Speech Recognition Versus Typing: A Mixed-Methods Evaluation

Johanna Kneifel
johanna_kneifel@sfu.ca
SIAT, Simon Fraser University
Vancouver, Canada

Ohoud Alharbi
King Saud University
Riyadh, Saudi Arabia
omalharbi@ksu.edu.sa

Wolfgang Stuerzlinger
SIAT, Simon Fraser University
Vancouver, Canada

ABSTRACT

Mobile text entry has become a main mode of communication. To make text entry as efficient as possible, helpful features, such as autocorrect and speech recognition have been developed. In our study, we confirmed previous results in that speech recognition was faster, while the error rate for typing was lower. To analyze this in more depth, we performed semi-structured interviews identifying participants' text-entry preferences, specific pain points that occur, and potential suggestions for improving the editing experience.

CCS CONCEPTS

• **Human-centered computing** → **Keyboards**.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Johanna Kneifel, Ohoud Alharbi, and Wolfgang Stuerzlinger. 2022. Speech Recognition Versus Typing: A Mixed-Methods Evaluation. In *TEXT2030: MobileHCI'22 Workshop on Shaping Text Entry Research in 2030, October 1, 2022, Vancouver, Canada.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Text entry on mobile devices is commonplace [4], but autocorrect can make text entry a frustrating experience [23]. Speech recognition for text entry is fast but can also result in frustration, e.g., when having to correct errors through typing [3, 21]. Ruan et al. [19] investigated text entry efficiency for both English and Mandarin Chinese. Speech recognition was quicker for both languages. Interestingly, despite fewer errors occurring with speech, there were more errors in the final transcribed text for speech-based text entry. Foley et al. [8] also found that both transcribing and composing text through speech recognition is faster than typing. With composition tasks, the time difference between the two modalities was smaller, potentially due to the additional mental load for composing sentences.

Our study aims to verify whether speech or typing is more efficient for text entry. We use a transcription task to reduce mental workload and also evaluate the error rate for both modalities. We expand on previous work through adding a semi-structured interview, documenting participant's personal experiences with typing and speech recognition, both positive and negative, as well as participant's suggestions for new text editing methods on mobile devices.

TEXT2030, October 1, 2022, Vancouver, Canada

2 RELATED WORK

Text entry errors fall into specific types: insertion, omission, and substitution errors, with the latter being the most salient type [16, 21]. Thus, common errors could be predicted and corrected depending on the user group. Wang et al. thus used an "elastic probabilistic model for input prediction," which made typing more efficient for users with Parkinson's symptoms [22]. Others explored multi-modal correction mechanisms, e.g., with gaze and speech or touchscreen and gestures, to enhance the error correction experience [18, 20]. Buschek et al. found that autocorrections account for only 0.6 percent of all events in an "in the wild" study [6], but other work identified substantial negative impacts of autocorrect errors on mental and physical demand, as well as effort and frustration [2, 16]. Still, within a study a certain amount of errors needs to occur to reliably gauge user reactions to autocorrect, which can be challenging especially for composition tasks. Gains et al. encouraged participants to increase the complexity of their text through guided instruction [10], with higher rates of error correction occurring when more complex terminology was encouraged.

To facilitate text entry, users rely on technology to correct some of their mistakes for them, or at least suggest a range of alternatives. Arnold et al. [5] found that phrase suggestions affected composition, as participants adjusted their text according to the phrases. As it is not always possible to reasonably regulate user input and measure error rates, we used a transcription task with a predetermined phrase set with a defined percentage of more complex words.

When designing text-entry interfaces and features, factors that can encourage users to adopt certain text entry features [7, 13, 17, 18], such as framing and presentation [7, 17], need to be considered. We thus made sure to present the tasks neutrally. To collect more information around personal experiences and preferences, we performed a semi-structured interview, as this method can yield a wide range of insights complementary to the main task.

Previous studies of speech recognition have defined key characteristics that are important to measuring and analyzing speech recognition performance and adoption [14–16]. Uses of speech recognition range from controlling a system, turning on lights, to entering text, or even more complex tasks [1, 12, 24], with speech-activated home assistants being most common [1, 9, 11]. Koester [12] identified that the fastest users also employed the best correction strategies. Thus, when evaluating typing speed and efficiency, correction strategies need to be considered as well.

3 METHOD

We used a repeated-measures design, in which participants completed both typing and speech recognition text entry tasks, in counterbalanced order, followed by an interview. We locally recruited ten participants (five male, five female, 18-24 years), but had to exclude two due to a logging issue. We created a locally hosted

web application for data collection, which prompts users to enter text into a field, while logging all occurring events at the keystroke and event level. Our study was conducted via Zoom, recorded with participant consent, and started with a demographic survey. While on Zoom, participants opened our custom page on their mobile phone. They were also asked to record their phone screen as they completed the tasks. Participants entered the prompted text into a text field, either through typing or speech recognition. They spent about 30 minutes on these tasks. After that, they sent us the screen recording of their phone and we proceeded to the semi-structured interview with 11 questions (Appendix A) which took between 10-15 minutes. We followed up naturally on interesting issues and encouraged participants to voice their opinions, too.

4 RESULTS

We used one-way ANOVA with $\alpha = 0.05$. The data was normally distributed and all other preconditions of ANOVA were met, too. We observed a significant effect on entry speeds in terms of words per minute (WPM), $F(1,195) = 343$, $p < .001$, with a large effect size $\eta_p^2 = .64$, and speech recognition being fastest, see Figure 1. Error rate was also significantly different, $F(1,195) = 14.80$, $p < .001$, with a medium effect size $\eta_p^2 = .07$. Speech recognition exhibited more errors, see Figure 1. The difference in the number of operations per character for each condition was significant, $F(1,195) = 36.24$, $p < .001$, with a large effect size $\eta_p^2 = .16$. Speech recognition needed fewer operations per character on average, see Figure 1. The rate of backspaces (either during typing or later corrections) was also significantly different, $F(1,195) = 26.97$, $p < .001$, with medium effect size $\eta_p^2 = .12$. Speech recognition had fewer error corrections, see Figure 1. We also measured verification time, which is the time participants took to review a phrase before moving to the next, but did not observe a significant effect between the conditions, $F(1,195) = .77$, $p = .38$.

4.1 Participant Interviews

During the semi-structured interview, participants were asked about their thoughts on autocorrect during typing and for speech recognition, for the transcription tasks and their daily natural mobile phone use. Most preferred text entry over speech recognition. One participant even remarked: "*Text is superior to speech,*" because it is "*not as disruptive.*" On the other hand, participants acknowledged the higher speed and accuracy of speech recognition in the experiment, sparking valuable discussion. Speech recognition enabled participants to enter text more quickly than with typing. Some even perceived speech recognition to be highly accurate during the task. Yet, when recalling previous experiences with speech recognition, it was scrutinized for its imperfections. If the system presented an inaccurate word (for whatever reason) a user has two choices: even though temporarily distracted, they can finish their sentence and make edits afterwards. Or, they can immediately make the correction through typing. In this case they might continue typing the rest of the phrase (to avoid another mode switch). Neither option was perceived as attractive by participants, as in both cases, the text entry process is disrupted, resulting in lost time and heightened frustration.

Without prompting, many users brought up using speech for home assistants. While a different domain, we identified parallels that could also improve speech-based text entry. In particular, words that are not in an English lexicon, such as an international name, can be problematic. One participant stated: "*[the device] does not recognize some types of names. My name, for example, is "name" but if I tell them to "Call name" it will not recognize [that]. I have to spell: "Call (proceeds to spell name).*"" Another participant identified the disruptive effect of errors: "*I was not that frustrated for the text input one. But [] the voice one made me a bit more frustrated. Because when there was like an error, I would have to switch from the dictation back to the keyboard to delete stuff. And then if there was an error, that was like the first three words in a ten-word sentence, I would have to delete seven or like eight words if I needed, right?*" Without an easy editing option, speech recognition seems tedious to most users, even if it was fairly accurate. When asked how frequently they used speech recognition for text entry, one participant responded: "*Not often... I only used it like 5 times. [] My main language is Chinese, so sometimes I speak too fast and it won't recognize it. And there will always be a typo.*"

Many remarked the choice of modality depends on location. While usable when driving a car, speech is not always perceived as useful. For low-stakes communication, speech recognition mistakes can even be humorous: "*There was one time where I was driving across the bridge []. And then I was texting my sister [], but Siri converted some of the words into like profanity. And it was all fine, she had a little laugh about it. And [my sister] was like: "I understood what you meant, but that was kind of funny.*"" Yet, sometimes mistakes need to be avoided, e.g., when communicating with our bosses. Unless input can easily be corrected, speech-based text entry will thus most likely not be used for important matters. As one participant stated: "*I don't trust my speech to do it.*"

Privacy concerns for speech recognition were mentioned by almost every participant, along with noise level issues. "*If I'm on my own in my room, or whatever, then maybe that is OK. But usually I'm around my family, so they don't need to hear everything I am texting.*" One participant stated that "*if I randomly say something that sounds similar to "Hey Google," [Google Home mini] will say, actually: "What do you need?" like... "I don't need you!".*" Thus, speech recognition can be disconcerting to use, e.g., with inadvertent activation or if text that was not meant to be written down suddenly appears. And speech recognition is not always practical for its intended, hands-free use. "*One problem is [...] if I tell them to stop, or change to the next song, I have to speak really loud, over the sound of the song, and I will end up even yelling in the next room. It still won't work, so I have to turn (the music) off on my phone.*"

Different usage approaches determine the user experience. One participant remarked: "*For speech recognition, the user has to be accurate for the result to be accurate. If you're not clear, the sentence will be all jumbled up.*" Another participant noted: "*You have to pronounce each word accurately. [] if your accent is [] thick, speech recognition won't be able to identify it.*" Also, one user stated: "*If you hesitate during speech recognition, or if you want to think out loud, the sentence you are trying to create will not appear as intended. There comes a point of no return if you stumble over a word or if you realize you mispronounced something. [] Once you make a mistake,*

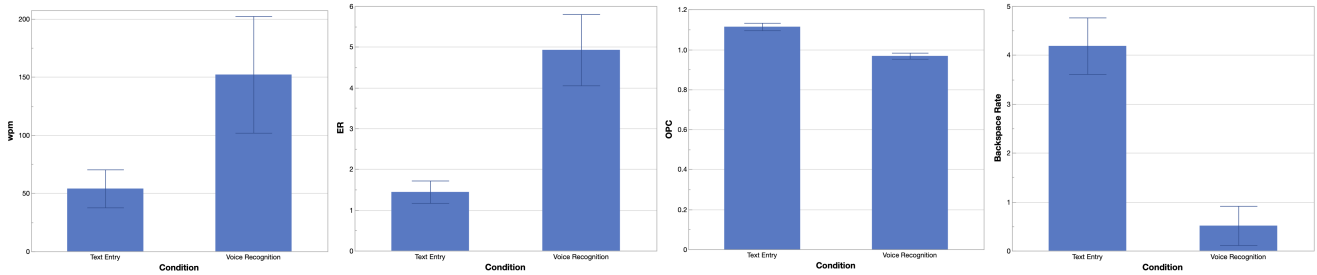

Figure 1: The average WPM, Error Rate, Operations per Character, and backspace rate for each condition and their standard deviation.

you either have to start over and discard the entire sentence, or send the sentence with a mistake."

5 DISCUSSION

We confirm that speech recognition resulted in a significantly faster input than typing, albeit at the cost of a higher error rate. Participants who were already in the process of typing corrected errors naturally in the transcription task. In contrast, with speech recognition, they carefully controlled what they said, as going back later to edit via typing was not generally perceived as desirable. Our findings from the interviews lend insights into the user's experience and perceptions of text entry methods. Many were satisfied with their experience during the experiment with speech recognition and reflected on their perceptions of it. Yet, most stated also that it was an imperfect system and did not use it regularly in their lives. Criticism focused on privacy, such as not wanting others to overhear the text, or feeling insecure about their speech. This was perceived to be worse than the frustration associated with autocorrect. Some suggestions for better interfaces for speech recognition arose as one participant wanted a "*better option for different accents*" while another suggested auto-complete for speech recognition. Thus, future work should look at improved speech interfaces, preferably with hands-free interaction, since most users engage with speech recognition this way.

For autocorrect, participants mentioned their frustration, but in the same breath would say they leave it on and use it whenever they are typing. Though there are mistakes and frustration had run high in the past, the only comment about turning autocorrect off came from a participant who constantly experienced cross-multi-language autocorrects.

6 CONCLUSION AND FUTURE WORK

Autocorrect and speech recognition aim to improve the users' experience and efficiency when entering text on mobile devices. Our initial study confirmed that speech recognition is faster than typing for text entry on mobile devices, but with an higher error rate. Our semi-structured interviews resulted in insights on common issues or concerns users were experiencing, both when typing and using speech recognition, as well as some design suggestions for speech-based text entry. Our work brings up the question what a seamless transition between speech recognition and then editing could be, both for hands-free and device-based contexts.

REFERENCES

[1] n.d.. Global Rationale for Choice of Voice Assistants over Websites and Apps 2017. https://www.statista.com/statistics/801980/worldwide-preference-voice-assistant-websites-app Accessed July 1, 2021.

[2] Ohoud Alharbi, Wolfgang Stuerzlinger, and Felix Putze. 2020. The Effects of Predictive Features of Mobile Keyboards on Text Entry Speed and Errors. *Proceedings of the ACM on Human-Computer Interaction* 4, ISS (2020), 1–16. https://doi.org/10.1145/3427311

[3] M. A. Anusuya and S. K. Katti. 2010. Speech Recognition by Machine, A Review. http://arxiv.org/abs/1001.2267

[4] Patrick Armstrong and Brett Wilkinson. 2016. Text Entry of Physical and Virtual Keyboards on Tablets and the User Perception. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction*. 401–405. https://doi.org/10.1145/3010915.3011006

[5] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 603–608. https://doi.org/10.1145/2984511.2984584

[6] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3173574.3173829

[7] Andy Cockburn, Blaine Lewis, Philip Quinn, and Carl Gutwin. 2020. Framing Effects Influence Interface Feature Decisions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11. http://doi.org/10.1145/3313831.3376496

[8] 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*. 1–11. https://doi.org/10.1145/3313831.3376861

[9] Santosh K. Gaikwad, Bharti W. Gawali, and Pravin Yannawar. 2010. A Review on Speech Recognition Technique. *International Journal of Computer Applications* 10, 3 (2010), 16–24. https://doi.org/10.5120/1462-1976

[10] Dylan Gaines, Per Ola Kristensson, and Keith Vertanen. 2021. Enhancing the Composition Task in Text Entry Studies: Eliciting Difficult Text and Improving Error Rate Calculation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8. https://doi.org/10.1145/3411764.3445199

[11] Shiyoh Goetsu and Tetsuya Sakai. 2019. Voice Input Interface Failures and Frustration: Developer and User Perspectives. In *The Adjunct Publication of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 24–26. https://doi.org/10.1145/3332167.3357103

[12] Horstmann Koester and Heidi. 2004. Usage, Performance, and Satisfaction Outcomes for Experienced Users of Automatic Speech Recognition. *The Journal of Rehabilitation Research and De-velopment* 41, 5 (2004), 739–754. https://doi.org/10.1682/JRRD.2003.07.0106

[13] Per Ola Kristensson and Thomas Müllners. 2021. Design and Analysis of Intelligent Text Entry Systems with Function Structure Models and Envelope Analysis. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3411764.3445566

[14] Graeme McLean and Kofi Osei-Frimpong. 2019. Hey Alexa ... Examine the Variables Influencing the Use of Artificial Intelligent In-Home Voice Assistants. *Computers in Human Behavior* 99 (2019), 28–37. https://doi.org/10.1016/j.chb.2019.05.009

[15] Debajyoti Pal, Chonlameth Arpnikanondt, Suree Funilkul, and Wichian Chutimaskul. 2020. The Adoption Analysis of Voice-Based Smart IoT Products. *IEEE Internet of Things Journal* 7, 11 (2020), 10852–67. https://doi.org/10.1109/JIOT.2020.2991791

[16] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How Do People Type on Mobile Devices?: Observations from a Study with 37,000 Volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12. https://doi.org/10.1145/3338286.3340120

[17] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of Speech and Silent Speech Input Methods in Private and Public. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3411764.3445430

[18] Shyam Reyal, Shumin Zhai, and Per Ola Kristensson. 2015. Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 679–88. https://doi.org/10.1145/2702123.2702597

[19] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23. https://doi.org/10.1145/3161187

[20] Korok Sengupta, Sabin Bhattarai, Sayan Sarcar, I. Scott MacKenzie, and Steffen Staab. 2020. Leveraging Error Correction in Voice-Based Text Entry by Talk-and-Gaze. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3313831.3376579

[21] R. William Soukoreff and I. Scott MacKenzie. 2001. 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*. 319–320. https://doi.org/10.1145/634067.634256

[22] Yuntao Wang, Ao Yu, Xin Yi, Yuanwei Zhang, Ishan Chatter-jee, Shwetak Patel, and Yuanchun Shi. 2021. Facilitating Text Entry on Smartphones with QWERTY Keyboard for Users with Parkinson's Disease. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3411764.3445352

[23] Nicola Wood. 2014. Autocorrect Awareness: Categorizing Autocorrect Changes and Measuring Authorial Perceptions. https://fsu.digital.flvc.org/islandora/object/fsu

[24] Dilawar Shah Zwakman, Debajyoti Pal, and Chonlameth Arpnikanondt. 2021. Usability Evaluation of Artificial Intelligence-Based Voice Assistants: The Case of Amazon Alexa. *SN Computer Science* 2, 1 (2021), 28. https://doi.org/10.1007/s42979-020-00424-4

A INTERVIEW QUESTIONS

(1) How often do you rely on autocorrect?
(2) What about voice recognition, how often do you rely on it?
(3) How well do you think the autocorrect and speech recognition works in smartphones?
(4) Does autocorrect influence the speed and correctness of text entry on mobile devices? What is the effect and why?
(5) Does speech recognition influence the speed and correctness of text entry on mobile devices? What is the effect and why?
(6) If there were any mis-predictions, do you think they affected your task?
(7) How frustrated do you think you got entering the text, all things considered?
(8) How satisfied do you think you were with entering the text, all things considered?
(9) What kinds of mistakes were the hardest to correct?
(10) Do you have any design recommendation/suggestion regarding auto-correction or voice recognition?
(11) Please share any stories with us about when you or someone you know encountered autocorrect or voice recognition.