The Effects of Errors in Speech Transcription: A User Study

TIANYI (VERA) BAO*, SIAT, Simon Fraser University, Vancouver, Canada, vera_bao@sfu.ca

AFSHAN AHMED, R.V. College of Engineering, Bangalore, India, afshanahmed.ec18@rvce.edu.in

WOLFGANG STUERZLINGER, SIAT, Simon Fraser University, Vancouver, Canada, w.s@sfu.ca

This paper investigates how errors that occur during speech recognition affect users' text entry performance. To study this, we implemented a speech recognition system that injects believable errors in a controlled manner. In our user study, participants were asked to transcribe a set of phrases using our speech recognition system, either with or without the insertion of errors. The results show that inducing 33% errors in a speech-based transcription task does not seem to affect users' performance and experience in a significant manner. Yet, according to participants' interview responses, our result might have been caused by the phrase set we used in the study. Our work thus motivates future research to develop a phrase set more suitable for speech-based transcription tasks.

CCS CONCEPTS • User studies • Human-centered computing

Additional Keywords and Phrases: speech recognition; text entry; error detection

ACM Reference Format:

Tianyi (Vera) Bao, Afshan Ahmed, and Wolfgang Stuerzlinger. 2022. The Effects of Errors in Speech Transcription: A User Study. In *TEXT2030: MobileHCI'22 Workshop on Shaping Text Entry Research in 2030*, October 1, 2022, Vancouver, Canada. 6 pages.

1 INTRODUCTION

In this digital age text entry efficiency is crucial for our everyday communication. State-of-the-art technologies that are commonly used to improve entry efficiency include auto-correction, word prediction, and voice-to-text input. Yet, studies have revealed that autocorrection and predictive features rarely increase text entry speed significantly, due to the time required to manually fix wrong predictions or corrections and/or the higher cognitive load required to fix such errors [2,3,11,14]. According to some studies, when such errors occur, users experience also an increase in frustration and physical and mental workload [2,5,14].

Beyond predictive features and autocorrection, and based on advances in speech recognition technology, voice-to-text input has become another widely used modality for text entry [6,19]. Ruan et al. discovered that transcribing short phrases with speech recognition can be almost 200% faster than typing on a touch-based smartphone [18]. Yet, there have been no studies that investigate how errors in such systems affect users' text entry performance and frustration.

2 RELATED WORK

Error correction plays a critical role in text entry. Advances in error correction algorithms have enabled improvements in the efficiency of typing-based text entry and the users' experience [16,20,26]. Still, voice-based text input can afford much higher text entry efficiency [4]. However, correcting errors via voice editing is more challenging than via typing, due to the linear and temporal nature of audio [7], which increases the mental and physical burden of the users [9].

TEXT2030, October 1, 2022, Vancouver, Canada

One major challenge in dictation tasks is specifying the location of the incorrect (part of the) phrase [10,13]. Users then typically resort to manually editing the dictation results at least 60% of the time [4,22]. To address this issue, McNair and Weibel [27] proposed an effective technique that only requires respeaking a phrase of a sentence to locate and correct it. With this, users can then simply respeak only the erroneous part of the sentence, e.g., correcting "I will **bang** it over tomorrow" by saying only "**bring**", instead of respeaking the full sentence. This technique was further improved by Ghosh et al. [9, appendix available in the ACM Library] by letting users speak a few additional words. In the above example, users could then say "**bring** it over" to correct the text. This method provides more context for the system – to better match the incorrect phrase – making the matching process more accurate and lets users speak natural phrases. We used this re-speaking technique in our evaluation, as participants are then able to edit the text using only verbal input.

Relative to text transcription, text composition is more common in real-world scenarios. Although the composition task has higher external validity than the transcription task, the latter outperforms the former by its higher internal validity and lower variability [25]. A recent study [8] let participants compose their own phrases, and used them for transcription tasks with other participants, which increased the logistic effort substantially. Most published text entry studies employed transcription to evaluate text entry efficiency. Yet, the most widely used phrase sets in such studies were all designed for typing, not for speaking [12,23,24]. According to Foley et al. [8], phrases for transcription tasks have to fulfill the following characteristics: memorable (users can enter a phrase after the prompt without referring to it), representative (resemblance of the actual text that is entered by people), and replicable (the phrase set is publicly available).

Spoken and written language also contrasts in various aspects. Spoken language is less abstract, has more finite verbs, and has fewer nouns of abstraction. There is also a contrast in syntax and sentence structure, and in terms of the manner and speed of production [1]. Moreover, entering text by tying versus speaking can lead to very different experiences for users [7,18]. Neuroscience research also found that written and spoken language involves two *distinct* systems that are controlled by different parts of the brain [15]. Therefore, phrase sets for spoken and written language might not be interchangeable.

3 USER STUDY

Twelve participants (six females, six males), aged between 21-29 years old, with an average of 24.5 (SD = 2.15) participated in the study. All participants were either completing or had completed a bachelor's degree in an English-speaking university in Canada. All data were collected over Zoom with participants sharing their screen, except for two who did not agree to share the screen.

The experiment used a web application housed on a local university server. We implemented the system using JavaScript, and PHP, building on the Google speech recognition API [28]. Normally, such a system would show the most likely recognition result for each user's utterance, but we sometimes injected errors by showing the second-most likely result returned by the Google API, which effectively generates a very believable misrecognition. For example, instead of showing "How was your trip to Florida?" (the most likely result), our system displays sometimes "How was your train to Florida?" (the second-most likely result). In the experimental conditions, we injected such an error either 0% or 33% of the time. We chose 33% to avoid inducing excessive frustration.

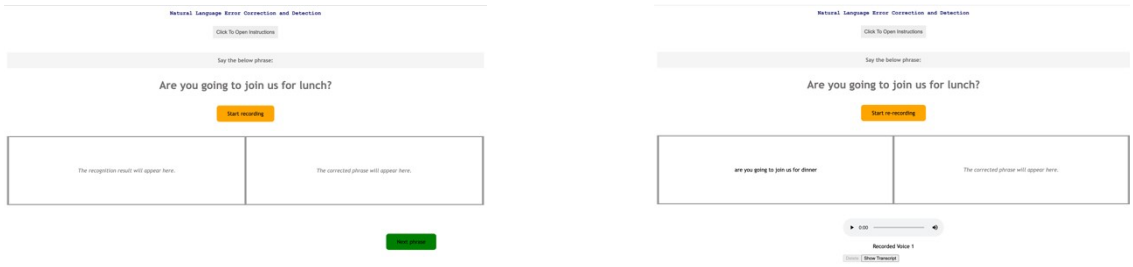
Our study used a between-subject design with the injected error rate (two levels, 0%, and 33%) as the independent factor. The dependent factors included entry speed (WPM), the (remaining) error rate (ER), as well as self-reported

frustration, physical demand, and mental demand (NASA TLX). Participants were randomly assigned to one of the 0% error and 33% error conditions. We collected 29 phrases for each participant, for a total of 348 phrases.

All phrases were randomly selected from the Enron MobileEmail phrase set [17]. We removed all punctuation marks, as they might introduce a confound in the dependent variables, which might undermine the internal validity [3,21].

Initially, participants were allowed to choose the most appropriate accent among English-US, English-UK, English-India, and English-Canada. Participants were then asked to speak each phrase that appeared on the screen. Participants clicked on "Start Recording" to record their utterances (Figure 1a) and finished with the Stop button. If the speech was transcribed incorrectly, they then could repeat part of the sentence by clicking on the "Start re-recording" button on the same page (Figure 1b).

We asked participants to re-speak the incorrect phrase from *at least one word before* the incorrect word and ending *at least one word after* the incorrect word. If the correction involved the ending or starting word, then they repeated from two words before or after. Participants were only given a single error correction attempt for each phrase, after which they had to proceed to the next phrase. After they completed all 29 phrases, participants completed a NASA TLX 7-point Likert scale questionnaire. Each participant was asked to self-report their subjective experience, followed by a brief interview at the end to assess participants' familiarity with speech-to-text systems and their experience with the re-speaking interface.



(a) (b)

Figure 1. Speech recognition task example (a) first attempt; (b) second attempt.

4 RESULTS

Overall, error rates for participants in the 33% error condition (M = 29.9%, SD = 13.7%) were higher than with 0% errors (M = 23.6%, SD = 6.3%). However, there was no significant difference between the two conditions, $t(7.03) = 1.02$, $p = 0.34$. See also Figure 2a.

Overall, participants exhibited a higher WPM in the 0% error condition (M = 119.90, SD = 13.77) compared to 33% errors (M = 116.08, SD = 10.10). Figure 2b shows the mean WPM for each group. Yet, a Wilcoxon test revealed no significant difference for entry speed (Z = -1.04, p = .30).

In the 0% error condition (M = 1.50, SD = .55) participants reported lower physical demand than with 33% errors (M = 1.67, SD = 1.63), but the difference was not significantly different (Z = - .76, p = .44). Figure 2c shows the mean physical demand for each group.

Figure 3d shows the mean mental demand for each group. Overall, mental demand for participants in the 33% error condition (M = 3.50, SD = 1.60) was higher than with 0% (M = 2.50, SD = 1.52). However, the result was not significantly different, $t(0.74) = 1.02$, $p = 0.48$.

Overall, frustration for participants in 33% error condition (M = 3.17, SD = 1.72) was higher than in the 0% condition (M = 1.33, SD = .52). Figure 2e shows the mean frustration for each group. A Wilcoxon signed-rank tests between the 0% and 33% error conditions revealed no significant differences for frustration (Z = 1.95, p = .05).

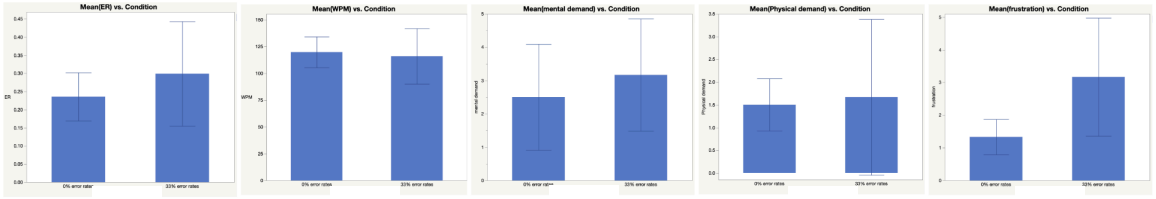


Figure 2: Mean performance metrics (±95% CIs) for each dependent measures: (a) error rate, (b) entry speed (WPM), (c) mental demand, (d) physical demand, e) frustration.

At the end of the experiment, we conducted a brief, semi-structured interview, focusing on participants' experience with our system. All participants reported at least some level of familiarity with speech-to-text tools. When asked about the experience with our system compared to other systems they had used before, participants had very diverse responses, regardless of which condition they experienced. For example, a participant in the control condition reported the system as "very dumb, needs more development", whereas another participant in the experimental condition said: "It has better accuracy, understood me better." When we asked them about the most challenging part of the study, seven out of twelve participants reported disliking the phrases we used in the study. Some said the phrases felt weird and unnatural, took them a long time to read, and they also made more mistakes. Others mentioned that some of the phrases were too long.

In addition, when we asked participants about their favourite part of the system, the most frequent answer was the correction feature. In general, participants reported that this feature was new to them, and they liked how they did not have to repeat the whole sentence to correct the recognition result.

5 DISCUSSION

One of the most striking findings from our study is that we failed to find a significant difference, even though we induced a non-trivial amount (33%) of errors. While using more participants may reveal significant results, our observed differences are small –only 3.25% difference in WPM. Yet, one of the key takeaways from our qualitative result was the apparent inappropriateness of the phrase set. Based on our participants' responses, we believe that the properties of the phrase set could be one of the most likely explanations for our results. As discussed in the Related Work section, the differences between written and spoken language are non-negligible, which likely contributed here. Also, given the high variability in text composition tasks, text transcription is more controllable and easier to study. However, the lack of available phrase sets designed specifically for speech-based transcription tasks makes such studies currently challenging.

6 CONCLUSION AND FUTURE DIRECTIONS

Although our study did not identify a significant effect for 33% induced errors, we can draw some valuable insights for how we could improve our work in the future. We are planning to develop a new phrase set that is more appropriate for speech transcription/dictation tasks. With such a phrase set, we plan to re-evaluate re-speaking interfaces for error correction.

REFERENCES

1. F. Niyi Akinnaso. 1982. On The Differences Between Spoken and Written Language. *Language and Speech* 25, 2: 97–125. https://doi.org/10.1177/002383098202500201
2. Ohoud Alharbi, Ahmed Arif, Wolfgang Stuerzlinger, Mark Dunlop, and Andreas Komninos. 2019. WiseType: A Tablet Keyboard with Color-Coded Visualization and Various Editing Options for Error Correction. *Proceedings of Graphics Interface 2019* Kingston: 10 pages, 423.17 KB. https://doi.org/10.20380/GI2019.04
3. 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: 1–16. https://doi.org/10.1145/3427311
4. Shiri Azenkot and Nicole B. Lee. 2013. Exploring the use of speech input by blind people on mobile devices. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (ASSETS '13), 1–8. https://doi.org/10.1145/2513383.2513440
5. Fei Cao, Jiaxi Zhang, Lei Song, Shoupeng Wang, Danmin Miao, and Jiaxi Peng. 2017. Framing Effect in the Trolley Problem and Footbridge Dilemma: Number of Saved Lives Matters. *Psychological Reports* 120, 1: 88–101. https://doi.org/10.1177/0033294116685866
6. G. E. Dahl, Dong Yu, Li Deng, and A. Acero. 2012. Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 20, 1: 30–42. https://doi.org/10.1109/TASL.2011.2134090
7. Jiayue Fan, Chenning Xu, Chun Yu, and Yuanchun Shi. 2021. Just Speak It: Minimize Cognitive Load for Eyes-Free Text Editing with a Smart Voice Assistant. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, 910–921. https://doi.org/10.1145/3472749.3474795
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. Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Di Chen, and Morten Fjeld. 2018. EDITalk: Towards Designing Eyes-free Interactions for Mobile Word Processing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–10. Retrieved December 15, 2021 from https://doi.org/10.1145/3173574.3173977
10. Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. 1999. 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), 568–575. https://doi.org/10.1145/302979.303160
11. H. H. Koester and S. P. Levine. 1994. Modeling the speed of text entry with a word prediction interface. *IEEE Transactions on Rehabilitation Engineering* 2, 3: 177–187. https://doi.org/10.1109/86.331567
12. I. Scott MacKenzie and R. William Soukoreff. 2003. Phrase sets for evaluating text entry techniques. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '03), 754–755. https://doi.org/10.1145/765891.765971
13. Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. 1997. Integration and synchronization of input modes during multimodal human-computer interaction. In *Referring Phenomena in a Multimedia Context and their Computational Treatment* (ReferringPhenomena '97), 1–13.
14. Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 83–88. https://doi.org/10.1145/2858036.2858305
15. Brenda Rapp, Simon Fischer-Baum, and Michele Miozzo. 2015. Modality and Morphology: What We Write May Not Be What We Say. *Psychological Science* 26, 6: 892–902. https://doi.org/10.1177/0956797615573520
16. 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*, 679–688. https://doi.org/10.1145/2702123.2702597
17. Paul Rozin and Edward B. Royzman. 2001. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review* 5, 4: 296–320. https://doi.org/10.1207/S15327957PSPR0504_2
18. 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: 1–23. https://doi.org/10.1145/3161187
19. Haşim Sak, Andrew Senior, and Françoise Beaufays. 2014. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *arXiv:1402.1128 [cs, stat]*. Retrieved December 14, 2021 from http://arxiv.org/abs/1402.1128

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* (CHI EA '01), 319–320. https://doi.org/10.1145/634067.634256
22. Bernhard Suhm and Alex Waibel. 1997. Exploiting repair context in interactive error recovery.
23. Keith Vertanen, Dylan Gaines, Crystal Fletcher, Alex M. Stanage, Robbie Watling, and Per Ola Kristensson. 2019. 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*, 1–14. https://doi.org/10.1145/3290605.3300821
24. Keith Vertanen and Per Ola Kristensson. 2011. 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*, 295. https://doi.org/10.1145/2037373.2037418
25. Keith Vertanen and Per Ola Kristensson. 2014. Complementing text entry evaluations with a composition task. *ACM Transactions on Computer-Human Interaction* 21, 2: 1–33. https://doi.org/10.1145/2555691
26. Yuntao Wang, Ao Yu, Xin Yi, Yuanwei Zhang, Ishan Chatterjee, 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
27. ICSLP 94 Abstract: McNair / Waibel. Retrieved April 13, 2022 from https://www.isca-speech.org/archive_v0/icslp_1994/i94_1299.html
28. Speech-to-Text: Automatic Speech Recognition. *Google Cloud*. Retrieved December 14, 2021 from https://cloud.google.com/speech-to-text

Author Bios:

Tianyi (Vera) Bao: Vera is currently a second-year MSc student at the School of Interactive Arts + Technology at Simon Fraser University in Vancouver.

Afshan Ahmed: Afshan is currently an undergraduate student in Engineering in Electronics and Communication at the Rashtreeya Vidyalaya College of Engineering in Bangalore, India.

Wolfgang Stuerzlinger: Building on his deep expertise in Virtual Reality and Human-Computer Interaction, Dr. Stuerzlinger is a leading researcher in Three-dimensional User Interfaces. He got his Doctorate from the Vienna University of Technology, was a postdoctoral researcher at the University of Chapel Hill in North Carolina, and professor at York University in Toronto. Since 2014, he is a full professor at the School of Interactive Arts + Technology at Simon Fraser University in Vancouver, Canada. His work aims to gain a deeper understanding of and to find innovative solutions for real-world problems. Current research projects include better 3D interaction techniques for Virtual and Augmented Reality applications, new human-in-the-loop systems for big data analysis (Visual Analytics and Immersive Analytics), the characterization of the effects of technology limitations on human performance, investigations of human behaviors with occasionally failing technologies, user interfaces for versions, scenarios, and alternatives, and new Virtual/Augmented Reality hardware and software.