Comparing Smartphone Speech Recognition and Touchscreen Typing for Composition and Transcription

by

Margaret Jean Foley

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2020

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis includes first-authored content from the following conference publication:

- Margaret Foley, Géry Casiez, and Daniel Vogel. 2020. **Comparing Smartphone Speech Recognition and Touchscreen Typing for Composition and Transcription**. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. DOI: https://doi.org/10.1145/3313831.3376861

The content from this paper has been adapted and extended for this thesis.

Abstract

Ruan et al. found transcribing short phrases with speech recognition nearly 200% faster than typing on a smartphone. We extend this comparison to a novel composition task, using a protocol that enables a controlled comparison with transcription. Results show that both composing and transcribing with speech is faster than typing. But, the magnitude of this difference is lower with composition, and speech has a lower error rate than keyboard during composition, but not during transcription. When transcribing, speech outperformed typing in most NASA-TLX measures, but when composing, there were no significant differences between typing and speech for any measure except physical demand.

Acknowledgements

First, I would like to thank my advisor, Daniel Vogel. I cannot overstate how immensely helpful and supportive Dan has been throughout these past few years. My thesis would not be what it is without him. He is an amazing advisor and professor, and I'm glad to have been one of his students.

Second, thank you to Géry Casiez and Edward Lank for reading my thesis and providing valuable guidance for my projects. In particular, I want to thank Géry for helping me with statistical analysis, and Ed for his feedback when I took his course.

I am also grateful to have worked in the HCI Lab alongside many talented and helpful people. We have a great lab community, and I'm sorry that I was unable to spend the last semester of my degree working in the lab (and going on Starbucks runs) with everyone in person. Thank you to everyone in the lab who helped and supported me throughout the past two years: Rina, Sasha, Damien, Jay; to everyone in the EXII group: Matthew, Hemant, Blaine, Jeremy, Quentin; and to everyone else in the lab! Thank you to Nikhita, Johann, and Greg for being my friends, and helping me stay motivated, whether through advice and support, or just sharing silly jokes and memes.

Finally, thank you to my family; Bill, Susan, and Mary, for always supporting me and for being my guinea pigs (when necessary). I would not be where I am today without them.

Table of Contents

List of Figures **viii**

List of Tables **ix**

1 Introduction **1**

1.1 Contributions 3

1.2 Outline 3

2 Background and Related Work **4**

2.1 Using a Transcription Task for Evaluations 4

2.2 Using a Composition Task for Evaluations 4

2.3 Comparing Speech Recognition and Keyboard Input 6

3 Experiment **8**

3.1 Participants 9

3.2 Apparatus 9

3.3 Tasks 9

3.4 Image and Phrase Stimuli 10

3.5 Study Design 13

3.6 Procedure 13

3.7 Measures 13

4 Results 15

4.1 Total Time 15

4.2 Prep Time 16

4.3 Input Time 16

4.4 Words per Minute 17

4.5 Uncorrected Error Rate 17

4.6 Corrected Error Rate 18

4.7 Corrected and Non-Corrected Sentences 18

4.8 NASA-TLX 19

4.9 Autocorrect Usage 21

4.10 Validating the Composition Task and Protocol 21

4.11 User Preferences 22

5 Discussion 23

5.1 Comparisons with Previous Studies 23

5.2 Subjective Measures 24

5.3 Limitations 25

5.4 Future Work 26

6 Conclusion 27

References 28

List of Figures

1.1 Composition task with keyboard (left) and speech (right). 2

3.1 Four examples of image triads. Semantically similar pairs are in columns, and triads are in different rows. 11

3.2 Transcription phrase set generation process. 12

4.1 Time-related measures: (a) Total Time; (b) Prep Time; (c) Input Time (all with 95% CI). 16

4.2 Words-per-minute (with 95% CI). 17

4.3 Error Rates: (a) Uncorrected; (b) Corrected (with 95% CI). 18

4.4 NASA-TLX Ratings. Lower values correspond to lower mental, physical, and temporal demand, as well as lower effort, lower frustration, and greater performance. 20

List of Tables

4.1 Examples of composed phrases for the image triads ``boy, boat, cat'' and ``woman, mom-and-baby, plane'' in Figure 3.1 22

Chapter 1

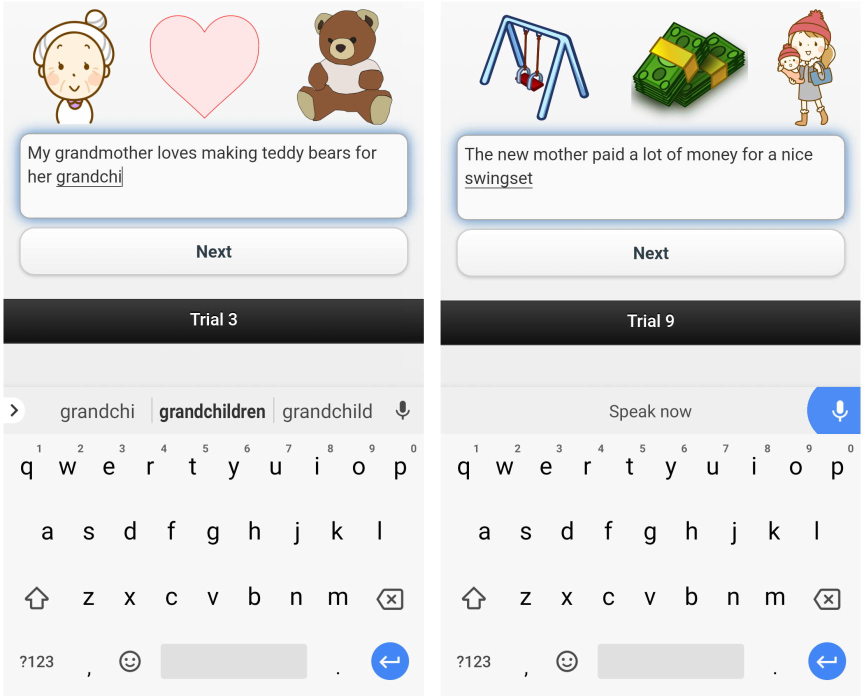
Introduction

Ruan et al. [14] found a state-of-the-art speech recognition system performed nearly 200% faster than touch screen typing when transcribing short phrases on a smartphone. While transcription is commonly used to evaluate text entry, it is less ecologically valid than text composition [9]. Shneiderman argues composing phrases with speech uses more cognitive resources than transcribing phrases with a keyboard [17], and in general, Kristensson and Vertanen show text entry speeds are “bottlenecked” by the time taken for users to conceive their input [7]. The question is whether this cognitive overhead and conception time creates a measurable difference when comparing speech recognition to keyboard typing. Previous work has reported different results, but these have not used state-of-the art speech technology, the composition tasks were not controlled, nor did they use a high number of repetitions.

Designing a composition task is challenging because it can introduce confounds [9]. Vertanen and Kristensson [29] provide a detailed examination of different composition tasks, and a method to measure error rates for composed phrases. However, their task prompt was very open-ended and did not change between trials. Furthermore, all their experiments took place on Mechanical Turk, with an unknown variety of input devices, and used a low number of repetitions. In contrast, we measure the effect of real, “in the moment,” creative composition by using a guided composition task to increase internal validity of our experiment, while keeping good external validity. Our task permits us to have a high number of repetitions in a controlled in-lab experiment with a within-subjects design.

The task presents simple image triads as a composition stimulus (Figure 1.1), and we introduce a protocol that enables a controlled comparison with transcription. Each participant first performs the composition task, then on a following day, they perform the transcription task. This allows us to create a controlled and comparable set of transcription phrases: half are average phrases composed by an initial group of participants, and the other half are phrases composed by the same participant. For a direct comparison to Ruan et al., the transcription portion of our protocol is a near replication.

A 28-participant experiment using this task and protocol found that speech is faster than typing on a keyboard when composing or transcribing. As predicted by Shneiderman, and Kristensson and Vertanen, we find composition with speech recognition requires more preparation time, but we also show the speed of speech entry makes up for it overall. However, the relative advantage in total entry time is less pronounced when composing, where speech is 29% faster than typing, compared to 45% faster when transcribing. NASA-TLX measures also showed there were no significant differences between typing and speech when composing, except for physical demand.


My grandmother loves making teddy bears for her grandchi

The new mother paid a lot of money for a nice swingset

Figure 1.1: Composition task with keyboard (left) and speech (right).

1.1 Contributions

This work contributes new evidence that although speech recognition is faster than typing in both composition and transcription, user perceptions when composing with speech are less clear. We believe these results have more internal validity from using our new protocol to make a direct comparison to transcription.

1.2 Outline

This thesis is organized as follows:

- Chapter 2 outlines previous works that used composition and transcription tasks on desktop and mobile environments, as well as studies that have compared speech recognition and typing.
- Chapter 3 describes our experiment, where we compare a composition task with a transcription task with speech recognition and touchscreen typing.
- Chapter 4 describes the results of this experiment.
- Chapter 5 discusses the findings of our work and how they compare to previous studies, the limitations of our experiment, and possible avenues for future work.
- Chapter 6 concludes by summarizing our findings.

Chapter 2

Background and Related Work

After discussing transcription and composition tasks, we review previous studies that compared speech recognition with typing.

2.1 Using a Transcription Task for Evaluations

Vertanen and Kristensson note "the transcription task is firmly entrenched as the de facto research methodology for text entry experiments" [29]. They explain the primary advantage is that all participants copy the exact same text, so variability decreases and internal validity increases. Phrases used for transcription should have three properties: they should be *memorable*, meaning that after a participant reads the phrase, they can enter it without referring back to the prompt; they should be *representative*, meaning they resemble text people might actually enter; and they should be *replicable*, meaning the phrase set should be publicly available. Many studies have used transcription to evaluate mobile phone text entry. Examples include studies in the wild [15, 24], novel text entry methods [18, 31], evaluating text entry when seated or moving [12, 27], and evaluating input decoders [5, 30].

2.2 Using a Composition Task for Evaluations

However, Vertanen and Kristensson [29] also argue that transcription has low external validity. In the real world, users rarely transcribe messages, they compose original text. A composition task is closer to real-world use, so it has better external validity, and each phrase is memorable since the participant creates it. For replicability, the set of composed phrases (or descriptive statistics characterizing those phrases) can be published. More challenging is designing a composition task to prompt participants to compose representative phrases that are similar across trials for good internal validity.

Studies using composition tasks in desktop evaluations prompt participants to compose multi-sentence or paragraph-length text [6, 11, 23]. This is not representative of typing on phones and controlling variability in long phrases is difficult. An early mobile study by Cox et al. [3] composed short phrases, but the specific prompt they used is not stated.

Vertanen and Kristensson [29] tested prompts for composition, with applications to mobile text entry evaluations. They found a composition task can produce phrases with a consistent length using the prompt: "Imagine you are using a mobile device and need to write a message. We want you to invent and type in a fictitious (but plausible) message. Use your imagination. If you are struggling for ideas, think about things you often write about using your own mobile device." Using this composition task, and a transcription task, Vertanen and Kristensson evaluated a novel desktop text entry technique. They found no difference in text entry speed between the tasks, and only a modest difference in phrase length. Later studies by Yeo et al. [33] and Vertanen, Fletcher, et al. [25] used tasks from Vertanen and Kristensson to evaluate novel mobile text entry methods, on a smartphone and a smartwatch respectively. They found composition to be faster than transcription.

The main focus of the evaluations above is to measure text entry speed (e.g. words-per-minute) independent of its overall impact on trial time. In addition, the prompt used is very open-ended. This may result in divergent phrases in terms of content, and the last sentence in the prompt may lead participants to recall phrases they used, rather than composing a truly original phrase. Vertanen and Kristensson also ran all their experiments on Mechanical Turk, meaning their participants used a wide variety of testing devices in uncontrolled environments. Their first two experiments only used 10 repetitions of their composition task, and their third let participants compose as many phrases as they could in 10 minutes. They did not examine learning effects, which would likely occur when trying to invent new phrases using a static prompt.

Three previous works have used longer composition tasks to evaluate text entry on desktop environments. Both Ogozalek et al. and Karat et al. asked participants to compose short letters, though Ogozalek placed no constraints on these letters [11], while Karat et al. required participants to incorporate three points into their composition [6]. Dunlop et al. [4] introduced a composition task where participants describe the scene in an image. By constraining the topic, composition are more controlled and unlikely to be based on recall. The composition tasks used in our study extend these ideas to increase internal validity.

Another thorny issue absent in transcription tasks is how to measure errors when the intended error-free target of a composed phrase is unknown [9]. Cox et al. [3] asked participants to write down their intended input after they entered each phrase, but this has obvious limitations. Instead, Vertanen and Kristensson [29] show that compositions can be judged by the experimenters or others, so the "correct" target phrase may be determined to calculate error-related measures. We also adopt this method.

2.3 Comparing Speech Recognition and Keyboard Input

Early studies simulated speech recognition with a hidden typist, or used older speech technology. Using a task to compose two letters, Ogozalek et al. [11] found no difference between simulated speech recognition and typing. Tsimhoni et al. [22] compared touchscreen typing with word- and character-based speech recognition when transcribing street addresses while driving, finding word-based speech fastest. In 1999, Karat et al. [6] asked participants to compose replies to specific prompts, and transcribe excerpts from novels, using three speech recognition systems and normal typing. Participants overwhelmingly disliked all speech systems, and speech was slower and more error prone. However, results using simulated or older speech technology are unlikely to generalize, and these studies did not use a mobile phone keyboard.

Cox et al. [3] used a 12-key numerical keypad phone to compare speech recognition, multitap typing, and predictive text typing, also with restricted visual feedback. In both transcription and composition tasks, they found speech fastest. However, the task and prompts are not described, and the 2008 speech recognition system is no longer state of the art.

Smith and Chaparro [19] used a transcription task to compare text entry using a physical keyboard on a mobile phone, a smartphone keyboard, tracing, handwriting, and speech recognition using a more current 2015 speech engine. The keyboard conditions used autocorrect and text prediction. Speech was fastest and, along with a physical keyboard, also the most preferred. But the most relevant previous study for our work is Ruan et al. [14], who used a state-of-the-art 2018 speech recognition system, Baidu Deep Speech 2 [1]. They found speech two times faster than touchscreen typing, both with autocorrect and text prediction and when transcribing English or Mandarin phrases. However, neither of these studies use a composition task, which may have an effect on speech recognition performance or preference.

In summary, it remains unclear if speech is more efficient than a keyboard when composing phrases using modern speech recognition systems. Most previous studies using composition tasks focus on text entry rate, like words-per-minute. Only four also report time measures [3, 11, 22, 25] that may also include additional composition overhead for preparing to enter text. Since our interest is in this overhead, we further decompose trial times into measures like preparation time to better understand differences between transcribing and composing.

Chapter 3

Experiment

Our main goal is to compare speech recognition and typing when composing text. A transcription task is included as a direct comparison to Ruan et al. [14] and to fulfill our secondary objective to compare transcription and composition with these two text entry methods.

A keystroke-level model (KLM) [2] indicates that input with speech recognition may be slower and more error prone, as speech input involves a larger amount of mental preparation before input can begin. In contrast, typing uses a larger number of small mental operations throughout input. Errors are also more time-consuming to fix with speech because one can only make corrections at the end of input. Work by Rochester et al. in the field of psychology also found that people tend to insert more pauses in their speech when performing difficult tasks, due to increased cognitive processing [13]. Participants may pause more when using speech in the composition task, increasing input times.

Based on this, we form two hypotheses:

H1: When composing a short phrase on a smartphone, it is faster to use a keyboard for text entry compared to speech recognition. This was evaluated using time-related measures while entering a phrase.

H2: When composing a short phrase on a smartphone, using a keyboard for text entry results in fewer corrected and uncorrected errors compared to speech recognition. This was evaluated using error measures.

Ruan et al.'s results suggest two more hypotheses regarding transcription:

H3: When transcribing a short phrase on a smartphone, using speech recognition is faster than using a keyboard. This was evaluated in the same manner as H1.

H4: When transcribing a short phrase on a smartphone, using a keyboard for text entry results in fewer corrected and uncorrected errors compared to speech. This was evaluated in the same manner as H2.

3.1 Participants

31 participants were recruited using word-of-mouth and email lists. Data from 3 were discarded due to technical difficulties, leaving 28 participants: 17 male, 11 female, ages 18-58 (M=25 SD = 7.2). All self-reported as fluent English speakers. If they were a non-native speaker, they needed a TOEFL score above 110 (the maximum possible score is 120 [16]) or an equivalent assessment. Three participants experienced occasional issues with the speech recognition software due to accents, or other speech impediments.

All participants owned a smartphone, with 17 using Android and 11 using iOS. In regards to dictation use: 10 participants had never used dictation; 9 participants said they tried it once or twice; 5 used it monthly; and 4 used it daily or weekly.

3.2 Apparatus

A Google Pixel 3 running Android 9.0 was used with the default GBoard keyboard and default speech recognition system. Following Ruan et al. [14], the gesture-based "Swype" keyboard input was disabled and auto-correct, spell check, and the word suggestion strip remained enabled. All tasks were delivered as HTML pages served from a local Node.js application using Ngrok. All events were logged, including characters added or removed from the text entry field and all key presses.

3.3 Tasks

There were two tasks, composition and transcription.

Composition Task — A triad of three clip art images were displayed on the phone's screen (examples are in Figure 3.1). Each image represented common objects or actions, like "boy", "boat", or "cat". The participant was prompted with "You have to compose a short sentence incorporating these three clip art images. You can use the images in any order you want, and you don't have to explicitly name every image. The phrase must make sense, though. The phrase must also be grammatically correct, and words must be spelled properly." The phrase also had to relate the subjects and objects represented by the images in a coherent and believable way. For instance, the sentence "The boy is a boat, and there is a cat" is not believable, and does not synthesize the images well, while the sentence "The boy takes his cat on a boat with him" is acceptable. We verified that all participant phrases were acceptable after the composition task was completed.

Our task is an extension of Dunlop et al.'s image description task [4]. Requiring the topic to be based on three things represented as images restricts the composition for internal validity, but still requires cognitive effort without resorting to simpler recall. Our emphasis on "short sentences" is supported by Lyddy et al.'s results [8] showing entering short phrases of about 70 characters is common on mobile devices.

Transcription Task — A short phrase was displayed on the phone's screen, and the participant was asked to enter it quickly and accurately.

In both tasks, a trial began when the participant pressed a "Start" button and the page loaded. The stimulus was shown in the top part of the screen with a multi-line, full-width text field in the middle of the screen. The participant tapped on the text field to focus it. This activated the keyboard so they could begin typing, or so they could press the "dictate" button to begin speech input, depending on the input condition. When done, they pressed a "Done" button located immediately below the text field. Note that the layout avoided any scrolling or occlusion from the keyboard. In all tasks and input conditions, the participant was instructed to correct spelling and grammar errors with the keyboard before completing the trial.

3.4 Image and Phrase Stimuli

We used a two step process to first generate image triads for the composition task, which all participants performed first. We then used a subset of the resulting composed phrases for the transcription task performed on a later day.

Image Triads for Composition — 56 pairs of royalty free clip art images were collected. Each pair of images portrayed the same object or action, but with visual differences (for example, two apples, but one is red, and the other is green, as in the bottom row of Figure 3.1). With these image pairs, two sets of 20 image triads were generated. Each triad pair had the same semantic meaning (e.g. `''boy,boat,cat''`) but used different images in each pair (see top row of Figure 3.1).

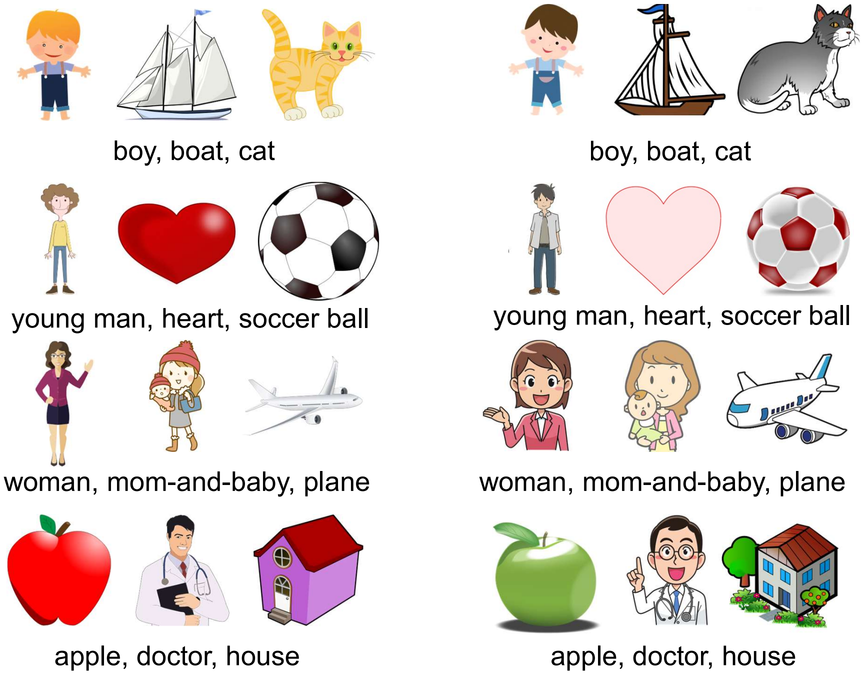

boy, boat, cat
boy, boat, cat
young man, heart, soccer ball
young man, heart, soccer ball
woman, mom-and-baby, plane
woman, mom-and-baby, plane
apple, doctor, house
apple, doctor, house

Figure 3.1: Four examples of image triads. Semantically similar pairs are in columns, and triads are in different rows.

Each triad pair was randomly generated by first partitioning the pairs of clip art images into "people", "animals", and "things" (any images not in the first two categories). A person or animal was randomly selected, then a randomly selected thing, and finally a random image from any of the three categories. The three selected image pairs were shuffled, and each half of the pairs formed a triad. Triads were qualitatively evaluated by two non-authors using the same experiment interface to assess how easily sentences could be generated from generated triads. Ambiguous or difficult triads were removed, leaving 20 image triad pairs for the experiment. Figure 3.1 shows four examples of semantically matching pairs of image triads used in the experiment.

Using two variations of each image triad avoids learning effects with our within-subject design. A participant sees the same semantic triad in both conditions, but created with different images. This way the participant is less likely to recognize images and re-use compositions between conditions.

Phrases for Transcription — For internal validity, we re-used a subset of composed phrases for the transcription task. Two sets of transcription phrases were selected for

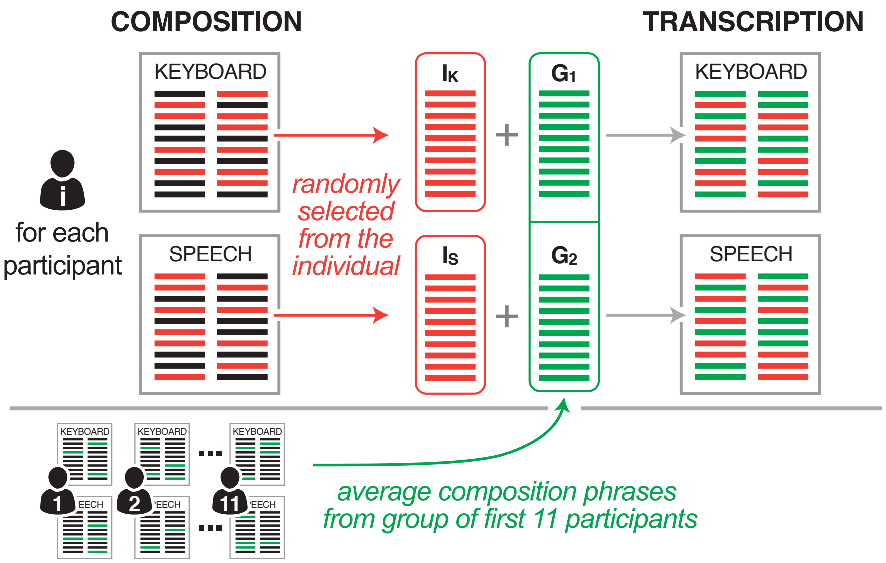


Figure 3.2: Transcription phrase set generation process.

each participant: 20 phrases were drawn from a pool of all phrases composed by the first 11 participants (set G); and 20 phrases were composed by the same participant (set I). Sentences in set G were the 20 closest to the pool's average sentence length. These were randomly partitioned into two 10-sentence subsets, G_1 and G_2. Whether G_1 or G_2 was the starting set for the transcription task alternated between participants. After removing any sentences already selected from the pool, the complete phrase sets were constructed by randomly selecting 10 sentences the participant composed in each input condition (I_S and I_K). These were shuffled with G_1 and G_2, depending on the order of input methods they were assigned. For instance, a participant who started with G_1 and KEYBOARD would first transcribe phrases from G_1 and I_K, and then phrases from G_2 and I_S. Figure 3.2 illustrates this process.

There are publicly available phrases sets for text entry evaluation [10, 26, 28], but constructing the transcription phrase sets in this way better controls our comparison of composition and transcription by reducing variance between participants.

3.5 Study Design

We used a within-subject design with two independent variables: TASK with two levels (COMPOSITION, TRANSCRIPTION), and INPUT with two levels (SPEECH, KEYBOARD). In SPEECH, participants entered the text by dictating it using speech recognition, and in KEYBOARD, they used a standard touchscreen keyboard. Each participant completed 20 trials for all combinations of TASK and INPUT, with the order of INPUT counterbalanced between participants. Image sets and phrase sets were also counterbalanced.

3.6 Procedure

The experiment consisted of two sessions. Participants first performed the composition task in a 40 minute session, followed by the transcription task in another 30 minute session at least one day later. This eliminated fatigue and learning effects, and enabled the phrase generation process described above. Both sessions were conducted in a quiet room to maximize the performance of the speech recognition software and to ensure the comfort of participants.

Participants completed six training examples to ensure they understood the task instructions, and how to use the interface and text-entry method. In the composition task, the examples were six image triads, with example phrase compositions. In the transcription task, the examples were six phrases to transcribe.

After completing all trials for an input method, participants completed a NASA-TLX assessment. At the end of each session, participants were asked which method they preferred, and their reasons for that choice. A brief survey was conducted after the composition task to gather information on smartphone use, whether they had used speech recognition before and their reasons for doing so, and demographic information.

3.7 Measures

Several time-related measures were collected for each trial:

Total Time: The time taken for a single trial, from page load to the participant's last input.

Prep Time: The time from page load, to the first input in the text area. For SPEECH, we found there is a delay after the participant begins to speak until that input appears in the text box. By examining videos and experiment logs, we calculate this average delay to be 1.69s, and subtract this from SPEECH *Prep Times* to compensate.

Input Time: The time from a participant's first input in the text area, to their last.

Words Per Minute (WPM): Defined per trial as the number of characters in the final phrase divided by the trial time in minutes, divided by 5 (the standard "word length" [32]).

Two error rates were considered: the *Corrected Error Rate*, which is the number of corrected characters divided by the sum of all correct and fixed characters, and the *Uncorrected Error Rate*, which is the number of uncorrected characters divided by the sum of all correct and fixed characters [20]. Both error rates were calculated with the same method as Soukoreff and Mackenzie [21].

Calculating error rates for TRANSCRIPTION is straightforward since the correct version is known beforehand. For COMPOSITION, the process was slightly more complex, using a modified version of the process outlined in Vertanen and Kristensson's work [29]. One of the experimenters and an external evaluator (who was compensated for their time) reviewed the composed phrases, and independently constructed two sets of "correct" phrases. The two evaluators had a 87% agreement rate, with less than 2 characters of difference in 95% of phrases. The final error rates for COMPOSITION were determined by averaging the error rates calculated from the two sets.

Chapter 4

Results

For each combination of TASK and INPUT, trials with a *Total Time* more than 3 standard deviations from the mean were excluded as outliers: 48 trials (2.14%) were removed.

According to the Shapiro-Wilk Normality test, none of the residuals of the collected data are normally distributed. To run repeated-measures ANOVAs, data was transformed either with Box-Cox tranformations, or the non-parametric Aligned Rank Transformation procedure (ART). Tukey's HSD was used for post-hoc comparisons. Results were considered statistically significant if $\alpha < .05$[1].

4.1 Total Time

We found that KEYBOARD trials were slower on average for both COMPOSITION and TRANSCRIPTION (Figure 4.1a). There is a significant main effect of INPUT ($F_{1,27} = 115.99$, $p < .001$, $\eta^2_G = .43$), and TASK ($F_{1,27} = 185.95$, $p < .001$, $\eta^2_G = .58$). More relevant, there is a significant interaction for TASK × INPUT ($F_{1,27} = 37.12$, $p < .001$, $\eta^2_G = .09$). Post-hoc comparisons found significant differences when comparing two INPUTS between a TASK, and when comparing two TASKS between an INPUT ($p < .05$). For COMPOSITION, SPEECH (20.8s) was faster than KEYBOARD (29.19s), and for TRANSCRIPTION, SPEECH (9.25s) was also faster than KEYBOARD (16.92s). This represents a 29% decrease in *Total Time* for SPEECH in COMPOSITION and a 45% decrease for SPEECH in TRANSCRIPTION.

[1] Detailed statistical analysis available at ns.inria.fr/loki/speech-type/

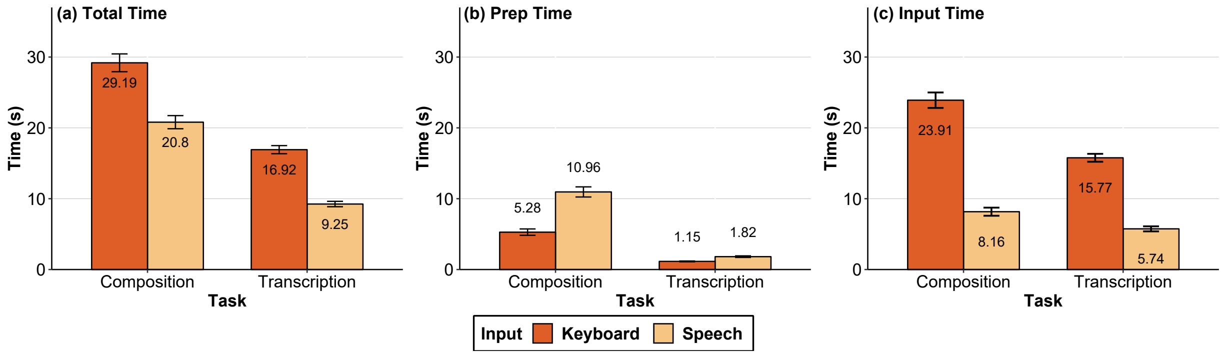


Figure 4.1: Time-related measures: (a) Total Time; (b) Prep Time; (c) Input Time (all with 95% CI).

4.2 Prep Time

SPEECH required more *Prep Time* than KEYBOARD for COMPOSITION. Though *Prep Times* were much smaller for TRANSCRIPTION, SPEECH was still found to have a longer average *Prep Time* (Figure 4.1b). There is a significant main effect of INPUT ($F_{1,27} = 60.62$, $p < .001$, $\eta^2_G = .28$), and TASK on Box-Cox transformed *Prep Time* ($F_{1,27} = 357.41$, $p < .001$, $\eta^2_G = .73$). There was also a significant interaction between TASK and INPUT ($F_{1,27} = 7.70$, $p < .01$, $\eta^2_G =$ .03). Post-hoc comparisons found differences when comparing two INPUTS between a TASK, and when comparing two TASKS between an INPUT ($p < .001$). For COMPOSITION, SPEECH (10.96s) was slower than KEYBOARD (5.28s), and for TRANSCRIPTION, SPEECH (1.81s) was also slower than KEYBOARD (1.15s). This represents a 52% decrease in *Prep Time* for KEYBOARD in COMPOSITION and a 36% decrease for KEYBOARD in TRANSCRIPTION.

4.3 Input Time

SPEECH had faster input times for both tasks (Figure 4.1c). Repeated-measures ANOVAs on the Box-Cox transformed data found a significant main effect of INPUT ($F_{1,27} = 374.04$, $p < .001$, $\eta^2_G = .75$), and TASK ($F_{1,27} = 67.27$, $p < .001$, $\eta^2_G = .26$), but not the interaction between the two. For COMPOSITION, SPEECH (8.16s) was faster than KEYBOARD (23.91s), and for TRANSCRIPTION, SPEECH (5.74s) was also faster than KEYBOARD (15.77s). This represents a 76% decrease in *Input Time* for SPEECH in COMPOSITION and a 64% decrease for SPEECH in TRANSCRIPTION.

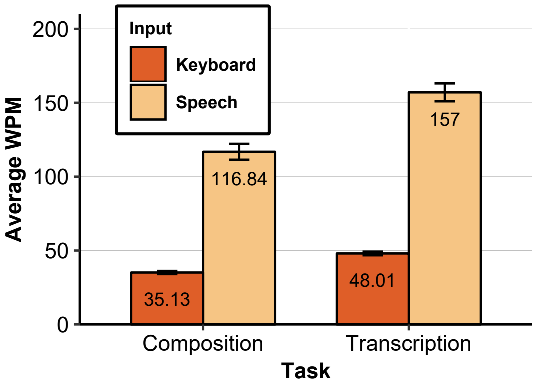


Figure 4.2: Words-per-minute (with 95% CI).

4.4 Words per Minute

SPEECH resulted in considerably higher *WPM* for both COMPOSITION and TRANSCRIPTION (Figure 4.2). A repeated-measures ANOVA run on the Box-Cox transformed data found a significant main effect of INPUT ($F_{1,27} = 488.49$, $p < .001$, $\eta_G^2 = .86$), and TASK ($F_{1,27} = 79.07$, $p < .001$, $\eta_G^2 = .28$), but not the interaction between the two. For COMPOSITION, SPEECH (116.5) was faster than KEYBOARD (35.12), and for TRANSCRIPTION, SPEECH (156.74) was also faster than KEYBOARD (48.13). This represents a 232% increase in *WPM* for SPEECH in COMPOSITION and a 226% increase for SPEECH in TRANSCRIPTION.

4.5 Uncorrected Error Rate

SPEECH had slightly higher *Uncorrected Error Rates* overall (Figure 4.3a). A repeated-measures ANOVA using ART data found a main effect of INPUT ($F_{1,81} = 9.23$, $p < .005$). Overall rates for KEYBOARD are 0.4% and SPEECH 0.7%, however this represents little difference from a practical point of view.

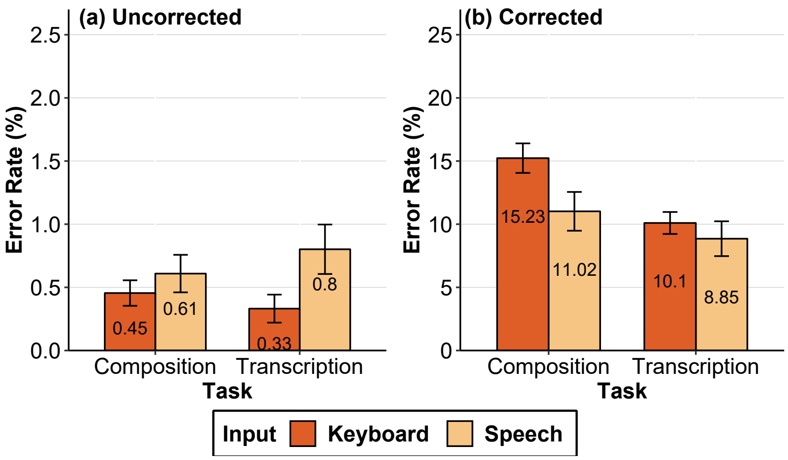


Figure 4.3: Error Rates: (a) Uncorrected; (b) Corrected (with 95% CI).

4.6 Corrected Error Rate

SPEECH had a 22% lower *Corrected Error Rate* than KEYBOARD overall, and COMPOSITION rates are slightly higher than TRANSCRIPTION overall (Figure 4.3b). A repeated-measures ANOVA using ART data found main effects of TASK ($F_{1,81} = 11.58$, $p < .001$) and INPUT ($F_{1,81} = 9.05$, $p < .005$), but no interaction. Overall rates for KEYBOARD are 12.7% and SPEECH are 10%. Overall rates for COMPOSITION are 13.1% and TRANSCRIPTION are 9.5%.

The 95% confidence error bars in Figure 4.3b suggest the overall main effect for INPUT is primarily due to the composition task. Since there was no interaction involving TASK, we divide the data into two sets, COMPOSITION only and TRANSCRIPTION only, then conduct separate analysis. As expected, when using composition data only, a repeated-measures ANOVA using ART data found a main effect of INPUT ($F_{1,27} = 10.04$, $p < .005$). Here, KEYBOARD rates are 15.2% and SPEECH is 11%, a 28% decrease for SPEECH. There was no significant effect when using transcription data only.

4.7 Corrected and Non-Corrected Sentences

We examined the differences between trials with and without corrections, finding that sentences with corrections were slower for all time-related measures. One-way ANOVAs on Box-Cox transformed data found a significant main effect of having corrections on *Total Time* ($F_{1,27} = 178.66$, $p < .001$, $\eta^2_G = .46$), *Input Time* ($F_{1,27} = 200.80$, $p < .001$, $\eta^2_G = .71$), and *WPM* ($F_{1,27} = 159.78$, $p < .001$, $\eta_G^2 = .63$). Post-hoc comparisons found differences between having corrections, and not having corrections for all three of these measures ($p < .001$). Trials with corrections had an average *Total Time* of 25.4s, an average *Input Time* of 18s, and an average *WPM* of 54.6. In comparison, trials without corrections were much faster and had higher *WPM*, with an average *Total Time* of 14.7s, an average *Input Time* of 6.57s, and a *WPM* of 141.1.

Trials with corrections also had higher *Uncorrected Error Rates* on average. A one-way ANOVA on the ART data found a significant main effect of correction on *Uncorrected Error Rates* ($F_{1,186} = 8.64$, $p < .005$). Post-hoc comparisons found a difference between having corrections, and not having corrections ($p < .005$). Trials with corrections had an average *Uncorrected Error Rate* of 0.65%, compared to an average rate of 0.52% for trials without corrections.

4.8 NASA-TLX

All measures except *Effort* were not normally distributed, according to the Shapiro-Wilk Normality test. As such, every measure was analyzed using Friedman analyses, and Wilcoxon signed-rank tests with Holm-Bonferonni corrections for post-hoc comparisons. To enable non-parametric tests between combinations of INPUT and TASK, we create a 4-level factor representing each combination, and use this in all tests below. For all TLX measures, median values are reported.

In general, we found that SPEECH outperformed KEYBOARD in most measures for TRANSCRIPTION. However, for COMPOSITION, SPEECH only outperformed KEYBOARD in *Physical Demand*, but not in any other measures (Figure 4.4).

Physical Demand: SPEECH was much less physically demanding than KEYBOARD regardless of TASK. A Friedman analysis found a significant effect on *Physical Demand* ($\chi^2(3) = 36.9$, $p < .001$), and post-hoc comparisons found differences between INPUTS in both TASKS ($p < .005$). For COMPOSITION, there was a 100% increase for KEYBOARD (30) compared to SPEECH (15), and for TRANSCRIPTION, there was a 183% increase for KEYBOARD (42.5) compared to SPEECH (15).

Mental Demand: For COMPOSITION, there was no significant difference in participant ratings between SPEECH and KEYBOARD, but for TRANSCRIPTION, SPEECH was perceived as less mentally demanding. A Friedman analysis found a significant effect on *Mental Demand* ($\chi^2(3) = 35.87$, $p < .001$). Post-hoc comparisons found a difference between SPEECH and KEYBOARD for TRANSCRIPTION only ($p < .01$): SPEECH (17.5) was 46% less mentally

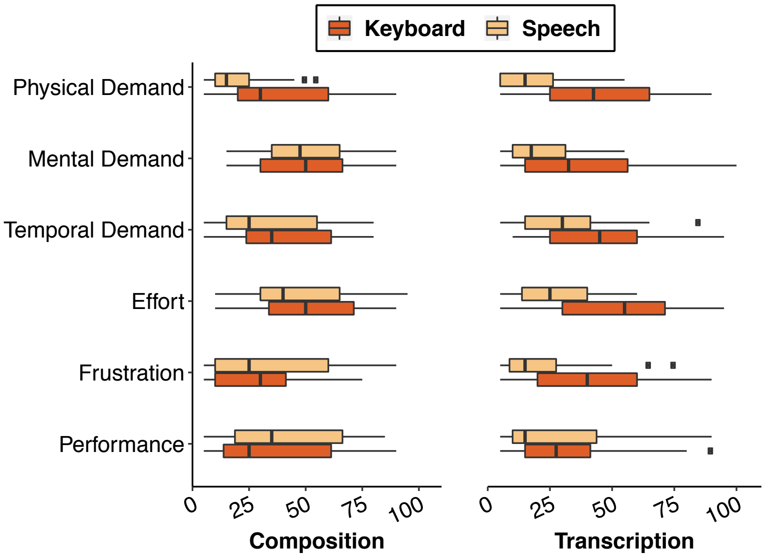


Figure 4.4: NASA-TLX Ratings. Lower values correspond to lower mental, physical, and temporal demand, as well as lower effort, lower frustration, and greater performance.

demanding than KEYBOARD (32.5). In addition, there was a difference between COMPOSITION and TRANSCRIPTION for SPEECH ($p < .001$): COMPOSITION (47.5) was 171% more mentally demanding than TRANSCRIPTION (17.5).

Temporal Demand: Again, there was no significant difference in participant ratings between SPEECH and KEYBOARD for COMPOSITION, but for TRANSCRIPTION, SPEECH was perceived as less temporally demanding. A Friedman analysis found a significant effect on *Temporal Demand* ($\chi^2(3) = 11.6$, $p < .02$). Post-hoc comparisons found SPEECH (15) had 33% lower *Temporal Demand* than KEYBOARD (40) for TRANSCRIPTION ($p < .05$).

Effort: Continuing the trend, there was was no significant difference between SPEECH and KEYBOARD for COMPOSITION, but there was again for TRANSCRIPTION. A Friedman analysis found a significant effect on *Effort* ($\chi^2(3) = 28.41$, $p < .001$), and post-hoc comparisons found a 120% increase in ratings for KEYBOARD (55) compared to SPEECH (25) for TRANSCRIPTION ($p < .001$). Post-hoc comparisons also found COMPOSITION (40) was rated 60% greater than TRANSCRIPTION (25) for SPEECH ($p < .01$).

Frustration: Once more, SPEECH was less frustrating for TRANSCRIPTION, but there was no differences detected for COMPOSITION. A Friedman analysis found a significant effect on *Frustration* ($\chi^2(3) = 15.33$, $p < .002$). For TRANSCRIPTION, post-hoc comparisons showed SPEECH (15) was 62.5% less frustrating than KEYBOARD (40).

Performance: There were no differences between KEYBOARD and SPEECH for either TASK. Although Friedman analysis found a borderline significant effect on *Performance* ($\chi^2(3) = 9.59$, $p < .05$), post-hoc comparisons did not detect any differences between TASK or INPUT.

4.9 Autocorrect Usage

We define autocorrect and word suggestion strip use as multiple characters appearing between timestamps in the keystrokes array for a trial. Around 40% of trials used autocorrect at least once for transcription and composition. Curiously, trials that used autocorrect at least once had lower WPM on average. A one-way ANOVA found a significant main effect of autocorrect use on WPM ($F_{1,27} = 31.77$, $p < 0.001$, $\eta^2_G = .14$). Post-hoc comparisons found a difference between using autocorrect, and not using autocorrect ($p < 0.005$).

4.10 Validating the Composition Task and Protocol

To validate our experimental protocol, we examined whether there were any learning effects, or if the different image triad sets and phrase sets had any effect on our results. We also examined aspects of the phrases composed by participants.

We created four blocks of five trials to investigate possible learning effects. Repeated-measures ANOVAs found a significant main effect of block on *Prep Time* and *Total Time*, but there were no significant interactions between block, INPUT, or TASK. Post-hoc analysis revealed no significant differences between blocks due to multiple comparison adjustments, leading us to conclude there were no observable learning or fatigue effects.

Repeated-measures ANOVAs did not find a significant effect of image sets or phrase sets on any of our collected measures. This highlights that our image and phrase sets can be considered as equivalent and do not represent confounding variables in our experiment.

The average length of a composed phrase was 58 characters (SD = 18), with a maximum length of 140 characters, and a minimum length of 16 characters. In comparison, Vertanen and Kristensson had an average length of 38 characters (SD = 25), 52 characters (SD = 27), and 39 characters (SD = 33) in their three experiments [29]. The triad with the longest average compositions was `''woman, mom-and-baby, plane''`, with an average phrase length of 70.75 characters, and the triad with the shortest average compositions was `''young man, heart, soccer ball''` with an average length of 41.1 characters. These triads can be seen

P11: The boy placed his cat in the miniature boat to take a picture
P22: While chasing a cat, the boy found himself on a boat.
P8: The boy's cat was afraid of going in a boat.
P27: Johnny sold his toy ship to buy a pet cat.

P9: The flight attendant was friendly to the family.
P4: A stewardess is welcoming a young mother and her baby onboard.
P25: A woman takes a flight to visit her sister and newborn nephew.
P16: The mom took a flight with her baby to see her friend

Table 4.1: Examples of composed phrases for the image triads `‘‘boy, boat, cat’’` and `‘‘woman, mom-and-baby, plane’’` in Figure 3.1

in Figure 3.1, in rows 2 and 3. Examples of phrases participants composed for the triads `‘‘boy, boat, cat’’` and `‘‘woman, mom-and-baby, plane’’` are found in table 4.1.

A one way ANOVA on the ART data found a main effect of triad on message length ($F_{19,20} = 3.93$, $p < .002$), indicating that image triads did not generate sentences of the same average length. Post-hoc tests revealed that only two image triads generated significantly longer sentences than others. Image triads did not have a significant main effect on any other measure.

Following Vertanen and Kristensson [29], we calculated an Out-Of-Vocabulary (OOV) rate with a lexicon of 64K most common words from an email corpus[2]. We had a 1% OOV rate compared to 2.3% (SD = 3.4) and 9.7% (SD = 12.9) in Vertanen and Kristensson's two experiments [29]. Our low OOV demonstrates that participants did not use great amounts of texting shorthand, or other forms of slang.

4.11 User Preferences

In the transcription task, 27 participants preferred speech to keyboard. However, for composition, a slim majority of participants (14) preferred typing over speech recognition, with 12 preferring speech, and 1 having no strong preference for either input.

[2] https://keithv.com/software/composition/

Chapter 5

Discussion

H1 is rejected. We found that speech was faster than keyboard for every time-related measure except *Prep Time*, and that speech resulted in significantly higher *WPM*.

H2 is rejected. Speech had a much lower corrected error rate, both in the composition task and overall, with only a slightly greater uncorrected error rate.

H3 is confirmed. As with composition, speech was faster than keyboard for every time-related measure except *Prep Time*, and resulted in significantly higher *WPM*.

H4 is inconclusive. Although speech had lower corrected error rates overall and only slightly higher uncorrected error rates, we did not detect differences for transcription specifically.

While we could replicate Ruan et al.'s results for speed, we found different results for error rate.

5.1 Comparisons with Previous Studies

For both tasks, we found similar results to Ruan et al. [14], Smith and Chaparro [19], and Cox et al.[3], with speech being superior for all measures except *Prep Time* and *Uncorrected Error Rate*. This contrasts with the findings of Ogozalek et al. [11], who found speech did not improve performance, and Karat et al. [6], who found that speech was slower. Our analysis demonstrates that transcription trials were, on average, faster, than composition trials. In comparison, Vertanen and Kristensson [29] did not find any difference between transcription and composition, Vertanen et al. [25] found transcription was faster, and

Yeo et al. [33] and Karat et al.found composition was faster. Similar to Ruan et al., we found that speech had a lower *Corrected Error Rate* and a (slightly) higher *Uncorrected Error Rate* when both composition and transcription are combined, but not when only considering transcription.

In their first experiment, Vertanen and Kristensson found there was a 57% decrease in transcription preparation times on a keyboard when compared to composition. In comparison, we found a 78% decrease in preparation times between our transcription and composition tasks on a keyboard. Of course, our experiment was conducted on a smartphone, while Vertanen and Kristensson used full-sized keyboards.

Even though preparation times are much higher for speech recognition, speech input was still faster than keyboard input in both tasks. As evidenced by the disparity between speech and keyboard input for input times and words-per-minute, speech input is so fast that it makes up for the penalty incurred by higher preparation times.

Ruan et al. also calculated a *utilized bandwidth* measure which is, in their words, "the proportion of keystrokes that contributed to correct parts of the final transcribed string". We also calculated and examined this measure, and saw the same patterns as the other measures we examined.

5.2 Subjective Measures

For transcription, we again found similar results to Ruan et al. Participants overwhelmingly preferred speech over typing, and the TLX results favour speech in most measures. In post-session interviews participants noted that speech was more comfortable, required less effort, and less physical strain than typing. Several mentioned that the speech recognition software did most of the work for them, requiring fewer corrections. In the words of one participant, *"[I just] had to read stuff"* [P1].

In contrast, participants favoured typing both in the composition task by a slim majority (14 versus 12). For composition, we did not find any significant differences for any TLX measure except *Physical Demand*, where speech rated much better than typing. Participants commented in post-session interviews that they felt speech required more mental effort, as they had to think of the sentence ahead of time.They also stated that typing gave them more freedom to edit text in real-time, as opposed to waiting for the speech recognition software to finish. Several participants also felt their phrases were more creative when typing.

While using speech recognition had great speed advantages over typing, there are several reasons why participants did not overwhelmingly prefer speech to typing for composition. In post-session interviews after both tasks, participants noted the privacy issues that result from using speech recognition in public. Even if privacy was not a concern, many mentioned feeling "awkward" or "embarassed" about speaking to their phones in public, making statements such as "it's weird to talk to my phone", and that "[they] don't like saying things aloud". Comments about privacy and perceived awkwardness were prevalent in the composition task. It is evident that even though speech has a significant speed advantage over typing, there are still factors that discourage people from using speech recognition in public. Indeed, a few participants stated that they felt uncomfortable using speech recognition even in the presence of an experimenter in an otherwise private room.

Several participants also mentioned that speech recognition performed better than they had expected. Even so, there were still several participants who experienced issues with speech recognition, which likely influenced those participants' preference of input method.

5.3 Limitations

Ruan et al.'s participants were all native English speakers [14]. In contrast, we allowed for non-native speakers with a TOEFL score of 110 or above to participate in our experiment. The maximum possible TOEFL score is 120, meaning that all non-native English speakers who participated in our experiment achieved 91% or higher on this assessment [16]. Still, allowing participants who were not fluent in English may have affected our speed and error rate results. It is also possible that some participants were not entirely truthful in reporting their English fluency levels, but a formal assessment of language proficiency was not possible due to time and resource constraints.

Our experiment was also conducted in a quiet room. While this helped participants feel comfortable, it may have artificially augmented the results for speech. Speech recognition would likely be affected if the experiment took place with more ambient noise.

A different composition task, such as asking participants for longer compositions as in Karat et al. [6] and Dunlop et al. [23], may have resulted in greater subjective preference for composition, or a decrease in the speed advantages of SPEECH. However, results from the demographic survey show that many of the participants who had previously used speech recognition only used it for tasks that required a short burst of input, such as sending text messages, or for Google searches. Lyddy et al.'s analysis of text messages sent by university students found they have an average length of 70 (SD = 59.4) characters [8].

As the average length of our composed phrases was 58 characters, this suggests our task is representative of typical text input on a smartphone.

5.4 Future Work

Future work could compare typing and voice input with a more elaborate composition task, such as asking participants to write a short paragraph based off a writing prompt, or displaying more images, or some other task that would require participants to compose multiple phrases as in Ogozalek et al. [11] and Dunlop et al. [23]. It is possible that speech recognition could lose its speed advantage if participants are forced to input more punctuation. Participants may also become more frustrated with speech recognition when using it for a longer period, leading to a greater preference for typing.

Our composition task could also be compared to the one used by Vertanen and Kristensson [29]. It would be interesting to examine whether participants find one task harder than the other, and if the two tasks produce sentences of similar length. Vertanen and Kristensson's task may also favour recall over true composition, with participants composing several phrases with similar subject matter. In contrast, our task presents participants with a new stimulus for each trial.

Chapter 6

Conclusion

Though speech is faster than typing for composition and transcription tasks, speech recognition results in higher preparation times. Speech has a lower error rate than keyboard during composition, but not during transcription. Speech did not have a significant advantage over typing for composing, except in physical demand. A slight majority of people preferred typing for composition.

While speech recognition is the more efficient text entry method for composing short phrases on a smartphone, our results also suggest that people may continue to use a keyboard given their subjective impressions of the experience.

References

[1] Dario Amodei, Rishita Anubhai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Jingdong Chen, Mike Chrzanowski, Adam Coates, Greg Diamos, Erich Elsen, Jesse Engel, Linxi Fan, Christopher Fougner, Tony Han, Awni Y. Hannun, Billy Jun, Patrick LeGresley, Libby Lin, Sharan Narang, Andrew Y. Ng, Sherjil Ozair, Ryan Prenger, Jonathan Raiman, Sanjeev Satheesh, David Seetapun, Shubho Sengupta, Yi Wang, Zhiqian Wang, Chong Wang, Bo Xiao, Dani Yogatama, Jun Zhan, and Zhenyao Zhu. Deep speech 2: End-to-end speech recognition in english and mandarin. *CoRR*, abs/1512.02595, 2015.

[2] Stuart K. Card, Thomas P. Moran, and Allen Newell. The keystroke-level model for user performance time with interactive systems. *Commun. ACM*, 23(7):396–410, July 1980.

[3] Anna L. Cox, Paul A. Cairns, Alison Walton, and Sasha Lee. Tlk or txt? using voice input for sms composition. *Personal and Ubiquitous Computing*, 12(8):567–588, Nov 2008.

[4] Mark D. Dunlop, Emma Nicol, Andreas Komninos, Prima Dona, and Naveen Durga. Measuring inviscid text entry using image description tasks. In *Inviscid Text Entry and Beyond*, CHI'16 Workshop, 2016.

[5] Christina L. James and Kelly M. Reischel. Text input for mobile devices: Comparing model prediction to actual performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 365–371, New York, NY, USA, 2001. ACM.

[6] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 568–575, New York, NY, USA, 1999. ACM.

[7] Per Ola Kristensson and Keith Vertanen. The inviscid text entry rate and its application as a grand goal for mobile text entry. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices &#38; Services*, MobileHCI '14, pages 335–338, New York, NY, USA, 2014. ACM.

[8] Fiona Lyddy, Francesca Farina, James Hanney, Lynn Farrell, and Niamh Kelly O'Neill. An Analysis of Language in University Students' Text Messages*. *Journal of Computer-Mediated Communication*, 19(3):546–561, 04 2014.

[9] I. Scott MacKenzie and R. William Soukoreff. Text entry for mobile computing: Models and methods,theory and practice. *Human-Computer Interaction*, 17(2-3):147–198, 2002.

[10] I. Scott MacKenzie and R. William Soukoreff. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '03, pages 754–755, New York, NY, USA, 2003. ACM.

[11] V. Ogozalek and J. Van Praag. Comparison of elderly and younger users on keyboard and voice input computer-based composition tasks. *SIGCHI Bull.*, 17(4):205–211, April 1986.

[12] Kathleen J. Price, Min Lin, Jinjuan Feng, Rich Goldman, Andrew Sears, and Julie A. Jacko. Data entry on the move: An examination of nomadic speech-based text entry. In Christian Stary and Constantine Stephanidis, editors, *User-Centered Interaction Paradigms for Universal Access in the Information Society*, pages 460–471, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[13] Sherry R Rochester. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research*, 2(1):51–81, 1973.

[14] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(4):159:1–159:23, January 2018.

[15] Richard Schlögl, Christoph Wimmer, and Thomas Grechenig. Hyper typer: A serious game for measuring mobile text entry performance in the wild. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, pages LBW0259:1–LBW0259:6, New York, NY, USA, 2019. ACM.

[16] Educational Testing Service. Toefl ibt test scores, 2019.

[17] Ben Shneiderman. The limits of speech recognition. *Commun. ACM*, 43(9):63–65, September 2000.

[18] Miika Silfverberg, I. Scott MacKenzie, and Panu Korhonen. Predicting text entry speed on mobile phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 9–16, New York, NY, USA, 2000. ACM.

[19] Amanda L. Smith and Barbara S. Chaparro. Smartphone text input method performance, usability, and preference with younger and older adults. *Human Factors*, 57(6):1015–1028, 2015. PMID: 25850116.

[20] R. William Soukoreff and I. Scott MacKenzie. Measuring errors in text entry tasks: An application of the levenshtein string distance statistic. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, pages 319–320, New York, NY, USA, 2001. ACM.

[21] R. William Soukoreff and I. Scott MacKenzie. Metrics for text entry research: An evaluation of msd and kspc, and a new unified error metric. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 113–120, New York, NY, USA, 2003. Association for Computing Machinery.

[22] Omer Tsimhoni, Daniel Smith, and Paul Green. Address entry while driving: Speech recognition versus a touch-screen keyboard. *Human Factors*, 46(4):600–610, 2004. PMID: 15709323.

[23] Keith Vertanen, Mark Dunlop, James Clawson, Per Ola Kristensson, and Ahmed Sabbir Arif. Inviscid text entry and beyond. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 3469–3476, New York, NY, USA, 2016. ACM.

[24] Keith Vertanen, Justin Emge, Haythem Memmi, and Per Ola Kristensson. Text blaster: A multi-player touchscreen typing game. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 379–382, New York, NY, USA, 2014. ACM.

[25] Keith Vertanen, Crystal Fletcher, Dylan Gaines, Jacob Gould, and Per Ola Kristensson. The impact of word, multiple word, and sentence input on virtual keyboard decoding performance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 626:1–626:12, New York, NY, USA, 2018. ACM.

[26] Keith Vertanen, Dylan Gaines, Crystal Fletcher, Alex M. Stanage, Robbie Watling, and Per Ola Kristensson. Velociwatch: Designing and evaluating a virtual keyboard for the input of challenging text. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 591:1–591:14, New York, NY, USA, 2019. ACM.

[27] Keith Vertanen and Per Ola Kristensson. Parakeet: A continuous speech recognition system for mobile touch-screen devices. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*, IUI '09, pages 237–246, New York, NY, USA, 2009. ACM.

[28] Keith Vertanen and Per Ola Kristensson. A versatile dataset for text entry evaluations based on genuine mobile emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, pages 295–298, New York, NY, USA, 2011. ACM.

[29] Keith Vertanen and Per Ola Kristensson. Complementing text entry evaluations with a composition task. *ACM Trans. Comput.-Hum. Interact.*, 21(2):8:1–8:33, February 2014.

[30] Daryl Weir, Henning Pohl, Simon Rogers, Keith Vertanen, and Per Ola Kristensson. Uncertain text entry on mobile devices. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2307–2316, New York, NY, USA, 2014. ACM.

[31] Jacob O. Wobbrock, Duen Horng Chau, and Brad A. Myers. An alternative to push, press, and tap-tap-tap: Gesturing on an isometric joystick for mobile phone text entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 667–676, New York, NY, USA, 2007. ACM.

[32] Hisao Yamada. A historical study of typewriters and typing methods: from the position of planning japanese parallels. *Journal of Information Processing*, 2(4):175–202, feb 1980.

[33] Hui-Shyong Yeo, Xiao-Shen Phang, Steven J. Castellucci, Per Ola Kristensson, and Aaron Quigley. Investigating tilt-based gesture keyboard entry for single-handed text entry on large devices. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 4194–4202, New York, NY, USA, 2017. ACM.