Effects of WER on ASR Correction Interfaces for Mobile Text Entry

Christine Murad
University of Toronto, TAGlab,
Department of Computer Science,
Toronto, ON, Canada
cmurad@taglab.ca

Wolfgang Stuerzlinger
School of Interactive Arts + Technology (SIAT)
Simon Fraser University
Vancouver, BC, Canada
w.s@sfu.ca

Cosmin Munteanu
University of Toronto, TAGlab and
University of Toronto Mississauga, ICCIT
Toronto, ON, Canada
cosmin.munteanu@utoronto.ca

MobileHCI '19, October 1–4, 2019, Taipei, Taiwan

ACM ISBN 978-1-4503-6825-4/19/10…$15.00

Abstract
Speech is increasingly being used as a method for text entry, especially on commercial mobile devices such as smartphones. While automatic speech recognition has seen great advances, factors like acoustic noise, differences in language or accents can affect the accuracy of speech dictation for mobile text entry. There has been some research on interfaces that enable users to intervene in the process, by correcting speech recognition errors. However, there is currently little research that investigates the effect of Automatic Speech Recognition (ASR) metrics, such as word error rate, on human performance and usability of speech recognition correction interfaces for mobile devices. This research explores how word error rates affect the usability and usefulness of touch-based speech recognition correction interfaces in the context of mobile device text entry.

ACM Classification Keywords
H.5.2 [User interfaces]: Voice I/O, Natural language

Introduction
Using speech for mobile text entry has become more practical over the past decade. Most current mobile devices, such as smartphones and tablets, allow a user to enter text using speech, where an automatic speech recognition (ASR) system decodes the acoustic data and provides a transcription of text back to the user.

Figure 2 (a) Study Procedure (repeated for each statement / independent variable combination)

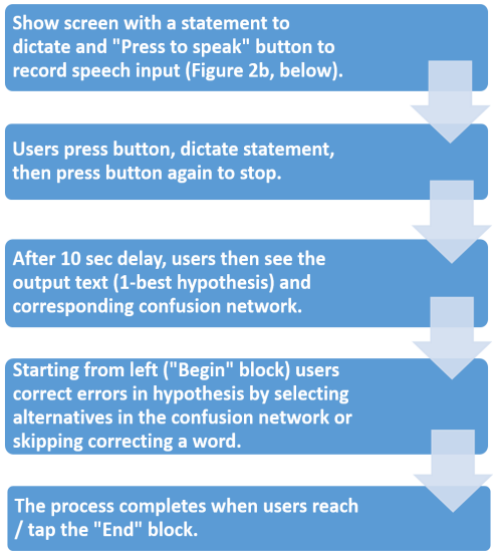
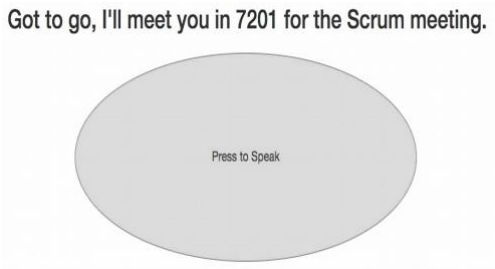


Figure 2 (b) GUI "Speak" Screen



While ASR has made many advances, errors in ASR systems are still frequent [3,7]. Variances in language, different accents, and the natural ambiguity of language are all factors that can cause ASR to be unable to deliver a perfect result, forcing users to use other means to correct the errors. The effects of speech recognition errors can greatly impact the viability of ASR as a text entry method [1,4]. As speech is often useful to quickly and more naturally enter text compared to using an on-screen keyboard [5], ASR failures can cause much frustration.

Due to these issues, a user must have some means to efficiently correct ASR errors. Interactive interfaces to correct ASR results can help mitigate accuracy errors in text entry [3,8]. Past research has explored different ways of helping users to correct ASR errors. Such interfaces to assist in ASR error correction are based on various approaches, such as re-speaking the incorrectly-recognized text [11], using a custom phonetic alphabet to input corrections [2] or more multimodal interfaces that provide alternative word candidates for users to select the correct transcription [3,6,8,10]. The goal of these interfaces is often to decrease the word error rate (WER) in final transcriptions of dictated text.

While much research has explored methods to reduce WER in ASR correction, little work has focused on
1) determining how easy it is for users to correct speech recognition errors in text input when the text is subject to ASR errors of varying WERs, and
2) measuring the influence of WERs (as a measure of ASR performance) on human performance when correcting text produced by ASR after user's speech input. Such research has been performed in different contexts, including webcast transcriptions [5].

In this paper, we conduct preliminary research on how WER affects human performance and acceptance when correcting ASR errors on mobile devices. We asked 10 participants to use a graphical interface that visualizes confusion networks to correct 3 different statements with multiple WERs. Visualizing confusion networks to aid speech-to-text correction has been proposed by other researchers, such as Ogata [6] and Vertanen [10]. While their work focused on how well such an interface may help the ASR system lower the WER, they did not extensively investigate how well users could use the interface and what factor WER plays in usability. Using confusion networks as a text correction aid, we collected data and observed how long it took to perform corrections, how many swipes/taps were performed, and the resulting WERs of the corrected sentences. We found that WER had a statistically significant effect on

Figure 1: Graphical Interface (Confusion Network Layout)

Our interface is a Node.js web app that works across mobile devices and OSes. This interface displays a confusion network that was generated from an ASR word lattice. Each word is displayed in a column, with the hypothesis outputted from the confusion network in the top row and other potential word candidates displayed within the same column, decreasing in posterior probability. One can tap or swipe the blocks across the interface in order to select alternative word candidates if there are ASR errors. Blank boxes allow users to skip a column.

Table 1: Statements and ASR metrics

Statement	Accent	Avg. confidence/ word	WER
1. "Got to go, I'll meet you in seven-two-oh-one for the Scrum meeting	American	0.865	0.28
	Scottish	0.850	0.44
2. "Hey, shop closes in 10, meet me at the back."	American	0.919	0.31
	Scottish	0.891	0.36
3. "I'm going study at Panera Bread, for a bit, catch you later."	American	0.828	0.20
	Scottish	0.742	0.52

Table 2: Ordering of conditions by sentence and accent

Ordering	Condition
1	Sentence 1, Scottish
2	Sentence 3, American
3	Sentence 2, American
4	Sentence 3, Scottish
5	Sentence 1, American
6	Sentence 2, Scottish

Table 3: Independent and Dependent Variables

Independent Variable	Dependent Variables
Word Error Rate of statement/ accent pair	Time required to complete correction
	Number of taps/swipes require to complete correction
	Word error rate post-correction

the effort required to correct sentences using this visualization. We also found that using an interface visualizing a confusion network to correct ASR errors (as in [6,10]) may be a challenging cognitive task. We argue that by analysis of the effort required to correct ASR errors through such visualizations, we can develop interfaces that leverage the accuracy of ASR systems and allow users to (more) easily correct recognition errors in mobile text dictation.

Study and Apparatus Setup

To give us access to the lattice and n-best hypotheses (as opposed to just a single output choice), we used the Kaldi speech recognition toolkit [12] for ASR. It also allowed us to construct confusion networks from the lattices. This was needed to be able to offer users choices for selecting possible alternative to correct the speech output. The statements used represented short text messages. These were generated to account for the exclusion of grammatical features (such as conjunctions) and out-of-vocabulary terminology that would not be common in regular speech and tend to typically throw off ASR.

We built a text correction interface that builds on the Parakeet system [7], as detailed in Figure 1. Our aim was not to propose a new interface, but to measure how human performance is affected by varying levels of speech recognition accuracy while correcting text using a confusion network visualization. In comparison, the original Parakeet work aimed to measure how much the ASR can self-improve its accuracy based on the error corrections carried out by users.

A confusion network [9] contains word "sausages", where each sausage represents alternatives considered by the ASR for each word (represented as columns in our interface). Although confusion networks may contain even more word candidates, we displayed only the first 8 candidates per word in each column. We made this choice based on available screen space and the previous literature [10].

Methods

Experimental Setup

We used a Quasi (offline) Wizard of Oz setup to explore the effect of different word error rates on human performance of speech recognition error correction. Experimental sessions were carried out with lattice and outputs that were generated offline. This information was *omitted* from the participant on-boarding briefing, resulting in a mild deceit setup. The study was run on a 10-inch tablet running Windows 10, with participants sitting for the entirety of the tasks. Participants were provided with plausible text messages as prompts and asked to dictate them. We used this procedure to be able to control (simulate) various WER levels. Participants were instructed to use the interface to correct the text to the closest result possible.

10 graduate computer science students, ages 22-28, were recruited to complete this study (7 male, 3 female). Their English proficiency varied from functionally professional to native.

Tasks and Measures

Figure 2 shows a description of the study procedure and the welcome ("begin") screen. The independent variable in this study was the *Word Error Rate* for each combination of statement & accent, yielding six tested conditions (see Table 2). The confidence scores within lattices and overall WER were controlled by manipulating various parameters in generating the audio (input) files from which text prompts were

Table 4: Average time required for correction

Statement	Accent	Avg. time to complete correction (ms)
1. "Got to go, I'll meet you in seven-two-oh-one for the Scrum meeting	American	21444
	Scottish	91323
2. "Hey, shop closes in 10, meet me at the back."	American	33530
	Scottish	23606
3. "I'm going study at Panera Bread, for a bit, catch you later."	American	28212
	Scottish	53100

Table 5: Average # of taps/swipes required for correction

Statement	Accent	Avg. # of taps/swipes to complete correction
1. "Got to go, I'll meet you in seven-two-oh-one for the Scrum meeting	American	16
	Scottish	27
2. "Hey, shop closes in 10, meet me at the back."	American	21
	Scottish	16
3. "I'm going study at Panera Bread, for a bit, catch you later."	American	15
	Scottish	24

Table 6: WER Post-Correction

Statement	Accent	WER post-correction
1. "Got to go, I'll meet you in seven-two-oh-one for the Scrum meeting	American	16
	Scottish	27
2. "Hey, shop closes in 10, meet me at the back."	American	21
	Scottish	16
3. "I'm going study at Panera Bread, for a bit, catch you later."	American	15
	Scottish	24

obtained, the most impactful manipulation being accent.

Three dependent variables were measured during the study (see Table 3). These measures are similar to those that have been collected in other usability studies of interfaces for ASR correction [3, 5, 7] T

The first is *time required to complete correction*. Speech dictation is often used as a quicker and more efficient means of entering text into a mobile device. Therefore, it was of interest to explore how word error rate affects how quickly someone is able to make a recognition correction using this interface. The interface recorded, in milliseconds, how long it took a participant from when they selected the "Begin" box (signifying the start of a correction) to when they selected the "End" box (signifying the completion of a correction).

The second is *number of taps/swipes required for correction*. Ideally, for an error correction to be quick and efficient, it should not require copious amounts of interaction to complete it. Therefore, we recorded the number of taps/swipes participants made for each error correction. A tap/swipe here is defined the action of touching the screen and lasts until a user lifts their hand from the screen. With this definition, both a tap and a swipe are recorded as the same type of action, as it was mainly of interest to explore how many times a participant had to lift their hand from the screen to complete an error correction.

The third is *word-error rate post-correction*. As mentioned earlier, WER is often used as a measure of how successful an ASR correction interface is. To explore how WER of automatic speech recognition affects the accuracy of the final result, we also recorded the final corrected sentence that users selected, along with its corresponding WER.

We also conducted informal interviews after the studies were completed, to qualitatively explore how participants felt about the layout of the hypothesis sentence and alternative word candidates, and the interaction experience of performing error corrections.

Results

Friedman's test was used to determine the statistical significance in the measured quantitative dependent variables between different statements and WERs. A post-hoc test of pairwise comparisons between statements was then performed where $p < 0.5$, with Bonferroni correction.

Time Required to Complete Error Correction

Table 4 displays the averages for time required to complete an error correction for each condition between all participants. A statistically significant difference was observed ($\chi 2 = 38.171$, $p < 0.05$, degrees of freedom = 5). A post-hoc comparison analysis, after Bonferroni correction, showed this difference existed between all 3 statements. It took, on average, longer to complete an error correction for confusion networks with higher word error rates for statements 1 and 3. Surprisingly, it took longer to correct <Statement 2, American>, which had a lower WER. This suggests that there may be more than just WER that affects these results.

Number of Taps/Swipes to Complete Error Correction

Table 5 displays the average number of taps/swipes performed for each statement across all participants. A statistically significant difference was observed ($\chi 2 = 19.406$, $p < 0.05$, degrees of freedom = 5). A post-hoc comparison analysis however, after Bonferroni

correction, showed this difference only existed between the WERs for <Statement 1, Scottish> and <Statement 3, American>. It took, on average, longer to complete an error correction for confusion networks with higher WERs for statements 1 and 3, except for <Statement 2, Scottish accent>, which has a higher WER. This again suggests that they may be other contributing factors other than WER.

Word-Error Rate Post-Correction

Table 6 displays the average post-correction WER for each statement between all participants. A statistically significant difference was observed ($\chi 2 = 20.597$, $p < 0.05$, degrees of freedom = 5). It took on average longer to complete an error correction for confusion networks with higher word error rates for statements 1 and 3, again except for <Statement 2, Scottish>. which had a higher WER. In this case, while differences exist between most pairs, none were conclusively strong. This again strengthens the evidence that there may be other factors affecting the time on top of WER.

Post-Study Qualitative Feedback

Through the qualitative data from post-study informal interviews, we identified that participants were initially overwhelmed with the size of the confusion networks, and with the amount of alternative word candidate options. Participants often forgot the sentence they had initially dictated after they had finally absorbed the confusion network that was presented to them.

We also found that participants preferred to tap on each box to complete a selection. Swipes were mostly only used when there was a contiguous string of words that were correctly recognized and a participant could swipe through that in a straight line. Participants commented that they felt that some word candidates were vertically too far away from each other to comfortable swipe from one block to the next in a subsequent column.

Discussion

While this was a preliminary study, these results suggest that WER does have an effect on how humans perform when correcting speech recognition errors on a visual interface. One major factor was the amount of word candidates presented to participants. Confusion networks with higher WERs often had lower lattice confidences, which increased the number of candidates presented. Participants often felt overwhelmed when presented with many word candidates, requiring time at the beginning to absorb all possible options.

Another major factor we observed was that WER itself is not the only factor that affects completing a correction using this interface, but also the position of an error within a sentence. When there was a continuous string of words that were correct, participants very quickly swiped through them, and took much less time to complete the correction. However, when there were multiple errors spread throughout a sentence, and when they were farther down in the confusion network (at lower probability) it took overall much longer to complete a correction. Another factor besides WER that impacted the length of correction completion was the density of the networks – particularly the fact that the density was not uniformly distributed for some sentences. This necessitated extra time to process word candidates - even before starting correction. These observations are also validated by our quantitative data, in which post-hoc analyses and counter-cases suggests that WER may not the only factor affecting the effort needed to make a correction.

Our results show that presenting many alternative word candidates to choose from can be a difficult cognitive task. The cognitive load required to complete the task increases as the WER and number of word candidates increase. There comes a point where an interface such as this becomes more frustrating and difficult to use especially compared to simply typing out the correction or the entire statement itself. While we limited the number of word candidates presented on the interface to 8 per word, participants still perceived this to be a large number to parse through and choose from.

Conclusion and Future Work

In this paper, we found that confusion networks with higher WERs (and in turn, lower confidences) had a statistically significant effect on human performance when correcting speech recognition errors. We also identified that these ASR metrics affect the cognitive load required to complete an error correction task.

As this is only preliminary research, there are many other factors to be explored. Future work will consider a larger sample size (allowing for more meaningful statistical inferences) and with additional data collected to support the development and validation of a rigorous formula of human performance under voice dictation/correction tasks (potentially similar to Fitt's law). Other factors such as in-depth English proficiency measures for the users and variation in the number of candidates shown to the user can also have an effect on correction effort needed, and future work will collect and explore such data. Future work will also explore the most useful range of WERs for confusion network interfaces to be helpful and usable for correcting ASR errors.

References

[1] Arnout R. H. Fischer, Kathleen J. Price, and Andrew Sears. 2005. Speech-Based Text Entry for Mobile Handheld Devices: An Analysis of Efficacy and Error Correction Techniques for Server-Based Solutions. *International Journal of Human-Computer Interaction* 19, 3: 279–304.

[2] Kazuki Fujiwara. 2016. Error Correction of Speech Recognition by Custom Phonetic Alphabet Input for Ultra-Small Devices. *Proc. of CHI EA '16*, 104–109.

[3] David Huggins-Daines and Alexander I. Rudnicky. 2008. Interactive ASR Error Correction for Touchscreen Devices. *Proc. of ACL-HLT '18*:17–19.

[4] Ying Liu. 2009. Will Input Style Affect Mandarin Short Messages in Mobile Device？: a Wizard of Oz Study.

[5] Cosmin Munteanu, Ronald Baecker, Gerald Penn, Elaine Toms, and David James. 2006. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. *Proc. CHI '06, ACM Press* c: 493.

[6] Jun Ogata and Masataka Goto. 2005. Speech repair: Quick error correction just by using selection operation for speech input interfaces. *Proc. of Interspeech '05*: 133–136.

[7] Ben Shneiderman. 2000. The limits of speech recognition. *Communications of the ACM* 43, 9: 63–65.

[8] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction* 8, 1: 60–98.

[9] Gökhan Tür, Jeremy Wright, Allen L. Gorin, Giuseppe Riccardi, and Dilek Z. Hakkani-Tür. 2002. Improving spoken language understanding using word confusion networks. *Proc. of Interspeech '02*.

[10] Keith Vertanen and Ola Kristensson. 2009. *Parakeet: A Continuous Speech Recognition System for Mobile Touch-Screen Devices*.

[11] Keith Vertanen and Per Ola Kristensson. 2009. Automatic selection of recognition errors by respeaking the intended text. *Proc. of ASRU '09*: 130–135.

[12] Kaldi ASR. Retrieved August 18, 2018 from http://kaldi-asr.org/.