Gaze Enhanced Speech Recognition for Truly Hands-Free and Efficient Text Input During HCI

Matheus Vieira Portela [*]
Universidade de Brasilia
Campus Universitario Darcy Ribeiro
Brasilia, Brazil
matheus.portela@aluno.unb.br

David Rozado [†]
ICT Centre, CSIRO
1 Technology Court Pullenvale QLD 4069
Brisbane, Australia
David.Rozado@csiro.au

ABSTRACT
The performance of current speech recognition algorithms is well below that of human speech recognition, with high number of misrecognized words in quiet environments and degrading even further in noisy ones. Therefore, hands-free interaction remains a deeply frustrating experience. In this work, we present an innovative form of correcting misrecognized words during a speech recognition task by using gaze tracking technology in a multimodal approach. We propose to employ the user's gaze to point at misrecognized words and select appropriate alternatives. We compare the performance of this multimodal approach with traditional modalities of correcting words: usage of mouse and keyboard and usage of voice alone. The results of the user study show that whereas the proposed system is not as fast as using mouse and keyboard for correction, gaze enhanced correction significantly outperforms voice alone correction and is preferred by the users, offering a truly hands-free means of interaction.

Author Keywords
Gaze tracking, gaze responsive systems, speech recognition, multimodal interaction

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces

INTRODUCTION
In computer science, automatic speech recognition (ASR) refers to the computational translation of spoken words into text. Using speech to create or edit documents offers the potential to be a faster and more natural way to interact with computers as well as a hands-free modality of Human Computer Interaction (HCI) with obvious positive implications for handicapped users or scenarios where computer users have their hands engaged in other tasks, for example surgeons in the operating room.

Although there has been significant increases in the accuracy of ASR, error rates still make the technology cumbersome to use for everyday interaction. Previous works have shown that several factors increase the error rates in conversational speech: infrequent words, very fast or very slow speech, and long words among others [4].

The correction of ASR errors or misrecognized words can be carried out manually with keyboard and mouse by retyping it, which is probably the most widely used method for the task. This modality can be perfectly valid for some scenarios, but it is not truly hands-free anymore.

The usage of voice for correction maintains the notion of exclusively hands-free input to the computer. In practice however, this modality can be very frustrating for the user since "certain" challenging words are extremely difficult to be properly recognized by ASR engines and requires the user to pronounce the same word multiple times until it appears in a list of similar alternatives. Moreover, repeating several times the same word makes the vocal cords go through the same pattern of folding and vibrations repeatedly which has been shown to cause voice strain [3].

In this work, we propose the enhancement of ASR with gaze tracking technology to speed up the correction of misrecognized words and to maintain speech interaction truly hands-free. A gaze tracking system tracks the point of regard (PoR) of the user on the screen by monitoring the users pupils while sitting in front of a computer [8]. With the proposed modality, the user is required to simply gaze at the misrecognized word and then select the correct word from an emerging panel of most likely alternative words just by looking at it.

The experimental part of this work compares the three aforementioned modalities of correcting misrecognized words during a speech recognition task: usage of the traditional keyboard and mouse, usage of voice alone and usage of gaze.

[*]Scholar of CNPq - Brazil
[†]Postdoctoral Fellow

OZCHI'14, December 02 - 05 2014, Sydney, NSW, Australia Copyright 2014 ACM 978-1-4503-0653-9/14/1215.00
http : *//dx.doi.org/*10.1145/2686612.2686679

To our knowledge, the idea of using gaze to correct misrecognized words in an ASR task has not been explored before in the research literature. There exists however work on gaze aware and multimodal systems [7], such as advanced display of text [1], gaze integration in first person shooter games [5], gaze-based interface for browsing and searching images [6], and Dasher, an innovative dynamic text input system controlled by gaze [11]. Multimodal interfaces in literature include a calendar system integrating speech, gesture and handwriting recognition systems[10] and the use of speech- and gesture-based systems to transform a single-user interaction into a multi-user one [9].

METHODOLOGY

The goal of the user study was to compare the correction of misrecognized words using the gaze-based correction method with the voice and mouse and keyboard correction methods in a previously not trained ASR system. For this purpose, a graphical user interface (GUI) was designed and used in tests for each correction modalities. The video at `http://www.youtube.com/watch?v=xdBoNsMthr8` provides a good overview of the experimental setup and the different correction modalities being compared in the user study. We encourage the interested reader to watch it in order to gain a good understanding of the work presented here.

Participants

Nineteen participants took part in the user study: 17 male and 2 female. Among them, there were 10 native English speakers and 9 non-native English speakers. This is mentioned to account for the fact that ASR performance varies significantly between native and non-native speakers.

Apparatus

The experiment used a GUI interface written in Python 2.6 using Qt 4.8.4 Framework, by Digia, and PyQt 4.9.6 bindings. The ASR system incorporated in the interface was provided by the Microsoft Speech Application Programming Interface (SAPI), using version 5.1 of the SDK, in Windows 7 operating system. A Tobii X1 Gaze Tracker was used together with Tobii SDK 3.0 RC1 for Windows.

Experimental task

Each participant was requested to dictate 10 sentences within an allotted time of 60 seconds per sentence. This task was repeated for each correction modality: gaze, voice and mouse and keyboard, making a total of 30 dictated sentences for each subject. The sentences and the correction modality were randomly shuffled between experimental trials in order to smooth out ordering effects on the correction modalities performance. Prior to the beginning of each experimental trial, we ran a few test trials of each correction modality to make the subject comfortable with the interface. Also, a new empty Windows speech recognition profile was created for each subject in order to avoid the previous trials to interfere in the recognition of the sentences.

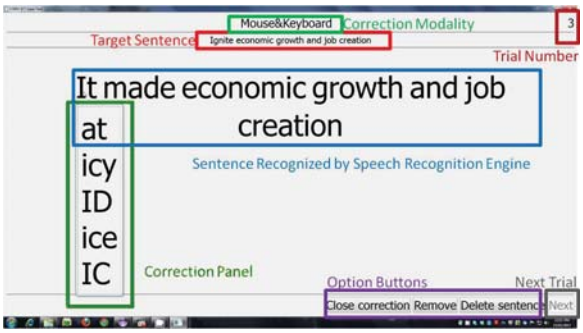

Figure 1. GUI used in the experiments.

The user interface, presented in Figure 1, shows correction modality, target sentence and trial number, on the top of the screen. When speech was recognized by the ASR system, words would be presented on the center of the screen. These words could be corrected in three different ways, determined by the correction modality:

- Mouse and keyboard: When clicking on a word with the mouse, a correction panel would pop up showing a list of at most 5 alternative words, which could also be selected by clicking. When the desired word was not presented as an alternative, clicking again on the word would generate a line edit and allow the subject to type the desired word with the keyboard.

- Voice: A correction panel was raised by saying the command "correct" followed by the desired word to correct. In this mode, alternative words presented in the correction panel are preceded by a number that, when pronounced, selected the corresponding word. When the desired word was not presented in the correction panel, pronouncing it again would refresh the menu with a new list of alternatives.

- Gaze: Fixating the gaze on a word for a dwell time of 2 seconds would pop-up a correction panel with the list of at most 5 alternative words. The alternative word could be selected by fixating the gaze for 2 seconds over it. Pronouncing the word again would refresh the menu with new alternatives.

On the bottom of the screen, three option buttons were available when the correction panel was opened: "Close correction", to close the correction panel, "Remove", to delete the word, and "Delete sentence", to delete the entire sentence. All buttons could be interacted with by using the same interface modality as the one being evaluated. When the user had correctly pronounced the target sentence, the "Next" button and the recognized sentence would become green to indicate that the user could go to the next trial. However, if the target sentence was not reached in less than 60 seconds, the "Next" button and the recognized sentence would become red,

indicating that the user has failed to correct the sentence in suitable time and could proceed to the next trial.

At the end of the experiments, the subjects involved in the user study were required to fill in a questionnaire about their subjective experience with the different correction modalities.

Measurements

During the experiment, time to complete the task was measured, which would stop when the active sentence being uttered matched the target sentence or if the user failed to achieve the target sentence within the allotted 60 seconds threshold. Each sentence also was marked indicating whether the trial has failed or not.

Furthermore, both the pronounced and the target sentence were recorded, data used later to calculate the Damerau-Levenshtein distance [2] between them. This distance reveals how far away two strings are from each other considering four basic operations: insertion, deletion, substitution of a single character, and transposition of two adjacent characters. This indicates how wrong the pronounced final sentence was in relation to the target sentence. Hence, small Damerau-Levenshtein distances means the recognized sentence is closer the target sentence, suggesting an interaction modality with less errors.

Lastly, the user questionnaire was composed by the following questions, where the subjects were able to answer either "Keyboard and mouse", "Voice" or "Gaze".

- Which method do you find the fastest to correct misrecognized words during speech recognition?
- Which method do you find the least error prone to correct misrecognized words during speech recognition?
- Which method do you find the most fatiguing to get the job done?
- Which method would you prefer to use to correct misrecognized words during speech recognition?
- If you could not use your hands during HCI, which method would you prefer to use to correct misrecognized words during speech recognition?

RESULTS

The average time of each experimental trial to achieve the target sentence using a given correction modality is displayed in Figure 2. A Levenes' test for equal variance for the 3 correction modalities failed ($p = 2.1$). Hence, the results of the ANOVA analysis need to be interpreted with caution. The F-test produced a value of $F(2, 54) = 17.43$, $p < 0.001$. A Posthoc Bonferroni-Holm test indicated significant differences between the voice-mouse and keyboard, gaze-mouse and keyboard, and gaze-voice modalities with $p < 0.001$ for the first two modalities, and $p = 0.02$ for the last one.

Figure 3 shows the average number of trials in which the user was unable to reach the target sentence in the allotted time of 60 seconds using the given correction modality. The Levene's test also failed to determine equal variance. The ANOVA analysis showed statistically significant differences between the results of the different correction modalities, $F(2, 54) = 17.92$, $p < 0.001$. A Posthoc Bonferroni-Holm test found significant differences between the voice-mouse and keyboard modalities, $p < 0.001$, gaze-mouse and keyboard modalities $p < 0.001$, and gaze-voice modalities $p = 0.004$.

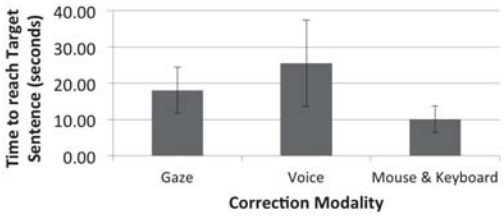

Figure 2. Average time required to achieve target sentence.

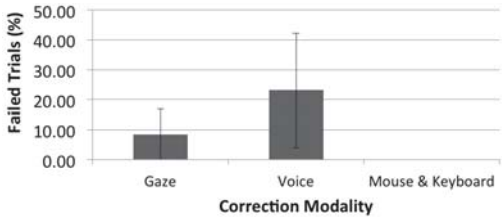

Figure 3. Percentage of trials where the user was unable to reach the target sentence within 60 seconds.

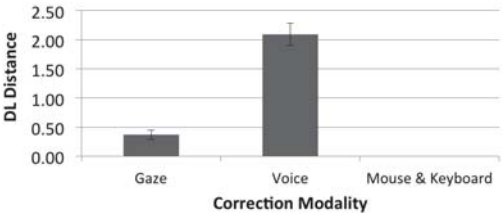

Figure 4. Average Damerau-Levenshtein distance between target and recognized sentences.

The average Damerau-Levenshtein distance between the target sentence and the recognized sentence is shown in Figure 4. The ANOVA analysis generated statistically significant differences between modalities with values $F(2, 54) = 11.60$, $p < 0.001$. A Posthoc Bonferroni-Holm test found significant differences between the voice-mouse and keyboard modalities, $p < 0.001$, gaze-voice modalities $p = 0.0044$, and gaze-mouse and keyboard modalities $p = 0.0171$.

Subjects involved in the user study expressed their subjective impressions about the different correction modalities being compared through a user questionnaire, the results of which are visible in Figure 5.

DISCUSSION

The results of the user study showed that the gaze enhanced correction modality for an ASR task is not as fast as using mouse and keyboard for correction. Yet,

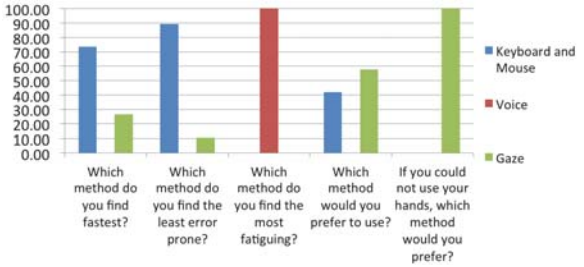

Figure 5. Subjective opinions expressed by the subjects on their perceptions of the different correction modalities.

the gaze based correction modality is significantly faster than using voice alone and it is truly hand-free as opposed to the mouse and keyboard modality. The time performance of gaze enhanced speech recognition can be improved in future research by implementing more sophisticated selection methods, replacing the 2 seconds fixation used in this work by dynamic and continuous selection of words, for instance.

The analysis of both the number of failures and the Damerau-Levenshtein distance for each modality reveals that mouse and keyboard provides the best HCI. However, when comparing the two hands-free modalities, gaze correction significantly outperforms voice correction, strongly indicating an easier to use interface that yields more accurate sentences.

It is important to emphasize that the most of the subjects involved in the user study had never been exposed to gaze trackers before, hence, they did not have time to properly familiarize themselves with the technology. Given enough time, learning effects would most likely improve the performance of gaze-based correction.

Moreover, algorithms that would respond to gaze behavior in a context aware manner could aid in disambiguating where the user is intending to point to. This could be done by opening the correction panel in the word nearest to the gaze position with the least amount of confidence in the recognition results. Innovative dynamic displays of alternative words could also help in this regard.

CONCLUSION

The gaze modality for correction of misrecognized words is not as efficient in terms of accuracy and time to completion as the traditional mouse and keyboard modality but it possesses the advantage of being truly hand-free. Furthermore, the gaze modality significantly outperforms the other hand-free modality to correct misrecognized words, using voice, in all measured variables. The gaze based correction modality also prevents the appearance of voice strain for correction of words since it prevents considerably the amount of utterances required to correct a word.

In light of the evidence presented here, we assert the advantages of the proposed multimodal approach to HCI that complements ASR with gaze interaction to create a multimodal interface for speech recognition tasks that is faster and more accurate than using voice alone to correct misrecognized words while remaining a truly hands-free form of interaction.

REFERENCES

1. Biedert, R., Buscher, G., Schwarz, S., Hees, J., and Dengel, A. Text 2.0. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, ACM Press (New York, New York, USA, Apr. 2010), 4003.
2. Damerau, F. J. A technique for computer detection and correction of spelling errors. *Commun. ACM 7*, 3 (Mar. 1964), 171–176.
3. Fritzell, B. Voice disorders and occupations. *Logopedics Phoniatrics Vocology 21*, 1 (1996), 7–12.
4. Goldwater, S., Jurafsky, D., and Manning, C. D. Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication 52*, 3 (2010), 181 – 200.
5. Koesling, H., Kenny, A., Finke, A., Ritter, H., McLoone, S., and Ward, T. Towards intelligent user interfaces. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications - NGCA '11*, ACM Press (New York, New York, USA, May 2011), 1–8.
6. Kozma, L., Klami, A., and Kaski, S. GaZIR. In *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*, ACM Press (New York, New York, USA, Nov. 2009), 305.
7. Rozado, D., Rodriguez, F. B., and Varona, P. Gaze Gesture Recognition with Hierarchical Temporal Memory Networks. In *Advances in Computational Intelligence*, J. Cabestany, I. Rojas, and G. Joya, Eds., vol. 6691 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011, 1–8.
8. Rozado, D., Rodriguez, F. B., and Varona, P. Low cost remote gaze gesture recognition in real time. *Applied Soft Computing 12*, 8 (Aug. 2012), 2072–2084.
9. Tse, E., Greenberg, S., and Shen, C. Gsi demo: multiuser gesture/speech interaction over digital tables by wrapping single user applications. In *Proceedings of the 8th international conference on Multimodal interfaces*, ICMI '06, ACM (New York, NY, USA, 2006), 76–83.
10. Vo, M. T., and Wood, C. Building an application framework for speech and pen input integration in multimodal learning interfaces. In *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 6 (May), 3545–3548 vol. 6.
11. Ward, D. J., and Mackay, D. J. C. Fast Hands-free writing by Gaze Direction. *Nature 418*, 6900 (2002).