Exploring the Differences Between Touch and Voice Input on Smartphones

Usability Engineering & Testing - Winter semester 2017/18

Adrian Teschendorf
Matriculation number:
60258
adrian.teschendorf@uni-weimar.de

Lisa Guth
Matriculation number:
100936
lisa.guth@uni-weimar.de

ABSTRACT

Writing text messages has mostly surpassed the classic phone call in popularity these days. With the aid of modern voice recognition methods, speech-to-text features are a possible alternative to using the standard touch-keyboard on smartphones. This paper describes an experiment comparing the performance of touch and voice input for text messaging. While seemingly being the faster input method, we wanted to figure out if voice input is also the more efficient way to prevent spelling errors.

KEYWORDS

Automatic voice recognition, Deep Learning, Likert scale

1 INTRODUCTION

In the recent years, mobile devices became more and more a solid part of people's everyday life. Sending text messages to family and friends is one of the main uses of current smartphones, which has even surpassed common face-to-face conversions in Great Britain [1].
To ease to task of writing sentences for the users, various technologies have been implemented in the past, like T9, auto-correction and word suggestions. Since typing a long text message on a smartphone can still be tiresome or even a nuisance, especially on smaller phones due to fat-finger errors, using voice as an input method appears to be a promising alternative.
Although beginning in the 1930s, where Bell Laboratories developed the first "system model for speech analysis and synthesis", it wasn't until the 1980s before noticeable advances were made in the field later called Automatic Speech Recognition [2]. While touchscreen keyboards only require finger contact with the screen to produce a letter, a user's voice has to be recorded, converted and further processed the be recognized as words [3].
With the advance of Deep Learning algorithms, Automatic Speech Recognition systems like Apple's Siri or Google Speech Recognition started to become more efficient in recognizing a user's voice by using cloud-based word-databases and learning algorithms that can also understand the various dialects of a language [4].
Because of these promising advantages we wanted to conduct a test study to find out if these voice recognition systems are more efficient for writing text messages than the standard touchscreen keyboard used in current smart-phones.

2 EMPIRICAL RESEARCH

Touch and voice input can be used for many similar tasks on a smartphone, for example, submitting search queries, taking notes or writing text messages, yet both use completely different ways to produce a sentence. In contrast to touch inputs, voice recognition systems can only identify whole words and not single letters. Furthermore, a user can correct his spelling immediately on a touch-keyboard using the backspace key, while voice input requires that the user repeats the entire sentence again in case of a mistake. Therefore, we had to adjust our research study so that both input methods can be compared to each other. For example, the auto-correction function of the touch-keyboard was disabled and the manually correction of misspelled words was prohibited. Also, we limited our test study to users that are already familiar with both input methods to concentrate only on the differences of these techniques and not the users' experience. With these restrictions in mind, we formulated our research question as: ***Is voice input more efficient than touch input for writing text messages on smartphones?***

In this case, the efficiency describes on the one hand, how fast a given sentence is entered and processed, and on the other hand, how many words contain mistakes.
We chose to split the efficiency into these two attributes while formulating the hypotheses H_1 and H_2, because it is possible for one input method to be faster, yet contain more mistakes than the other one. Since voice recognition systems are more advanced nowadays and use cloud-based dictionaries, we favored the voice input method in the our hypotheses because it appeared to be the more efficient method in theory. Our two hypotheses and the null hypothesis are formulated as follows:

H_0: *Using voice input is equal efficient than using touch input for writing text messages on a smartphone.*

H_1: *Voice input takes less time to enter a text message than touch input.*

H_2: *Voice input leads to less misspellings than touch input.*

After that, we determined suitable variables for our test study. As an independent variable we chose to concentrate only on the *type of the input method*. In our case, we varied between voice and touch input. The dependent variables we wanted to measure are on the one hand the *time for entering and processing* a given sentence, and on the other hand the *number of misspelled words*. Additionally, we asked the participants of the study for their *preferred method* by letting them fill out a questionnaire afterwards.

While conducting the test study, there were various confounding factors that needed to be considered. Since the quality of voice recognition is influenced negatively by background noises, we chose to conduct the test study in a controlled environment that suppressed most of the possible noise. Furthermore, the participant of the test study had their own way of speaking. While some of them spoke faster or quieter, other were slower or louder. Also, all users were non-native English speakers, but were on an advanced level in this language. However, we didn't want to eliminate this confounding factor on purpose, because we wanted to identify if the voice recognition system could handle different ways of speaking. It was important for us that the participants spoke with their regular tone without appearing too fake, because the experiment should simulate a situation in the everyday life.
To keep consistency throughout the whole study, we used the same mobile device (iPhone 5) with the standard english keyboard layout for every participant. Because some participants might perform different in an observed environment, which is described by the *Hawthorne effect*, we decided to leave the participant alone in the room and observed them anonymously by only monitoring the mobile device via camera. To minimize a possible *learning effect*, we let the users get comfortable with the environment and the device first for a few minutes. Also, the first task in the test study did not count towards the collected data, it was merely a tutorial for the participants to get used to the study.
To sum up, the research included the following traits:

Independent variable:
- Input method (voice / touch)

Dependent variables:
- Time to write and process a given sentence
- Number of misspelled words
- Participants' preferred input method

Confounding factors:
- Environment properties (e.g. background noise)
- Participants' way of speaking
- Participants' different mobile devices and settings
- Hawthorne effect
- Learning effect

3 EXPERIMENTAL DESIGN

For our test study, we chose a *one-factorial within-subjects* experimental design because we wanted every participant to use both input methods and compare the results within one group. Additionally, we were also interested in the users' preferred input method, which could only be obtained by solving both tasks. Overall, we had 12 users (7 males and 5 females) participating in our study, whose age lied between 20 and 35 years. Two of these participants were part of a pilot study and their result data was not included in the final statistics.

3.1 Setup of the experiment

The components of the experiment consisted of an iPhone 5 using Apple's standard messaging application *iMessage* with the standard english keyboard layout for touch input. Also, Siri's dictation function was used for voice input. A camera above the mobile device captured the user's interaction with the smartphone and an additional phone, used as a stop watch, was placed in the field of view of the camera and was invisible for the test person. Figure 1 shows the default setup of the experiment.
The participants of the test study had to solve two different tasks. At first, they had to use the touch-keyboard to input one of six different sentences, which were common english

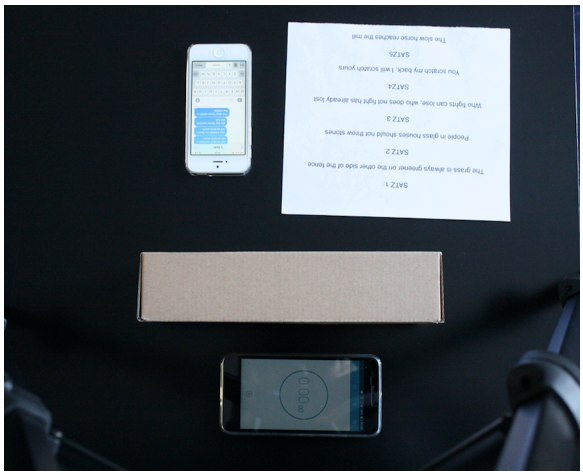


Figure 1: The experiment's default setup.

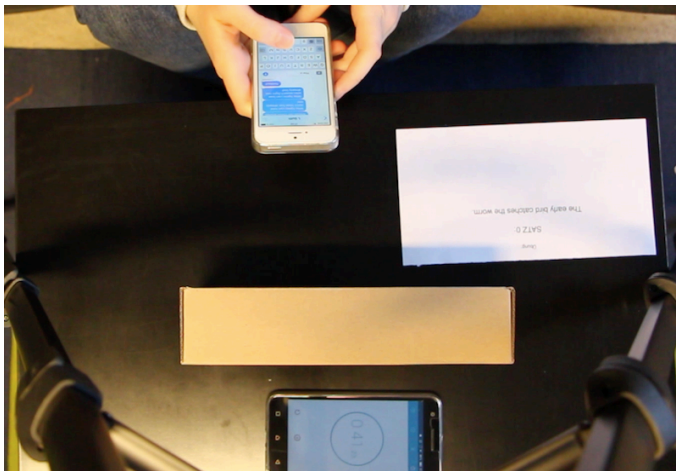


Figure 2: A capture of the experiment during the pilot study.

proverbs, like *"The grass is always greener on the other side of the fence"*. After entering the sentence via touch-keyboard, the participants had to use voice input, by activating Siri's dictation function, and speak the same sentence aloud into the microphone of the mobile device. The first sentence was not tracked and only used as an exercise to let the user get comfortable with the task and situation. These two tasks were then repeated five times with different phrases and tracked via camera. After that, the test persons were handed a questionnaire, which asked for their preferred input method. There were also additional questions using a Likert scale on how well the participants thought they solved the tasks.

3.2 Pilot study

After we set up the test study and planned the sequence of the experiment theoretically, we chose to perform a pilot study to figure out if everything works as intended and to eliminate possible flaws.

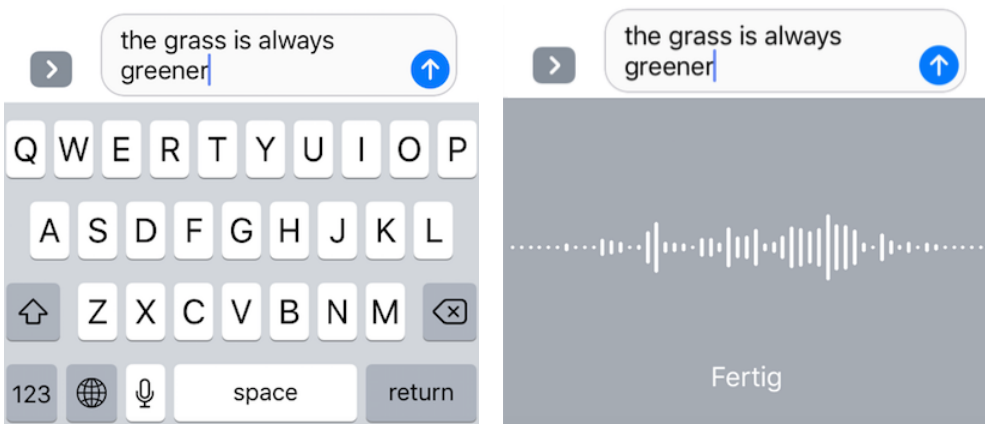
At first, we completed the tasks of the study ourselves and invited two test persons later to detect problems we might have overlooked. When we performed the tasks, there were a lot of things that needed adjustment. For example, to compare the speed of touch and voice input, the user had to write the sentence without using upper case letters or punctation. Writing these characters required switching between different keyboard modes, which takes certain time. Figure 3 shows the interface for both input methods. Also, voice recognition systems cannot differentiate between a normal sentence, an exclamation or a question. To be consistent, both input methods focus only on the words themselves ignoring the grammar of the overall sentence.

Initially, we wanted to allow the user to use auto-correction functions of the keyboard because this is the way most smartphone owners use their phone these days. However, the implemented keyboard can learn certain words over time and auto-correct into those that were typed most often. Therefore, we chose to disable to auto-correction function to keep each task of the study independent from each other without needing to reset the device every time.

The two participants of the pilot study helped us in attaining a routine to conduct the whole user study without forgetting important tasks, like handing out the questionnaire at the end of the study. A capture of the experiment setup during the pilot study is shown in Figure 2.

3.3 Conduction of the experiment

The performance of the whole test study took place within three days. One day was needed setting up the experiment and conducting the pilot study. The following two days covered the experiments of the other 10 participants, which were invited separately and each experiment took around 30 minutes. At first, we explained the test persons that their performance in the experiment will be anonymous and they could quit the study at any possible time. Each participant got an ID number that was associated with their test data. After that, the test person had 5 minutes to get comfortable with the device and the environment. Then, we explained the two tasks of the study to the participants and they could ask questions. Afterwards, we handed out a sheet of paper including the six sentences that had to be entered via touch and voice input. All participants got the same sentences, so that their performance could be compared to each other. However, the order of the phrases, except for the first, differed from each experiment to minimize a possible learning effect and not letting the users' fatigue influence the same sentence. As mentioned earlier, the first phrase was merely a tutorial to get used to the task and the participants could ask questions if they had problems with the experiment.



(a) Interface of the touch-keyboard. (b) Interface of the dictation function.

Figure 3: Interfaces of the input methods.

After that, we left the room and let the test person solve all tasks while recording their performance.

When the participant was finished, we handed them a questionnaire to fill in their age, gender, preferred input method and asked them a few questions to give us an estimation on how well they performed with the aid of Likert-type scale options. For example, we wanted to know how accurate they thought they entered the sentences giving them five possible answers ranging from not accurate to very accurate.

3.4 Measurement of the values

While the participant's preferred input method is measured with the help of a questionnaire, the other two variables were observed through evaluating the video footage. To determine the needed time to solve a task, the stop watch function of the second mobile device was used to calculate the difference between the start and the end time of each experiment. To measure the misspellings of the sentences, all the entered phrases were send to a dummy messenger account and evaluated by comparing each entered word with the correct spelling. This was done for both the touch and the voice input task. All the obtained data was then copied into an Excel spreadsheet for further analysis.

4 DATA ANALYSIS

The whole group of participants consisted of 58,3% males and 41,6% females. The majority studied at Bauhaus University Weimar while the rest were acquaintances of us. 25% of the users were between 20 and 25, 58,3% between 26 and 30 and 16,7% between 30 and 35 years old, with the mean of 27,3 years and the median of 27 years. 66,7% of the test persons were using an iPhone in their daily routine while 33,3% were Android users. All of them were experienced in using the touch-keyboard and the voice-to-text function of their phones.

During the experiment, we measured four values for each of the given sentences: the time between touching the first key and sending the text, the number of misspelled words entered via touch-keyboard, the time between pressing the key of Siri's dictation function and sending the text, and the number of misspelled words using the dictation function.

Their mean, median, variance, standard deviation and Shapiro-Wilk value are shown in Table 1. The box-plots in Figure 4 display on the one hand the comparison between the times of the touch and voice inputs, and on the other hand the differences of the misspelled words for both input methods.

	Time (Touch)	Time (Voice)	Misspelling (Touch)	Misspelling (Voice)
Mean	20.8	7.34	1.32	1.12
Median	19.5	7.4	0.7	1.1
Variance	25.66	1.01	1.42	0.45
Standard deviation	5.06	1.005	1.19	0.67
Shapiro-Wilk	0.859	0.921	0.849	0.887

Table 1: Mean, Median, Variance, Standard deviation and Shapiro-Wilk value of the dependent variables.

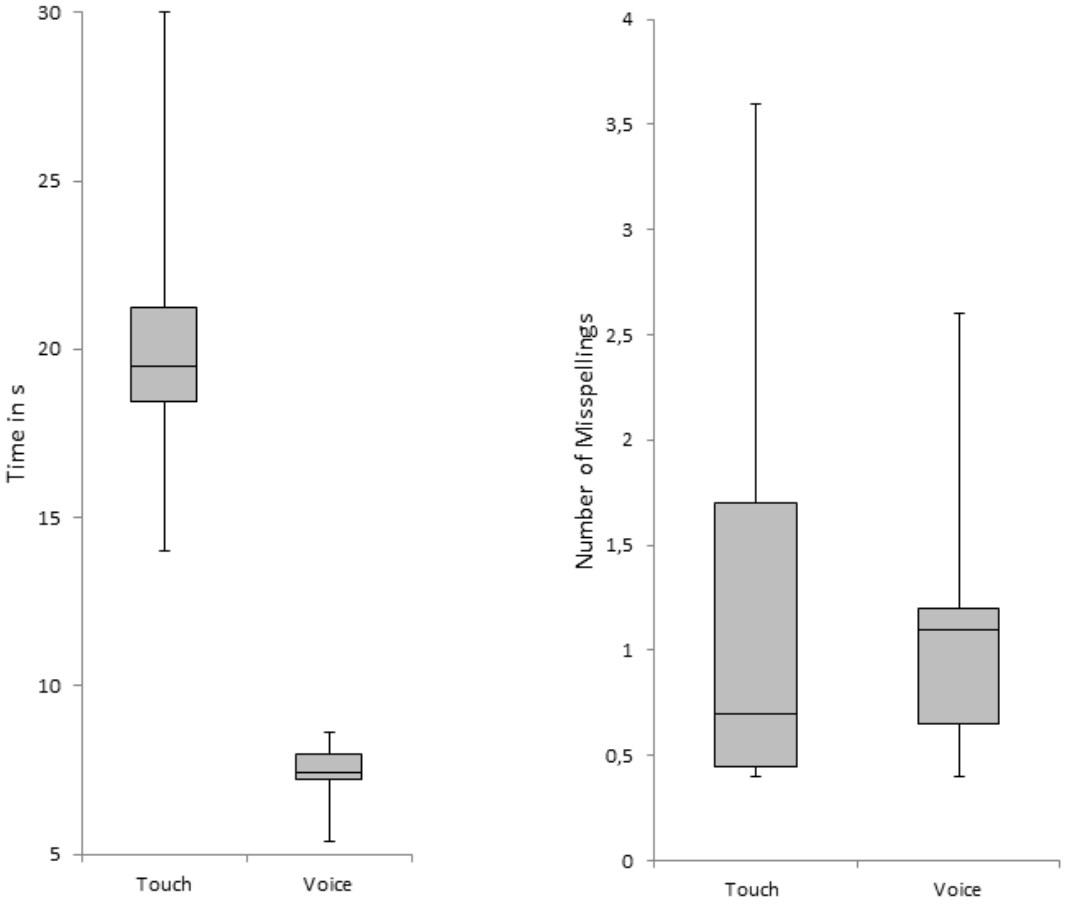


Figure 4: Box-plots comparing the time and misspellings of both input methods.

Since the Shapiro-Wilk values show that the data is normally distributed for a p-value of 0.05, we chose to conduct a two-tailed t-test. However, the standard deviations differed from each other, so we used Welch's t-test to calculate the p-values. The results are shown in Table 2.

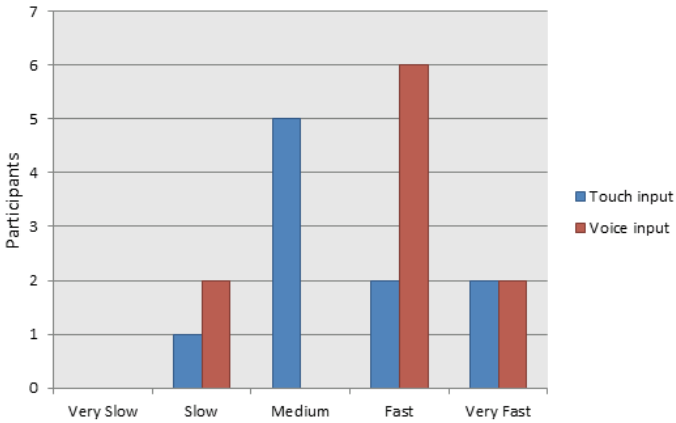
The Likert scale data of the questionnaire ist visualized in Figure 5. When asked for their preferred input method, 70% of the participants favored touch input, while 30% preferred the voice dictation function.

	Time (Touch)	Time (Voice)	Misspelling (Touch)	Misspelling (Voice)
p-value	0.000071		0.579	

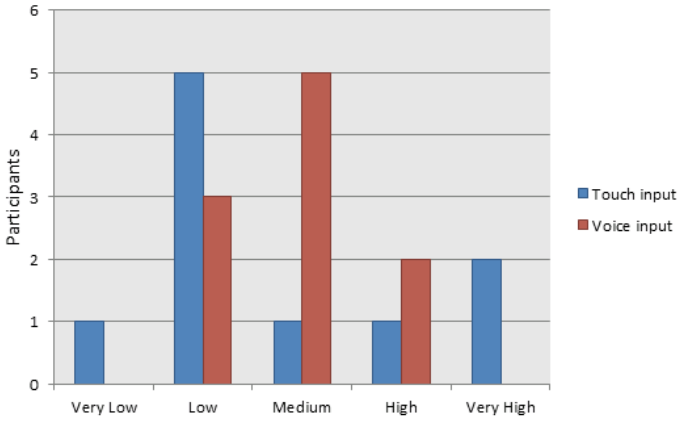
Table 2: The p-values of Welch's t-test.

5 INTERPRETATION OF THE DATA

Comparing the times of both tasks, the voice-to-text function was noticeably the faster method. Most of the participants entered the sentences in a similar amount of time with only a few outliers. These occurred mostly due to the test person checking the sentence for a few seconds before submitting it. The data of the touch input method showed a wide range of values. One person took half the time for all tasks then another participant. Since there was no restriction on how the test persons should use the keyboard, participants using two hands were generally faster than those using only one hand. These data and the p-value of Welch's t-test, which shows a high significance, support our hypothesis H_1. However, evaluating the data of the misspellings for both input methods, the results are not as clear. As shown in the box-plot in Figure 4, the test persons made less misspellings in general using the voice-to-text method that the touch-keyboard. However, the average performance showed less misspelled words using a touch-keyboard. This is particularly obvious when we examined the result for one particular sentence, which was *"The slow horse reaches the mill"*. While this seemed to be



a) Users' estimation of input speed.



b) Users' estimation of misspellings.

Figure 5: Visualization of the Likert scale data.

a fairly easy sentence to type and most participants wrote this one without mistakes on the touch-keyboard, the sentence had an average of three misspelled words using the voice-to-text method. There was not a single test person who entered this sentence without a mistake. We could only assume the explanation that automatic voice recognition systems use feature detection in the voice of the speaker and estimate the corresponding word [5]. In our case, the word "reaches" would often be recognized as "which is". Both words share similar sounds and would appear almost identical to the system when spoken into a microphone. Since "which is" is a very common used phrase, the voice recognition system might have assumed that those were the spoken words. Because of occurrences like that, we had to assume that there might be many more of these voice recognition errors in different sentences. Additionally, the p-value of Welch's t-test shows no significance and doesn't support our hypothesis H_2. Therefore, we rejected it. All in all, the null hypothesis could not be rejected. Even though voice input is the more efficient method in case of input speed, this cannot be said about the number of misspellings.

The data of the questionnaire showed that the participants' estimation of their performances often conflicted with the measured data. Even though voice input was clearly faster than touch input, some test persons perceived using the touch-keyboard as a faster or an equally fast method. Interestingly, a few participants performed both faster and with less mistakes entering the phrases via voice, yet they preferred using a touch-keyboard instead of voice input.

6 VALIDITY ANALYSIS

In our experiment we were trying to maximize the *internal validity* by measuring only the values we need, which were in our case time and the number of misspellings, and trying to control and minimize the possible confounding factors.

However, by controlling these factors we violated the *external validity* of our study. The participants had to use specific ways to input given sentences during the tasks, which often didn't correspond to the reality. Many persons make use of the auto-correction function in their daily routine or use voice input to send the complete spoken audio file without the need for a speech-to-text function. Also, people like to use slang words or abbreviations that cannot be found in a dictionary. Using the english keyboard layout was troublesome especially for german participants, since they sometimes pressed the letter "z" instead of "y" out of habit. In the real world, people are hardly alone while using their phone, so they are exposed to background noises all the time, which would reduce the efficiency of a voice recognition system. Therefore, our test study may not be generalizable.

7 CONCLUSION

All in all, the experiment confirmed some of the assumptions we had before we started the test study, but there were also unexpected results. While doing the tasks of the study ourselves, we were surprised how accurate the voice recognition worked, which motivated us to conduct

the experiment in the first place. However, we didn't expect that the majority of the test persons preferred the standard touch input, even if their results showed a better performance using the voice-to-text function. We expected voice input to be superior regarding both the time to enter a sentence and a lesser number of misspellings, yet this was not the case.

To sum up, using voice input in an appropriate environment is a faster way to write text messages than the touch-based method. However, there were no significant differences regarding the number of misspellings. Nevertheless, voice-to-text is still an input method that is worth using in the everyday life.

8 REFERENCES

[1] The Telegraph, *"Texting more popular than face-to-face conversation"*, July 18, 2012, *https://www.telegraph.co.uk/technology/9406420/Texting-more-popular-than-face-to-face-conversation.html* last visited: March 22, 2018

[2] B.H. Juang, L. R. Rabiner, *Automatic Speech Recognition – A Brief History of the Technology Development,* October 8, 2004, USA, pp.2-6

[3] D. Yu, L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, Springer-Verlag London, 2015 , pp.1-7

[4] J. Schalkwyk et al., *Google Search by Voice: A case study*, 2010, Google Inc., USA, pp.6-28

[5] T. B. Martin et al., *Speech Recognition by Feature Abstraction Techniques*, Tech. Report AL-TDR-64-176, Air Force Avionics Lab, 1964