs would react. The question *"What is the meaning of life?"* was another special case of a question with a complex answer, since it is a philosophical question and there is no one correct answer for it. The last special query in this category was a question with incomplete information. It was *"What is the difference between a duck?"* and it belongs to the special cases since it is a commonly known joke-question based on the fact, that it does not make sense without additional information.

6.3 Phone Control

The Phone Control category covers tasks that your phone can perform for you. Many MVAs provide interfaces to other apps so these functions are accessible via voice input. We thought of seven different tasks to perform and used three examples each in German and English, one of which being *"Create a new appointment tomorrow at three p.m."*. The total amount of test cases for this category was 42 (14.5).

6.4 Natural Conversation

An MVA is per definition (3) supposed to simulate a person. One of the most difficult tasks in simulating a real person is the act of natural conversation. There is a lot of information to be taken from one's surroundings, gestures, facial expressions and tone of voice. This information is therefore not obtainable from the words alone, but always requires additional context. That's why the MVA has to be able to gather this context and make cognitive connections by itself or at least be able to be programmed with such context by the user. We thought of seven different kinds of context information and small talk and tested three examples for each category in German and English. Combined we had 42 (14.6) test cases for this category. An example would be *"Call me 'Versuch'."*. In this category there is one query that needs to be mentioned specifically. The first kind of queries to be tested in natural conversation was the ability to tell jokes, since this feature was explicitly advertised by multiple of the test candidates and it therefore seems to be important. There were two queries dedicated to the feature of the MVA telling a joke and as last example for this kind of query we thought it might be interesting to see if an MVA could recognize a joke that is being told to it, we used the joke *"I was drinking at a bar, so I took the bus home. Seemed like a good Idea at the time, but I have never driven a bus before."* because it is short and works well in English and German alike.

6.5 Personalized Recommendations

Personalized Recommendations are suggested products, articles or websites, which are adapted to the user's personal habits and preferences. These preferences are learned by the MVA throughout the time of use. That is why this category can only be tested after a long period of training, which could not be achieved in the time frame of this experiment. Therefore Personalized Recommendations were excluded from the test.

7 Procedure

The test was done manually without automization. The phone was placed at a distance of 55cm from the speakers for the Voice Recognition tests. Also the testers were in a constant distance of approximately 30cm to the device when speaking. We tested in blocks whereby one block covered one category in one language. Only after a block was finished for all of the different MVAs, the next block was tested. In the Voice Recognition category, one block consisted of all queries with the same background noise and language. The order of queries was kept constant to provide equal conditions for the MVAs.
After each query, the response of the MVA was rated using the following system:

7.1 Rating of Voice Recognition

2 Points no errors

1 Point one error

0 Points more than one error

With the example sentence *"How does Outlier Detection using Gaussian Mixture Models work?"* the MVA would receive two points if all words of the sentence were recognized correctly, excluding articles. One point would be received for one mistake, for example if "Gaussian" was mistaken for "Goshen" and zero points for more than one mistake. Articles were excluded because of their phonetic and semantic proximity. With the 216 test cases of this category (6.1) we get a maximum amount of reachable Points (MRP) of $216 \cdot 2 = 432$.

7.2 Rating of Online Search

3 Points required plus additional information

2 Points required information

1 Point list of search results

0 Points no result

The example query of *"Who was Leonardo DaVinci?"* would lead to three points if there was a mention of the most important facts about his life, two points if the MVA just named his profession, one point if the question was searched through a search engine and the results presented without any kind of further processing or filtering and zero points if the MVA returned no result at all. With the 42 test cases of this category (6.2) we get an MRP of $42 \cdot 3 = 132$.

7.3 Rating of Phone Control

2 Points task completed with feedback

1 Point task completed without feedback

0 Points task not completed

If the MVA could process the query *"Create a new appointment tomorrow at three p.m."* by automatically creating the appointment and then asking for a name for it, it would receive two points, one point if it could create the appointment but without that additional feedback or it just opened the calendar app and started the creation of an appointment. It received zero points if it was not possible to create an appointment using this query and it either returned no result at all or searched for the query online. With the 42 test cases of this category (6.3) we get an MRP of $42 \cdot 2 = 84$.

7.4 Rating of Natural Conversation

1 Point query recognized and treated appropriately

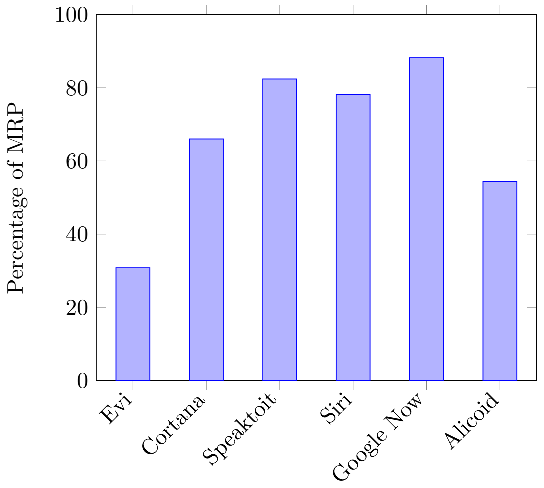

Figure 1: Results in the Voice Recognition category

0 Point query not recognized

If the query *"Call me 'Versuch'."* was recognized and the MVA then continued to call us 'Versuch' it would receive one point. If the MVA did not understand the query or searched for it online, it would receive zero points. With the 42 test cases of this category (6.4) we get the MRP of $42 \cdot 1 = 42$.

8 Results

The tests in each individual category led to the following results:

8.1 Voice Recognition

Table 1 shows the total scores in the Voice Recognition category.

MVA	Score
Evi	133
Cortana	285
Speaktoit Assistant	356
Siri	338
Google Now	381
Alicoid	235

Table 1: Total scores in the Voice Recognition category

The scores as percentages of the MRP are shown in Figure 1. The full results of the tests can be found in the appendices in Table 6 and Table 8. Splitting up these results by language provides the percentages shown in Figure 2.

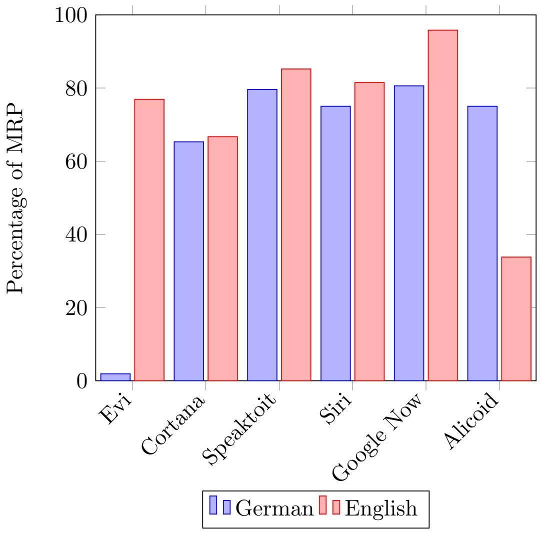

Figure 2: Results in the Voice Recognition category split by language

8.2 Online Search

Table 2 shows the total scores in the Online Search category.

MVA	Score
Evi	28
Cortana	52
Speaktoit Assistant	53
Siri	77
Google Now	77
Alicoid	31

Table 2: Total scores in the Online Search category

The scores as percentages of the MRP are shown in Figure 3. The full results of the tests can be found in the appendices in Table 10 and Table 11.

8.3 Phone Control

Table 3 shows the total scores in the Phone Control category.

MVA	Score
Evi	0
Cortana	52
Speaktoit Assistant	48
Siri	56
Google Now	44
Alicoid	3

Table 3: Total scores in the Phone Control category

The scores as percentages of the MRP are shown in Figure 4. The full results of the tests can be

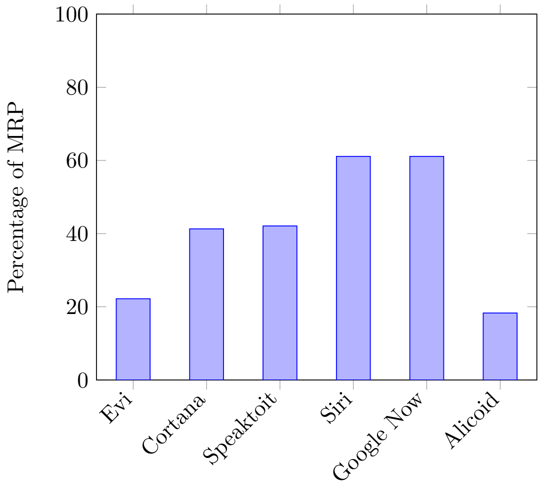

Figure 3: Results in the Online Search category

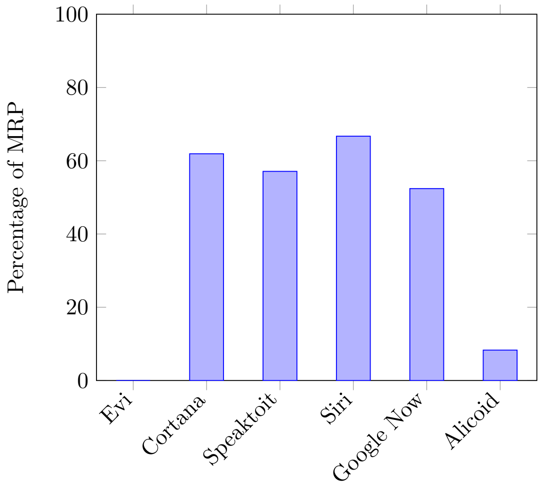

Figure 4: Results in the Phone Control category

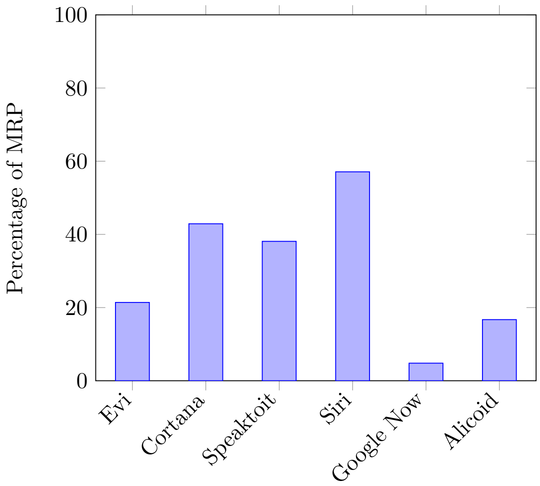

Figure 5: Results in the Natural Conversation category

found in the appendices in Table 12 and Table 13.

8.4 Natural Conversation

Table 4 shows the total scores in the Natural Conversation category.

MVA	Score
Evi	9
Cortana	18
Speaktoit Assistant	16
Siri	24
Google Now	2
Alicoid	7

Table 4: Total scores in the Natural Conversation category

The scores as percentages of the MRP are shown in Figure 5. The full results of the tests can be found in the appendices in Table 14 and Table 15.

9 Evaluation

We can see in the Voice Recognition results (8.1), that Google Now is leading in points with 88.2% of the MRP. Evi and Alicoid are in the last two places. The reason becomes apparent, when looking at the point distribution across the two languages in Figure 2. Evi only has four points in German while Alicoid has only 73 points in English. We can therefore assume that these languages are not supported by the corresponding MVA. Alicoid however uses Google's Voice

Recognition technology, which is able to process English sentences. Since Alicoid however can not handle these English queries any further, it prefers a German alternative sentence that is phonetically close, but not the correct one. This is the reason Alicoid still has this high of an amount of points for the English queries, which have no phonetically close German alternative.

		#1	#2	#3	#4	#5	#6
silence	slow	1	1	1	1	1	0
	medium	0	0	0	0	0	0
	fast	0	0	0	0	2	0
music	slow	0	1	1	0	2	0
	medium	0	0	0	0	2	0
	fast	0	0	0	0	0	0
train	slow	0	0	1	1	2	0
	medium	0	0	0	0	2	0
	fast	0	0	0	0	1	0
crowd	slow	0	0	1	0	2	0
	medium	0	1	0	0	2	0
	fast	0	0	0	0	2	0

Table 5: Results for test-query 3.2 of the Voice Recognition category in English

Another feature of MVAs is very good to see in one particular column of Table 8 and Table 9. Table 5 shows this column as an excerpt from the full tables. The results from query 3.2 *"How does Outlier Detection using Gaussian Mixture Models work?"* show that Google Now (#5) is in first place in this category, because it employs the most sophisticated Machine Learning algorithms and it is therefore the only MVA that ever recognized the sentence correctly throughout the tests. Looking at Google Now's results in this column, we can see that at the first try, there was still an error in its result. It mistakenly recognized the word "Gaussian" as "Goshen"[2]. With the second try Google's Voice Recognition software tried a different interpretation of the (now faster) input. In this interpretation there were however still errors. While it recognized the word "Gaussian" correctly this time, the words "How does" were mistaken for the word "photos". At the third try, even though this was spoken even faster and was therefore harder to correctly identify, it recognized the sentence correctly and from there on, it only ever made mistakes in recognizing this sentence, when it was spoken very fast. But even if we are only looking at the tests in fast speed after the initial training phase with a silent background, there is a linear improvement in the scores noticeable. This technology and this test case in particular is what put Google Now in front of the other test candidates. In German however no MVA could recognize the sentence and only Siri ever could get one point once. The reason for this is that the scientific terms are in English while the other words are in German and MVAs only expect sentences with one language in them.

In the Online Search category we can see that Siri and Google Now are tied for the first place at 61.1% of the MRP. Cortana and Speaktoit Assistant have just over 40% and Evi as well as Alicoid are trailing behind at about 20%. Considering that testing in this category was done in English and German and Alicoid and Evi only support one of the two languages each, it becomes apparent, that this is the reason they have about half as many points as Cortana and Speaktoit. The leaders in this category are ahead of the other candidates because they did very rarely return a list of search results but were able to extract information from these results. Siri only received one point for a query twice and Google Now thrice. They were also more regularly able to deliver some additional information, which is why they scored three points for a particular query more often. Furthermore Siri and Google Now are the only MVAs to produce results on queries that are tied to their last output in English. Speaktoit Assistant and Cortana were also both able to respond correctly to one of those queries in German. The question *"What is the square root of 45 plus four?"* showed that most MVAs were not able to process it, but those that were all preferred the term $\sqrt{45+4}$ over the term $\sqrt{45}+4$ and none asked, which of the two terms was the correct one. To the question *"What is the meaning of life?"* most MVAs responded with definitions, articles and in some cases the number 42 in reference to Douglas Adams' "Hitchhiker's Guide to the Galaxy"[24]. The query *"What is the difference between a duck?"* resulted mostly in lists of search results. Only Evi and Siri were able to recognize the joke-question in English language and respond accordingly.

The results in Phone Control look very close

[2] A place in Egypt that is named in the Bible as well as a city and a village in the USA

between Cortana, Speaktoit Assistant, Siri and Google Now but with Siri taking the lead at 66.7% with a 4.8% lead on Cortana, the runner up. Evi and Alicoid are very far behind with 8.3% and 0%. These low scores can not be attributed to the missing of language support alone, but there have to be features missing as well.

The Natural Conversation category has Siri as its clear leader at 57.1% with a 14.2% lead on the second place, Cortana with Speaktoit Assistant closely behind. Evi as well as Alicoid have about half as many points, which is because of the language support. Google Now is in last place in this category with only two points because it has only one of the 21 tested features implemented. Siri was in this category more consistent than the other candidates and was the only MVA to have scripted responses to questions about its feelings, which is the main reason for Siri's lead on the competition. None of the candidates were able to recognize *"I was drinking at a bar, so I took the bus home. Seemed like a good idea at the time, but I have never driven a bus before."* as a joke in either language.

To sum up these results we needed to calculate a combined score for each MVA, but since we wanted to keep this test as fair as possible, we cannot total up the points and divide them by the number of combined MRP, because the MRP differs between the four categories. A fairer solution would be to calculate the averages over the percentages of the individual categories for each MVA. This procedure leads to the following ranking:

1. Siri (65.8%)
2. Speaktoit Assistant (54.9%)
3. Cortana (53.0%)
4. Google Now (51.6%)
5. Alicoid (23.3%)
6. Evi (18.6%)

These scores are only valid if every category is valued as equally important. So it might not be sufficient data to advise a potential buyer, since they could have their own preferences and may not need features of a particular category as much as others. Also these numbers can lead to false assumptions since another important factor to look at is the worst-case scenario. If we ranked the MVAs after their lowest individual score, the ranking would look as follows:

1. Siri (57.1%)
2. Cortana (41.3%)
3. Speaktoit Assistant (38.1%)
4. Alicoid (08.3%)
5. Google Now (04.8%)
6. Evi (00.0%)

Both these numbers display valuable information to the consumer or critic, and should not be viewed independently. They reinforce however the the fact that Siri is the overall leader with only an 8.7% difference between the average and the worst-case. This information is helpful in the process of answering the initial question of, how good MVAs really are.

To answer this question we can use a set of different approaches depending on the situation we are in. We could assume, that we want to display the worst-case scenarios to the consumer as a counter-weight to the commercials. In this case, we could use the average of the worst-case scenarios or even the worst score of the worst-case scenarios if we really wanted to hamper enthusiasm for the product. Another circumstance would be if we were to pick a random MVA and try to give an approximation of its success rate. In this case our safest bet would be to use the average value of all scores. If we were to talk about MVAs as a field of research in a more grand scale however, we should use the score of the best MVA to represent the current progress in research, because it is always the best product, that sets the scale for others to follow. The last approach would come in handy if we wanted to represent the technology inside MVAs and their advancement. To show this best, we would need to combine the best scores of each individual category to create the imaginary ideal MVA that contains all the best technology on the market and use its score as a representation.

So as a whole we now have five different approaches to answering the question. We have the worst of the worst-cases, which would be Evi in the phone control category, the average worst case, the average over all categories, the best MVA, which would be Siri, and the ideal MVA, which would be Siri, but with Google Now's Voice Recognition technology. These are the scores representing the individual approaches:

1. worst worst-case: 0.00%
2. average worst-case: 24.93%
3. average: 44.5%
4. best: 65.8%
5. ideal: 68.3%

Since the institute for text technology is focused on research and not consumer counseling, we thought of the 'best' and 'ideal' approaches as more appropriate for this thesis.

10 Statistical Significance

After we have established a final result for this experiment, we may now investigate whether these results coincide with our expectations (4) or not. We need to perform a test for significance on our result, which means that we have to calculate the probability of our result occurring if the initial hypothesis was correct. If this probability is below 5% our results show a sufficient deviation from our expectations to invalidate the latter with only a 5% chance for error on our part. If the probability is above 5%, our initial hypothesis is reinforced by the test results.
We tried two different approaches of modeling our test with a statistical method. In the first approach we decided to treat each point of the MRP as a coin flip, which can either be a success or a failure, meaning that the point can either be received or not. In this case we are dealing with a Binomial Distribution [25], given by the formula:

$$P(X=k) = \binom{n}{k} * p^k * (1-p)^{n-k}$$

whereby n is the total amount of experiments, k is the number of successes and p is the probability of a success. All of these parameters are given, since the amount of experiments is the total amount of MRP (648), the number of successes is our final result extrapolated over the total MRP and the probability for a success is given in our initial hypothesis (4).
The final results for the best and ideal MVAs returned the following value for k:

best: $k_{best} = 648 \cdot 0.658 = 426$

ideal: $k_{ideal} = 648 \cdot 0.683 = 442$

With all of these values set into the formula, we get the following results:

best: $P(X = 426) = 1.348 \cdot 10^{-21}$

ideal: $P(X = 442) = 3.008 \cdot 10^{-22}$

The other approach we tried was to model every test-case as the throw of a die, with the individual test categories corresponding to different dice, since the rating system allows for differing amounts of points per question. Therefore we would use a three-sided die for the tests in the Voice Recognition and Phone Control categories, a four-sided die for the Online Search category and a two-sided die for the Natural Conversation category. Probabilities for the sum of a series of die rolls is a very complex concept to calculate statistically, but since the sums for die-rolls converge on the Normal Distribution[26], we can approximate a result via their Expected Value and Variance.
The Expected Value of dice-roll-sums usually equals $E(X) = \frac{a+b}{2}$ whereby a is the smallest and b the largest possible sum. But we are not using fair dice in this calculation, since we started with an initial expectation of 85% of the MRP. Using this percentage instead, we get an Expected Value of 367 points for Voice Recognition, 112 for Online Search, 71 for Phone Control and 36 for Natural Conversation.
The Variance can be calculated by first determining the Variance of a single die and then multiplying with the amount of dice thrown, since multiple dice are statistically independent. The Variance to a single die equals[27]

$$V(X) = \sum_{i=0}^{n-1} \frac{(i - E(X))^2}{n}$$

whereby n is the amount of sides on the die and $E(X)$ the Expected Value for the die (In this case 85% of the largest result). This formula delivers 1.16 as Variance of a single die in the Voice Recognition and Phone Control categories, 2.35 in Online Search and 0.37 in Natural Conversation. Multiplied with the amount of dice thrown, we get 250.56 as Variance in Voice Recognition, 98.7 in Online Search, 48.72 in Phone Control and 15.54 in Natural Conversation.
Since the categories are statistically independent as well, we can just add up the Expected Values and Variances to get the values for the whole test.

- $E(X) = 586$
- $V(X) = 413.52$

We can then use these parameters to define a Normal Distribution and determine the probability for our results. The cumulative probability for all values up to an upper margin in a Normal Distribution is calculated by:

$$P(X <= k) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{k} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt$$

With σ being the Standard Deviation, k being the upper margin and μ the Expected Value. Since discrete values have a probability of 0 in continuous probability distributions we use the interval of $[k-0.5, k+0.5]$ instead and receive the following results:

best: $P(425.5 <= X <= 426.5) = 8.9 \cdot 10^{-4}$

ideal: $P(441.5 <= X <= 442.5) = 9.1 \cdot 10^{-4}$

Through the second approach we get less exact and a lot higher probabilities, but even these are still below 5%, which states that our test results show a significant difference to the expected percentage.

11 Conclusion

We can state on the basis of our test results that MVAs have a significantly lower success rate, when challenged with a very diverse set of tasks, than we would have expected from them and than they need to be used by the majority of people. This shows that a lot of features are still to be integrated and improved to widen their currently limited field of use.

12 Horizon

As in (9) showed, a lot of the potency of MVAs comes down to their ability to adapt and learn from various inputs. As time goes on and adoption rates rise, MVAs will become better trained and be able to produce better results without any additional development. Also there are some promising new projects in development with Soundhound's 'Hound' [15] [16] [17] and Viv Labs' 'Viv' [18] [19], which can supposedly process more complex queries and will come with third-party app integration, which gives them a much wider field of use.
These prospects seem very promising from a technological point of view and if they come without ethical quandaries and privacy concerns, improvements of such a scale might convince more people of the idea of MVAs and their ability to replicate what we only know as fiction up until now.

13 Acknowledgments

At this point we would like to thank:

Jan Friebertshäuser For beta reading.

Dr. Bernhard Mertz For providing his Windows Phone for the duration of the tests.

Tobias Schlott For helping me with the tests and providing his iPhone as test device in the process.

Markus Sütter For the constructive criticism.

References

[1] Samuel Butler. *Erwewhon.* published anonymously, 1872.

[2] anonymous. “Definition of Virutal Assistant”. In: *PC Magazine Encyclopedia* (2015). URL: http://www.pcmag.com/encyclopedia/term/61228/virtual-assistant.

[3] Margaret Rouse. “Definition: Digital Assistant”. In: *WhatIs Personal Computing Glossary* (2014). URL: http://whatis.techtarget.com/definition/digital-assistant.

[4] *Google Now alternatives for All Platforms*. English. published online. last access 2015-09-17. URL: http://alternativeto.net/software/google-now/.

[5] *Evi - App*. English. published online. last access 2015-09-17. URL: https://www.evi.com/app.

[6] *Gestatten: Cortana*. German. published online. last access 2015-09-17. URL: http://www.windowsphone.com/de-de/how-to/wp8/cortana/meet-cortana.

[7] Tom Warren. “The story of Cortana, Microsoft’s Siri killer”. English. In: *The Verge* (2014). URL: http://www.theverge.com/2014/4/2/5570866/cortana-windows-phone-8-1-digital-assistant.

[8] *Assistant by Api.ai*. English. published online. last access 2015-09-17. URL: https://assistant.ai/.

[9] *Apple - iOS 8 - Siri*. German. published online. last access 2015-09-17. URL: https://www.apple.com/de/ios/siri/.

[10] Eric Herrmann. “Google Sprachbefehle: Nützliche Kommandos für Google Now”. German. In: *AndroidPIT* (2015). URL: http://www.androidpit.de/google-now-sprachbefehle.

[11] *Alicoid Dein persönlicher Sprachassistent*. German. published online. last access 2015-09-17. URL: http://www.kkdevs.com/doku/.

[12] *Dragon Go! Speech Recognition App for iOS and Android*. English. published online. last access 2015-09-17. URL: http://www.nuancemobilelife.com/apps/dragon-go.

[13] Hilmar Schmundt. “Charmante Maschinen: Siri, lass mich nicht allein”. German. In: *Spiegel Wissen* 3 (2015). URL: http://www.spiegel.de/spiegelwissen/siri-cortana-co-wie-intelligent-sind-digitale-sprachassistenten-a-1041324.html.

[14] *SHERPA Personal Assistant*. English. published online. last access 2015-09-17. URL: http://sher.pa/.

[15] Lea Weitekamp. “Shame on you, Siri und Cortana: Warum seid ihr nicht so gut wie Hound?” German. In: *t3n* (2015). URL: http://t3n.de/news/fantastisch-siri-cortana-google-now-hound-zukunft-614646/.

[16] *Hound Internal Demo*. English. published on YouTube.com. last access 2015-09-17. URL: https://www.youtube.com/watch?v=M1ONXeaOmXg.

[17] *Hound beta vs. Google Now*. English. published on YouTube.com. last access 2015-09-17. URL: https://www.youtube.com/watch?v=2uMzhWumLFs&feature=youtu.be.

[18] John H. Richardson. “Viv Will Replace Your Smartphone With Your Fridge And Then Take Over The World”. English. In: *Esquire* (2015). URL: http://www.esquire.com/lifestyle/a34630/viv-artificial-intelligence-0515/.

[19] Sarah Perez. “Viv, Built By Siri’s Creators, Scores $12.5 Million For An AI Technology That Can Teach Itself”. English. In: *TechCrunch* (2015). URL: http://techcrunch.com/2015/02/20/viv-built-by-siris-creators-scores-12-5-million-for-an-ai-technology-that-can-teach-itself/.

[20] *Frozen Synapse: Original Soundtrack*. published on Bandcamp.com. last access 2015-09-17. URL: http://nervoustestpilot.co.uk/album/frozen-synapse-original-soundtrack.

[21] *Frozen Synapse - OST MiX*. published on YouTube.com. last access 2015-09-17. URL: https://www.youtube.com/watch?v=PAln2rGiOF4.

[22] *train sound effect*. published on YouTube.com. last access 2015-09-17. URL: https://youtu.be/FTBC57laBy8.

[23] *RELAX - Crowd Sound Calm*. published on YouTube.com. last access 2015-09-17. URL: https://www.youtube.com/watch?v=9DzewZDtz7Q.

[24] Douglas Adams. *The Hitchhiker’s Guide to the Galaxy*. Pan Books, 1979.

[25] *The Binomial Distribution*. English. published online. last access 2015-09-17. URL: http://www.stat.yale.edu/Courses/1997-98/101/binom.htm.

[26] *Normal Distribution*. English. published online. last access 2015-09-24. URL: http://mathworld.wolfram.com/NormalDistribution.html.

[27] *Probability in Games 04: Variance in Dice Sums*. English. published online. last access 2015-09-27. URL: http://sugarpillstudios.com/wp/?page_id=1004.

[28] *Definition of Virtual*. English. published online. last access 2015-09-17. URL: http://www.oxforddictionaries.com/definition/english/assistant.

[29] *Definition of Digital*. English. published online. last access 2015-09-17. URL: http://www.oxforddictionaries.com/definition/english/digital.

[30] *Definition of Intelligent*. English. published online. last access 2015-09-17. URL: http://www.oxforddictionaries.com/definition/english/intelligent.

[31] *Definition of Mobile*. English. published online. last access 2015-09-17. URL: http://www.oxforddictionaries.com/definition/english/mobile.

[32] *Definition of Personal*. English. published online. last access 2015-09-17. URL: http://www.oxforddictionaries.com/definition/english/personal.

[33] *Definition of Virtual*. English. published online. last access 2015-09-17. URL: http://www.oxforddictionaries.com/definition/english/virtual.

14 Appendix

14.1 List of Acronyms and Abbreviations

AI Artificial Intelligence (First use: (2))

MVA Mobile Virtual Assistant (First use: (3))

MRP Maximum amount of Reachable Points (First use: (7))

14.2 Component Definitions

Assistant:
A person who helps in particular work.[28]

Digital:
(General): (Of signals or data) expressed as series of the digits 0 and 1, typically represented by values of a physical quantity such as voltage or magnetic polarization. Often contrasted with analogue.
(Alternative): Involving or relating to the use of computer technology.[29]

Intelligent:
(General): Having or showing intelligence, especially of a high level.
(Of a device or building): able to vary its state or action in response to varying situations and past experience.[30]

Mobile:
Able to move or be moved freely or easily.[31]

Personal:
Belonging to or affecting a particular person rather than anyone else.[32]

Virtual:
(General): Almost or nearly as described, but not completely or according to strict definition.
(Computing): Not physically existing as such but made by software to appear to do so.[33]

14.3 Voice Recognition Test-Queries

14.3.1 German

1. Simple

   1.1 Wie wird das Wetter morgen?
   1.2 Ich habe Hunger.
   1.3 Wie geht es dir?

2. Colloquial / Vulgar

   2.1 Die Pommes schmecken scheiße.
   2.2 Ich hab' keinen Bock mehr auf den Mist.
   2.3 Was is'n das für'n beschissenes Wetter?

3. Scientific

   3.1 Was ist das Raue Endoplasmatische Retikulum?
   3.2 Wie funktioniert Outlier Detection unter Verwendung von Gaussian Mixture Models?
   3.3 Was ist ein Malignes Melanom?

14.3.2 English

1. Simple

   1.1 How is the weather going to be tomorrow?
   1.2 I am hungry.
   1.3 How are you?

2. Colloquial / Vulgar

   2.1 The fries taste like shit.
   2.2 I just don't give a fuck anymore.
   2.3 What about this shitty weather?

3. Scientific

   3.1 What is the rough endoplasmic reticulum?
   3.2 How does Outlier Detection using Gaussian Mixture Models work?
   3.3 What is a malignant melanoma?

14.4 Online Search Test-Queries

14.4.1 German

1. Answering a simple question
   1.1 Wie wird das Wetter morgen?
   1.2 Wie viel Uhr ist es?
   1.3 Welcher Wochentag ist morgen?
2. Calculating mathematical tasks
   2.1 Was ist Drei mal Sieben?
   2.2 Was sind Drei Prozent von 27?
   2.3 Was ist die Wurzel von 45 plus Vier?
3. Answering a question with a complex answer
   3.1 Was ist der Sinn des Lebens?
   3.2 Wer war Leonardo DaVinci?
   3.3 Was ist ein Virus?
4. Processing questions with incomplete information
   4.1 Was ist der Unterschied zwischen einer Ente?
   4.2 Wie hoch sind meine Zinsen auf 500 Euro?
   4.3 Was ist das durchschnittliche Gehalt?
5. Answering questions relating to the last output
   5.1 (Wie wird das Wetter morgen?) Und übermorgen?
   5.2 (Was ist Drei mal Sieben?) Und das mal Vier?
   5.3 (Was sind Drei Prozent von 27?) Und von 30?
6. Navigating
   6.1 Zeig mir den Weg nach Frankfurt.
   6.2 Wie komme ich am schnellsten nach Frankfurt?
   6.3 Navigiere mich nach Frankfurt.
7. Searching for images
   7.1 Zeig mir Bilder von Pinguinen.
   7.2 Zeig mir Fotos mit Autos.
   7.3 Zeig mir Bilder von Leuten mit Hüten.

14.4.2 English

1. Anwering a simple question
   1.1 How is the weather tomorrow going to be?
   1.2 What is the time?
   1.3 What day of the week is tomorrow?
2. Calculating mathematical tasks
   2.1 What is seven times three?
   2.2 What are three percent of 27?
   2.3 What is the square root of 45 plus four?
3. Answering a question with a complex answer
   3.1 What is the meaning of life?
   3.2 Who was Leonardo DaVinci?
   3.3 What is a virus?

4. Processing questions with incomplete information

    4.1 What is the difference between a duck?
    4.2 How high is my interest on 500 Euros?
    4.3 What is the average salary?

5. Answering questions relating to the last output

    5.1 (How is the weather tomorrow going to be?) And the day after?
    5.2 (What is seven times three?) And that times four?
    5.3 (What are three percent of 27?) And of 30?

6. Navigating

    6.1 Show me the way to Frankfurt.
    6.2 How will I get to Frankfurt the fastest?
    6.3 Navigate me to Frankfurt.

7. Searching for images

    7.1 Show me pictures of penguins.
    7.2 Show me photographs with cars.
    7.3 Show me images of people with hats.

14.5 Phone Control Test-Queries

14.5.1 German

1. Creating Appointments

    1.1 Neuer Termin.
    1.2 Erstelle einen Termin morgen um Drei Uhr Nachmittags.
    1.3 Ich habe einen Termin morgen um Drei Uhr Nachmittags.

2. Creating reminders

    2.1 Erinnere mich in Zwei Minuten, weiter zu testen.
    2.2 Neue Erinnerung. Ich will in Zwei Minuten weiter testen.
    2.3 Kannst du mich in Zwei Minuten daran erinnern weiter zu testen?

3. Repeating past reminders

    3.1 Wiederhole die Erinnerung bitte.
    3.2 Wiederhole diese Erinnerung in zehn Minuten.
    3.3 Erinnere mich morgen zur gleichen Zeit nochmal daran.

4. Setting up alarms

    4.1 Wecke mich in Fünf Minuten.
    4.2 Stelle einen Wecker auf in Fünf Minuten.
    4.3 Ich will um Acht Uhr geweckt werden.

5. Calling contacts

    5.1 Rufe "Test" an.
    5.2 Rufe meinen Kontakt "Test" an.
    5.3 Rufe in Zwei Minuten "Test" an.

6. Creating notes

    6.1 Erstelle eine Notiz.
    6.2 Notiere folgendes.
    6.3 Neue Notiz.

7. Managing versions and installing updates

    7.1 Gibt es Updates, die ich installieren sollte?
    7.2 Welche Version bist du?
    7.3 Gibt es eine neuere Version von dir?

14.5.2 English

1. Creating Appointments
   - 1.1 New appointment.
   - 1.2 Create a new appointment tomorrow at three p.m.
   - 1.3 I have an appointment tomorrow at three p.m.
2. Creating reminders
   - 2.1 Remind me in two minutes to proceed testing.
   - 2.2 New reminder. I want to proceed testing in two minutes.
   - 2.3 Can you remind me in two minutes to proceed testing?
3. Repeating past reminders
   - 3.1 Please repeat that reminder.
   - 3.2 Please remind me again in ten minutes.
   - 3.3 Remind me again tomorrow at the same time.
4. Setting up alarms
   - 4.1 Wake me in five minutes.
   - 4.2 Set up an alarm for five minutes from now.
   - 4.3 I want to be woken up at eight a.m.
5. Calling contacts
   - 5.1 Call ”Test”.
   - 5.2 Call my contact ”Test”.
   - 5.3 Call ”Test” in two minutes.
6. Creating notes
   - 6.1 Create a note.
   - 6.2 Note the following.
   - 6.3 New notice.
7. Managing versions and installing updates
   - 7.1 Are there any updates I should install?
   - 7.2 What version are you?
   - 7.3 Is there a new version of you?

14.6 Natural Conversation Test-Queries

14.6.1 German

1. Telling a joke
   - 1.1 Erzähle mir einen Witz.
   - 1.2 Erzähle mir einen anderen Witz.
   - 1.3 Ich war in einer Bar etwas trinken und habe dann den Bus nach Hause genommen. Klang erst nach einer guten Idee, aber ich habe vorher noch nie einen Bus gefahren.
2. Creating aliases
   - 2.1 Ich will einen Alias für ”Test” erstellen.
   - 2.2 Ich will ”Test” auch ”Training” nennen können.
   - 2.3 Wenn ich ”Training” sage, meine ich ”Test”.
3. Monitoring current position
   - 3.1 Wo bin ich gerade?
   - 3.2 Wann war ich das letzte mal in Frankfurt?
   - 3.3 Was ist die nächste Stadt?

4. Saving names
    - 4.1 Was ist mein Name?
    - 4.2 Wie nennst du mich?
    - 4.3 Wie heißt du?
5. Changing names
    - 5.1 Nenne mich "Versuch".
    - 5.2 Ich will dich "Versuch" nennen.
    - 5.3 Ab sofort heiße ich "Versuch".
6. Talking about mood
    - 6.1 Wie geht's dir?
    - 6.2 Alles klar?
    - 6.3 Wie fühlst du dich?
7. Having 'thoughts'
    - 7.1 Was denkst du gerade?
    - 7.2 Woran denkst du?
    - 7.3 Was geht in deinem Kopf vor?

14.6.2 English

1. Telling a joke
    - 1.1 Tell me a joke.
    - 1.2 Tell me another joke.
    - 1.3 I was drinking at a bar, so I took the bus home. Seemed like a good idea at the time, but I have never driven a bus before.
2. Creating aliases
    - 2.1 I want to create an alias for "Test".
    - 2.2 I want to call "Test" "Training" as well.
    - 2.3 If I say "Training", I mean "Test".
3. Monitoring current position
    - 3.1 Where am I?
    - 3.2 When was the last time I was in Frankfurt?
    - 3.3 What is the nearest city?
4. Saving names
    - 4.1 What is my name?
    - 4.2 What do you call me?
    - 4.3 What is your name?
5. Changing names
    - 5.1 Call me "Versuch".
    - 5.2 I want to call you "Versuch" from now on.
    - 5.3 From this point on, my name is "Versuch".
6. Talking about mood
    - 6.1 How are you?
    - 6.2 What's up?
    - 6.3 How are you doing?
7. Having 'thoughts'
    - 7.1 What are you thinking about?
    - 7.2 What's on your mind right now?
    - 7.3 What's your brain busy with?

14.7 Results

German

		1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3
Evi										
silence	slow	0	0	0	0	0	0	0	0	0
	medium	0	0	0	0	0	0	0	0	2
	fast	0	0	0	0	0	0	0	0	2
music	slow	0	0	0	0	0	0	0	0	0
	medium	0	0	0	0	0	0	0	0	0
	fast	0	0	0	0	0	0	0	0	0
train	slow	0	0	0	0	0	0	0	0	0
	medium	0	0	0	0	0	0	0	0	0
	fast	0	0	0	0	0	0	0	0	0
crowd	slow	0	0	0	0	0	0	0	0	0
	medium	0	0	0	0	0	0	0	0	0
	fast	0	0	0	0	0	0	0	0	0
Cortana										
silence	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	0	2	2	1	0	2
music	slow	2	2	2	2	2	0	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	1	1	2	0	0	0	0
train	slow	0	2	0	2	2	2	1	0	2
	medium	2	0	0	0	2	2	0	0	2
	fast	2	0	2	0	2	0	0	0	0
crowd	slow	2	2	2	2	2	2	0	0	0
	medium	2	2	2	2	2	2	0	0	2
	fast	2	2	2	0	1	0	0	0	0
Speaktoit Assistant										
silence	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	2	2	2	0	0	2
music	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	0	2	1	0	0	0
train	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	0	2	2	2	0	1	0	2
crowd	slow	2	2	2	2	2	0	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	0	2	2	0	0	2

Table 6: Results in the Voice Recognition category in German.

German

		1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3
Siri										
silence	slow	2	2	2	2	2	2	1	0	2
	medium	2	2	2	2	2	2	0	1	2
	fast	2	2	2	2	2	1	0	0	1
music	slow	2	2	2	2	2	2	1	0	2
	medium	2	2	2	2	2	2	0	0	2
	fast	2	2	2	2	0	2	0	0	1
train	slow	2	2	2	2	2	2	1	0	2
	medium	2	2	2	2	2	2	1	0	2
	fast	2	2	2	0	2	0	0	0	2
crowd	slow	2	2	2	2	0	2	1	0	2
	medium	2	2	2	2	2	2	1	0	2
	fast	2	2	2	2	2	0	0	0	2
Google Now										
silence	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	2	2	2	2	0	2
music	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	0	0	2	0	0	2
train	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	0	2	0	2	2	0	2
crowd	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	0	2	0	0	1	1	0	2
Alicoid										
silence	slow	2	2	2	2	1	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	0	2	0	2	0	2
music	slow	2	2	2	2	2	2	2	0	1
	medium	2	2	2	1	2	2	2	0	2
	fast	2	2	2	0	0	0	0	0	2
train	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	0	2	0	2	2	1	0	0	2
crowd	slow	2	2	2	2	2	2	2	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	0	2	0	2	0	0	0	2

Table 7: Results in the Voice Recognition category in German.

English										
		1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3
Evi										
silence	slow	2	0	2	0	2	2	0	1	2
	medium	2	2	2	1	2	1	0	0	2
	fast	2	2	2	1	2	1	1	0	2
music	slow	2	1	2	0	0	1	0	0	2
	medium	2	2	2	1	2	2	1	0	1
	fast	1	2	2	0	2	1	0	0	2
train	slow	2	1	2	0	2	1	0	0	2
	medium	2	2	2	0	2	1	0	0	2
	fast	2	2	2	0	1	1	0	0	2
crowd	slow	1	2	2	1	2	2	0	0	2
	medium	2	2	2	0	2	0	0	0	1
	fast	2	2	2	0	2	0	0	0	2
Cortana										
silence	slow	2	2	2	2	2	2	2	1	2
	medium	2	2	2	2	2	1	2	0	1
	fast	2	2	2	0	2	1	0	0	0
music	slow	2	2	2	2	2	0	0	1	0
	medium	2	2	2	2	2	2	2	0	2
	fast	0	0	2	0	2	0	2	0	0
train	slow	2	2	2	0	2	2	2	0	2
	medium	2	2	2	2	2	2	0	0	2
	fast	2	2	2	0	2	1	0	0	1
crowd	slow	2	0	2	2	2	2	0	0	2
	medium	2	2	0	2	2	2	2	1	2
	fast	2	2	0	0	0	1	0	0	1
Speaktoit Assistant										
silence	slow	2	2	2	2	2	2	2	1	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	2	2	2	2	0	2
music	slow	2	2	2	1	2	2	2	1	2
	medium	2	2	2	1	2	2	2	0	2
	fast	2	2	0	1	2	2	2	0	2
train	slow	2	2	2	2	2	2	2	1	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	0	2	0	2	0	2
crowd	slow	2	2	2	1	2	2	2	1	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	0	2	2	2	0	2

Table 8: Results in the Voice Recognition category in English.

English

		1.1	1.2	1.3	2.1	2.2	2.3	3.1	3.2	3.3
Siri										
silence	slow	2	2	2	2	2	2	1	1	1
	medium	2	2	2	1	2	1	2	0	2
	fast	2	2	2	2	2	2	2	0	2
music	slow	2	2	2	2	2	2	2	0	1
	medium	2	2	2	2	2	2	2	0	2
	fast	1	2	2	0	2	1	2	0	2
train	slow	2	2	2	2	1	2	2	1	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	2	2	1	2	0	2
crowd	slow	2	2	2	1	0	2	1	0	2
	medium	2	2	2	2	2	2	2	0	2
	fast	1	2	2	0	2	1	2	0	2
Google Now										
silence	slow	2	2	2	2	2	2	2	1	2
	medium	2	2	2	2	2	2	2	0	2
	fast	2	2	2	2	2	2	2	2	2
music	slow	2	2	2	2	2	2	2	2	2
	medium	2	2	2	2	2	2	2	2	2
	fast	2	2	2	0	2	2	2	0	2
train	slow	2	2	2	2	2	2	2	2	2
	medium	2	2	2	2	2	2	2	2	2
	fast	2	2	2	2	2	2	2	1	2
crowd	slow	2	2	2	2	2	2	2	2	2
	medium	2	2	2	2	2	2	2	2	2
	fast	2	2	2	2	2	1	2	2	2
Alicoid										
silence	slow	2	2	2	0	2	1	0	0	0
	medium	2	2	2	0	0	1	0	0	0
	fast	0	1	2	0	0	0	0	0	0
music	slow	2	2	2	0	2	1	1	0	0
	medium	2	2	2	0	1	1	2	0	0
	fast	0	0	0	0	0	0	0	0	0
train	slow	2	2	2	0	1	1	2	0	0
	medium	2	2	2	0	0	1	1	0	0
	fast	0	0	0	0	0	0	0	0	0
crowd	slow	2	2	2	0	1	0	2	0	1
	medium	2	0	2	0	0	0	0	0	0
	fast	0	0	2	0	0	0	0	0	0

Table 9: Results in the Voice Recognition category in English.

German

	Evi	Cortana	Speaktoit Assistant	Siri	Google Now	Alicoid
1.1	0	2	2	2	2	3
1.2	0	1	2	3	2	2
1.3	0	1	3	3	1	2
2.1	0	2	2	2	1	2
2.2	0	1	2	2	2	0
2.3	0	1	1	1	1	2
3.1	0	2	0	2	2	2
3.2	0	1	2	2	2	2
3.3	0	1	0	2	2	2
4.1	0	1	1	1	1	0
4.2	0	1	0	1	1	0
4.3	0	1	1	1	1	0
5.1	0	2	2	2	2	0
5.2	0	0	0	2	1	0
5.3	0	0	0	0	2	0
6.1	0	2	1	2	3	2
6.2	0	2	0	2	3	2
6.3	0	2	2	2	3	2
7.1	0	1	2	2	2	0
7.2	0	1	1	1	2	0
7.3	0	1	1	2	2	0

Table 10: Results of the Online Search category in German.

English

	Evi	Cortana	Speaktoit Assistant	Siri	Google Now	Alicoid
1.1	1	2	2	2	2	0
1.2	2	2	2	3	2	0
1.3	3	1	3	3	3	0
2.1	2	2	2	2	2	0
2.2	1	1	2	2	2	0
2.3	1	1	2	2	2	0
3.1	2	2	2	2	2	0
3.2	2	2	3	2	1	0
3.3	2	1	2	2	2	0
4.1	2	1	0	2	1	0
4.2	1	1	0	1	1	0
4.3	1	1	0	2	2	0
5.1	0	0	0	2	0	0
5.2	0	0	0	2	2	0
5.3	0	0	0	0	0	0
6.1	1	2	1	2	3	0
6.2	1	2	0	1	3	0
6.3	1	2	2	2	3	0
7.1	2	1	2	2	2	0
7.2	2	1	1	2	2	0
7.3	1	1	2	2	2	0

Table 11: Results of the Online Search category in English.

German

	Evi	Cortana	Speaktoit Assistant	Siri	Google Now	Alicoid
1.1	0	2	2	2	2	0
1.2	0	2	2	2	2	1
1.3	0	2	2	2	0	0
2.1	0	2	2	2	2	0
2.2	0	2	2	2	2	0
2.3	0	2	2	2	2	0
3.1	0	0	0	0	0	0
3.2	0	0	0	0	0	0
3.3	0	0	0	0	0	0
4.1	0	2	2	2	2	1
4.2	0	2	2	2	2	0
4.3	0	0	0	2	0	0
5.1	0	2	2	2	2	1
5.2	0	2	2	2	2	0
5.3	0	0	0	0	0	0
6.1	0	2	2	2	2	0
6.2	0	2	2	2	2	0
6.3	0	2	2	2	2	0
7.1	0	0	0	0	0	0
7.2	0	0	0	1	0	0
7.3	0	0	0	0	0	0

Table 12: Results of the Phone Control category in German.

English

	Evi	Cortana	Speaktoit Assistant	Siri	Google Now	Alicoid
1.1	0	2	2	2	2	0
1.2	0	2	2	2	2	0
1.3	0	2	2	2	2	0
2.1	0	2	2	2	2	0
2.2	0	2	2	2	2	0
2.3	0	2	2	2	2	0
3.1	0	0	0	0	0	0
3.2	0	0	0	0	0	0
3.3	0	0	0	0	0	0
4.1	0	2	2	2	2	0
4.2	0	2	2	2	2	0
4.3	0	2	0	0	0	0
5.1	0	2	2	2	0	0
5.2	0	2	0	2	0	0
5.3	0	0	0	0	0	0
6.1	0	2	2	2	2	0
6.2	0	2	2	2	2	0
6.3	0	0	0	2	0	0
7.1	0	0	0	0	0	0
7.2	0	0	0	1	0	0
7.3	0	0	0	0	0	0

Table 13: Results of the Phone Control category in English.

German

	Evi	Cortana	Speaktoit Assistant	Siri	Google Now	Alicoid
1.1	0	1	0	1	0	1
1.2	0	1	0	1	0	1
1.3	0	0	0	0	0	0
2.1	0	0	0	0	0	0
2.2	0	0	0	0	0	0
2.3	0	0	0	0	0	0
3.1	0	1	1	1	1	1
3.2	0	0	0	0	0	0
3.3	0	0	0	0	0	0
4.1	0	1	1	1	0	1
4.2	0	1	0	1	0	0
4.3	0	1	1	1	0	1
5.1	0	0	1	1	0	0
5.2	0	0	0	0	0	0
5.3	0	0	0	0	0	0
6.1	0	1	1	1	0	1
6.2	0	1	0	1	0	0
6.3	0	1	0	1	0	1
7.1	0	0	0	1	0	0
7.2	0	0	0	1	0	0
7.3	0	0	0	0	0	0

Table 14: Results of the Natural Conversation category in German.

English

	Evi	Cortana	Speaktoit Assistant	Siri	Google Now	Alicoid
1.1	1	1	1	1	0	0
1.2	1	0	1	0	0	0
1.3	0	0	0	0	0	0
2.1	0	0	1	0	0	0
2.2	0	0	0	0	0	0
2.3	0	0	0	0	0	0
3.1	1	1	1	1	1	0
3.2	0	0	0	0	0	0
3.3	0	0	0	1	0	0
4.1	0	1	1	1	0	0
4.2	0	1	0	1	0	0
4.3	1	1	1	1	0	0
5.1	1	0	1	1	0	0
5.2	0	0	1	0	0	0
5.3	0	0	0	0	0	0
6.1	1	1	1	1	0	0
6.2	1	1	1	1	0	0
6.3	1	1	1	1	0	0
7.1	1	0	0	1	0	0
7.2	0	0	0	0	0	0
7.3	0	0	0	1	0	0

Table 15: Results of the Natural Conversation category in English.