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

Matheus Vieira Portela
Mechatronics Engineering Student
University of Brasilia, Brazil
Scholar of CNPq
matheus.v.portela@gmail.com

David Rozado
Postdoctoral Fellow
ICT Centre
CSIRO, Australia
David.Rozado@csiro.au

March 18, 2013

1 Abstract

Algorithms based on modeling speech as a finite-state hidden Markov process have been the most successful approach used to solve the automatic speech recognition problem. Nonetheless, the performance of current speech recognition algorithms is well below that of humans. Error rates of computational speech recognition are an order of magnitude greater than human speech recognition in quiet environments. Machine performance degrades even further in environments with noise channel variability and spontaneous speech. Humans can also better recognize sounds with very little high-level grammatical information. The error rates of current speech recognition systems are frustrating for the humans using them. In this work, we present an innovative form of correcting misrecognized words during a speech recognition task by using gaze tracking technology. Specifically, we propose to use the user's gaze to point at misrecognized words during a speech recognition task and to select appropriate word alternatives also by gaze. We compare the performance of this multimodal approach to speech recognition with traditional modalities of correcting misrecognized words: usage of the mouse and the keyboard and usage of voice alone. The results of the user study show that while the proposed system is not as fast as using the mouse and the keyboard for correction, it remains a truly hands-free means of interaction with the computer with obvious advantages for certain types of users such as those who lack motor mobility or dexterity of the hands or those suffering from repetitive strain injuries. The system is also advantageous for scenarios that prohibit or discourage the usage of the hands to interact with the computer. The gaze enhanced correction of misrecognized words also significantly outperforms the other truly hands-free correction modality: voice. This results suggest the advantage of the proposed gaze enhanced speech recognition modality to improve the speech recognition experience for a variety of users and scenarios.

2 Keywords

Eye tracking, gaze tracking, gaze aware interfaces, pervasive computing, speech recognition, multimodal interaction, human computer interaction, HCI, gaze responsive systems

3 Introduction

In computer science, speech recognition 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 a significant increase in the accuracy performance of speech recognition software over time, error rates still make the technology cumbersome to use for everyday interaction and functional speech recognition still requires a relatively long learning curve on the part of the user to properly master it. Still, misrecognized words occur often during speech recognition tasks even for advance users. For instance, 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 [1].

The correction of speech recognition errors or misrecognized words can be carried out manually with a traditional keyboard and mouse to select the misrecognized word and retype it. The usage of the keyboard and mouse to correct misrecognized words during speech recognition is probably the most widely used method for the task. This modality can be perfectly valid for some scenarios, but this interaction modality is not truly hands-free anymore.

The usage of voice for correction of misrecognized words during speech recognition 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" difficult words are extremely difficult to be properly recognized by speech recognition engines. Hence, this correction modality is extremely error-prone and frustrating for users since the user has to often repeat over and over again the same words until it is properly recognized. Users with a foreign accent struggle markedly with this form of interaction. Furthermore, 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 [2].

In this work, we propose the enhancement of speech recognition software with gaze tracking technology to speed up the correction of misrecognized words and to maintain speech recognition 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 [3]. With the proposed modality, the user is required to simply gaze at the misrecognized word and then simply select the correct word from a gaze dependent emerging panel of most likely alternative words. If the correct word is not in the panel of alternative words, the user has the chance of trying to utter it again. The correct word is selected from the panel 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 and usage of gaze.

To our knowledge, the idea of using gaze to correct misrecognized words in a speech recognition task has not been explored before in the research literature. There exists however work on gaze aware systems that use gaze tracking data to adapt their behavior to the patterns of user's gaze.

The notion of gaze attentive interfaces is neatly explained in [4]. That work focused on examining the benefits and limitations of using eye movements in the human computer interface (HCI). Authors tested the performance of gaze aware applications and concluded that they had a potential to be more pleasing and effective than traditional application interfaces.

In [5], authors introduced a gaze-based interface for browsing and searching images that refined search engine prediction results using relevant images obtained by monitoring the user's gaze patterns over the first images returned by the search engine. The work showed that there was sufficient information in the gaze patterns to complement or even replace explicit feedback given by the mouse on images of interest.

Authors in [6] explored the potential of a Computationally Aided Diagnostics System (CADS) through making it context sensitive by providing decision-support to radiologists' focus of attention during visual search in a mammography analysis task. Their system combined machine learning prediction algorithms to obtain a list of coordinate places on the image with potentially cancerous tissue to direct the attention of the radiologist towards those locations

The work from [7] studied gaze behavior in a first person shooter game and suggested that game engines might use this knowledge to anticipate actions that players have not executed yet. Authors provided a convincing argument that a gaze dependent anticipation module should enhance game character behaviors and make them much ”smarter”.

The project text 2.0 from [8] studies the advanced display of text by monitoring user’s gaze during reading while providing extra features to enhance the reading experience. In this manner, gaze behavior is used to make text responsive to readers’ gaze patterns by, for instance, providing dynamic pop up definition boxes when a user gazes at a word for an amount of time exceeding a predefined threshold.

Authors in [9] described the usage of focus points to improve the visual effects in virtual environments by adapting the rendering technique of a 3-D model to a user’s gaze position using models of visual attention. This was done with the intention of improving users’ sensations during first person navigation in the 3-D environment. In a similar work, authors in [10] described a dynamic display system that naturally and interactively adapts the display properties as the user’s eyes move around a high definition panoramic scene.

Gaze has also been used to input text or commands in a variety of scenarios [11]. Authors in [12] showed an innovative approach for fast text input using gaze named Dasher. Dasher is a dynamic text input system controlled by gaze that uses a language model to facilitate letter completion and selection within a word/sentence. Users enter text by tracking the desire character with the eyes, rather than by performing discreet gestures. Dasher can write up to 25 words per minute (wpm) after considerable familiarization with the system in comparison to the maximum of 15 wpm for expert users using an on-screen keyboard.

Finally, the work from [13] studied the effectiveness of gaze guidance on the visual performance of drivers. Their system monitors drivers’ gaze to provide them with alerts of potential dangers on the road (obstacles, incoming pedestrians, nearby cars) if the gaze monitoring system determined that the driver is not paying attention to them. Such a system detects driver’s distraction with respect to an incoming pedestrians coming from the right and it will display an arrow pointing towards the pedestrian in the area of the windshield where the user is paying attention to.

Once considered pure science fiction, automatic speech recognition (ASR) has been well studied over the past few decades over various distinguishing perspectives. This includes the acoustic-phonetic approach, pattern recognition approach, knowledge based approach, connectionist approach, and using support vector machines. Ghai et al. summarizes these various points of view in their work [14].

The speaking mode is an important factor that has to be considered in ASR systems, since its constraints require different designs. An isolated word speech recognition (IWR) assumes that one single word with well defined beginning and ending points is being uttered, which imposes great limitations on ASR. In order to deal with several words separated by pauses, connected word recognition (CWR) should be considered. Continuous speech recognition (CSR) deals with the scenario where words are connected, i.e., the boundaries of each individual word are unclear, requiring more sophisticated systems, such as the usage of Hidden Markov Models (HMM) [15]. Finally, there is the main objective for ASR systems: spontaneous speech recognition - the most natural speech mode. The difficulty here lies on the various odd sounds humans produce when speaking in a natural fashion, such as when expressing hesitation or uncertainty [14].

Even with state-of-the-art approaches to ASR, external noise degrades considerably the quality of the recognition. Gong [16] presents several mathematical approaches that deals exclusively with improving recognition in noisy situations, specially the most successful technique: speech model compensation. Still in this area of research, Chan et al. [17] proposed to use myo-electric signals to refine ASR in extremely noisy environments, with promising results.

Another aspect that quickly degrades quality of speech recognition is the presence of accent in the user speech. In their work, Kat and Fung [18] employ accent detection and accent-adaptive recognition for Chinese users to transform a native accent dictionary in an accent-adapted dictionary using as prior information the native language of the speaker. Using this methodology, the error rate of recognition could drop by up to 13.5%.

When these variables that rise the recognition error rate are slightly controlled, ASR systems find appropriate niche applications, specially when integrated with different interaction modalities. Vo and Wood [19] present a multimodal calendar integrating speech, gesture and handwriting recognition systems. The main advantage of multimodal systems is that the final hypothesis can integrate recognition results from each modality alone and, in the end, produces a better and more complete recognition. Even though recognition errors can degrade performance, the multimodal approach counterbalances it.

Tse, Greenberg and Shen [20] use speech- and gesture-based recognition over digital tables to transform a previously single-user interaction into a multi-user one. In this exciting research, they enabled well-known commercial software such as the map-viewer and games to be used simultaneously by many users, where speech was responsible for command and control interaction (e.g., 'stop' or 'fly to Boston') and gestures could enable selection, zooming or dragging, for instance.

In summary, we present here a multimodal approach to HCI that augments speech recognition with a gaze attentive interface for correction of misrecognized words. This approach offers the advantage to maintaining a truly hands-free paradigm of inputting text into a computer. To obtain empirical validation to our hypotheses, we explore through a comprehensive user study how this multimodal paradigm compares to the two traditional modalities used to correct misrecognized words, using voice alone and using the mouse and the keyboard, in terms time, task completion rate, and physical effort required to complete the task.

4 Methodology

The goal of the user study was to compare the correction of misrecognized text using the gaze-based correction method with the voice and mouse and keyboard correction methods. For this purpose, a graphical user interface (GUI) was designed and implemented to be shown to the subjects for each one of the 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.

4.1 Eye Tracking

Eyes are used by humans to obtain information about the surroundings and to communicate information. When something attracts our attention, we position our gaze on it, thus performing a *fixation*. A fixation usually has a duration of at least 150 milliseconds (ms). The fast eye movements that occur between fixations are known as *saccades*, and they are used to reposition the eye so that the object of interest is projected onto the fovea. The direction of gaze thus reflects the focus of *attention* and also provides an indirect hint for *intention* [21].

A video-based gaze tracking system seeks to find where a person is looking, i.e. the Point of Regard (PoR), using images obtained from the eye by one or more cameras. Most systems employ infrared illumination that is invisible to the human eye and hence it is not distracting for the user. Infrared light improves image contrast and produces a reflection on the cornea, known as corneal reflection or glint. Eye features such as the corneal reflections and the center of the pupil/iris can be used to estimate the PoR. Figure 1 shows a screenshot of an eye being tracked by the open-source ITU Gaze Tracker [22, 23]. In this case, the center of the pupil and two corneal reflections are the features being tracked.

4.2 Speech Recognition

Speech is one of the most natural communication skills. Humans use the vocal tracts to produce sounds, the air as a communication channel, and the auditory system as a pass-band for frequencies in the 0-20kHz spectrum. The basic unit of speech are the *phonemes* which, when sequenced, are decoded as words by another person. However, the pronunciation of phonemes leads to *coarticulation*, i.e., to utter them without an intervening pause, what affects the neighbor phonemes in several ways [24].

In order to perform recognition, many pattern recognition techniques have been employed. When words are uttered, hidden Markov models compare the collected data to a set of probabilistic density

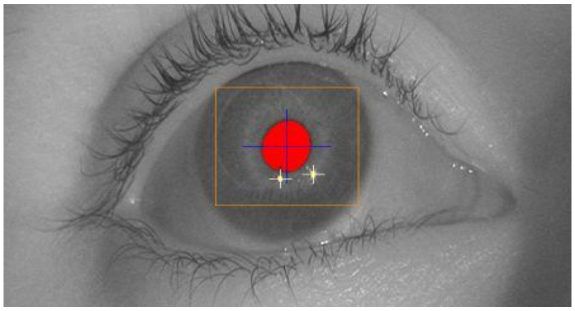


Figure 1: **The Open Source ITU Gaze Tracker tracking one eye.** The features tracked in the image are the pupil center and two corneal reflections. These features I used by the gaze estimation algorithms to determinine the PoR of the user on the screen.

functions (PDFs) of *acoustic models* and select the one with highest probability. Afterwards, the result is integrated to *language models*, which incorporates syntax and semantics to ASR [24].

4.3 Participants

Nineteen participants took part in the user study. Among them, there was 10 native English speakers and 9 non-native English speakers (male and female). This is mentioned to account for the fact that speech recognition performance varies significantly between native and non-native speakers.

4.4 Apparatus

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

4.5 Experimental task

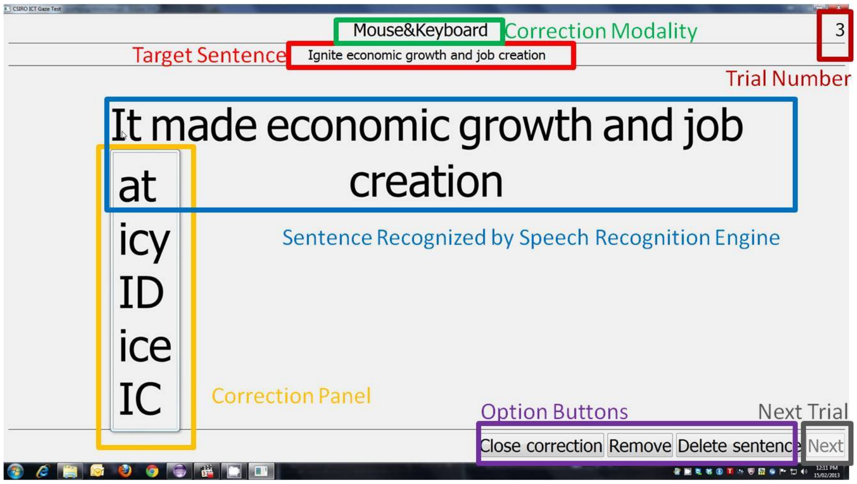
Each participant was requested to try dictating 10 sentences with an allotted time of one minute maximum per sentence. This task was repeated three times, one for each correction modality: gaze, voice and mouse and keyboard making a total of 30 dictated sentences for each subject. The sentences and the correction modality were randomly assigned from a previously written dataset in order to smooth out ordering effects of the correction modalities in speech adaptation. Previews to the beginning of each experimental trial, we ran a few test trials of each correction modality for each subject to make the subject comfortable with the interface.

The user interface, displayed in Figure 2 shows the correction modality, the target sentence and the number of the trial (from 1 to 30), on the top of the screen. When a speech utterance was recognized by the ASR, the 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 pop-up menu would be shown with similar 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 pop-up menu could be generated by saying the command “correct” followed by the desired word to correct. In this mode, alternative words are preceded by a number that, when pronounced, selects the corresponding word. When the desired correction was not presented

in the pop-up menu, pronouncing again the word would refresh the menu with a new list of best-guess alternatives produced by the speech recognition engine.

- Gaze: Fixating on a word for a dwell time of 2 seconds would generate the pop-up menu with alternatives. Similarly, the alternative word could be selected by fixating the gaze for 2 seconds. Pronouncing again the word would refresh the menu with new alternatives.



Ignite economic growth and job creation

It made economic growth and job creation

Figure 2: **GUI Used in the Experiments.** This figure shows highlighted in different colors the different parts composing the user interface exposed to the subjects participating in the user study.

On the bottom of the screen, three options buttons were available when a word was being corrected, i.e., with the pop-up menu opened: "Close correction", which would close the pop-up menu of alternative words to substitute a misrecognized word, "Remove", that would delete the word, and "Delete sentence", which would delete the entire recognized sentence. Also, a "Next" button would become active when the recognized sentence matched the target sentence or when the allotted. When the user had correctly pronounced the desired sentence, the "Next" button and the target sentence would become green to indicate that the user could go to the next trial. However, if the desired sentence was not reached in less than 60 seconds, the "Next" button and the target sentence would become red, indicating that the user has failed to correct the sentence in suitable time and could start the next trial. All these buttons could be activated either by using the mouse or the active given correction modality.
At the end of the experiments, the subjects involved in the user study were required to fill a questionnaire asking them about their subjective experience with the different correction modalities.

4.6 Measures

During the experiment, time to complete the task was measured. Timing would stop when the active sentence being uttered and edited matched the target sentence or if the user failed to achieve the target sentence within the allotted 60 seconds threshold. When correcting by mouse and keyboard, the mouse displacement and keystrokes were recorded. Each sentence also had a mark indicating whether the trial has failed or not.
Also, both the pronounced and the target sentence were recorded, this data was later used to calculate the Damerau-Levenshtein distance [25] between them. This distance revels how far two strings are from each other considering four basic operations: insertion, deletion, substitution of a single character, and transposition of two adjacent characters.
Lastly, the user questionnaire was composed by the following five questions, where the subject were able to answer either "Traditional keyboard and mouse", "Voice" or "Gaze".

- Which method do you find 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 more 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?

5 Results

The average time of each experimental trial to achieve the target sentence using a given correction modality is displayed in Figure 3. 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 $F(2, 54) = 17.43$, $p = 1.44 \times 10^{-06}$. A Posthoc Bonferroni-Holm test indicated significant differences between the voice-mouse&keyboard, gaze-mouse&keyboard and gaze-voice modalities with $p = 4.24 \times 10^{-06}$, $p = 3.6 \times 10^{-05}$ and $p = 0.02$, respectively.

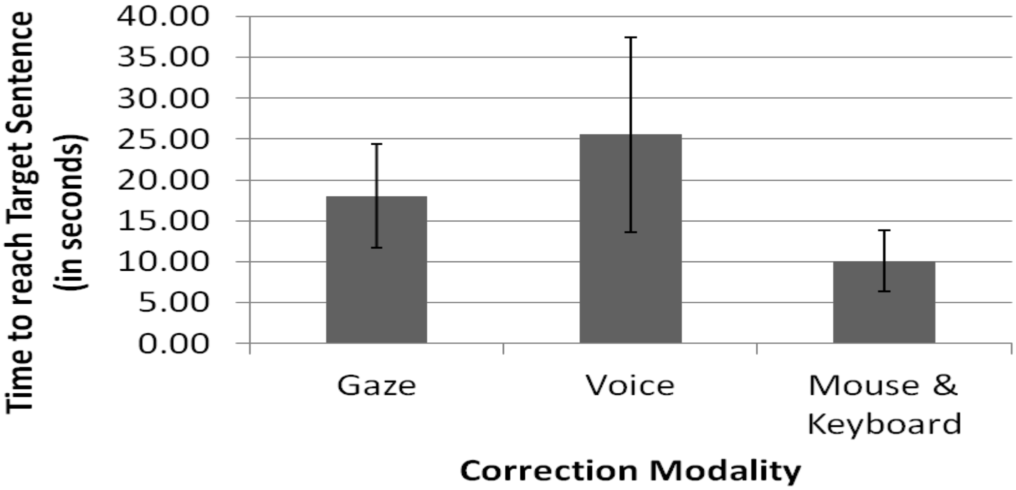


Figure 3: **Average Time Required to Achieve Target Sentence.** The figure displays the average time that the different correction modalities required to reach the target sentence.

Figure 4 shows the average displacement in pixels of the mouse during each experimental trial for each correction modality. Again, a Levene's test failed to find equal variance among experimental conditions. The ANOVA analysis showed significant differences between groups, $F(2, 54) = 25.06$, $p = 2.0 \times 10^{-08}$. A Posthoc Bonferroni-Holm test found significant differences between the gaze-mouse&keyboard and voice-mouse&keyboard modalities, $p = 1.48 \times 10^{-05}$, $p = 1.48 \times 10^{-05}$ respectively.

The average number of keystrokes required from the subjects to achieve the target sentence by each correction modality during the experimental trials is shown in Figure 5. Again a Levene's test failed to assume equal variance. The ANOVA analysis showed statistically significant differences between the results of the difference correction modalities, $F(2, 54) = 36.79$, $p = 8.29 \times 10^{-11}$.

Figure 6 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 difference correction modalities, $F(2, 54) = 17.92$, $p = 1.07 \times 10^{-06}$.

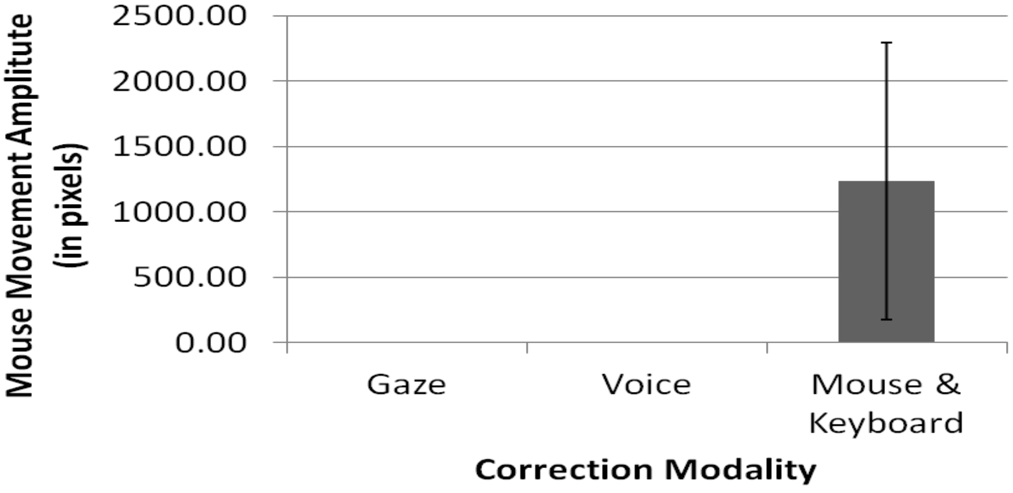


Figure 4: **Average Mouse Movement Amplitude Required to Complete the Task.** The figure displays the average mouse movement amplitude in pixels required by each correction modality to reach the target sentence during the experimental trials.

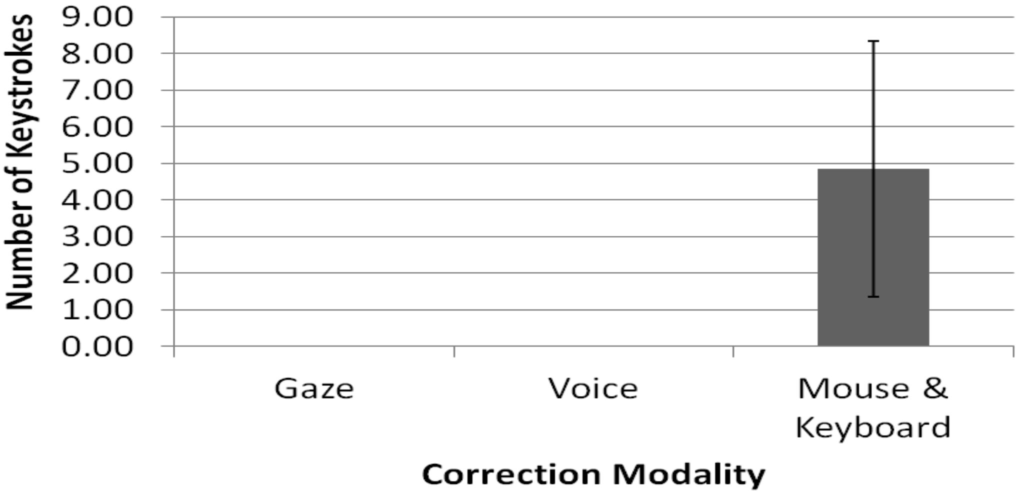


Figure 5: **Number of Keystrokes Required to Complete the Task.** The figure displays the average mouse movement amplitude (in pixels) required by each correction modality to reach the target sentence during the experimental trials.

A Posthoc Bonferroni-Holm test found significant differences between the voice-mouse&keyboard modalities, $p = 6.62 \times 10^{-06}$, gaze-mouse&keyboard modalities $p = 9.23 \times 10^{-05}$ and gaze-voice modalities $p = 4.02 \times 10^{-03}$.

The average Damerau-Levenshtein distance between the target sentence and the generated sentence within the 60 seconds time window for each correction modality is shown in Figure 7. The ANOVA analysis generated statistically significant differences between modalities with values $F(2, 54) = 11.60$, $p = 6.44 \times 10^{-05}$. A Posthoc Bonferroni-Holm test found significant differences between the voice-mouse&keyboard modalities, $p = 5.18 \times 10^{-04}$, gaze-voice modalities $p = 4.44 \times 10^{-03}$ and gaze-mouse&keyboard modalities $p = 1.71 \times 10^{-02}$.

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 8.

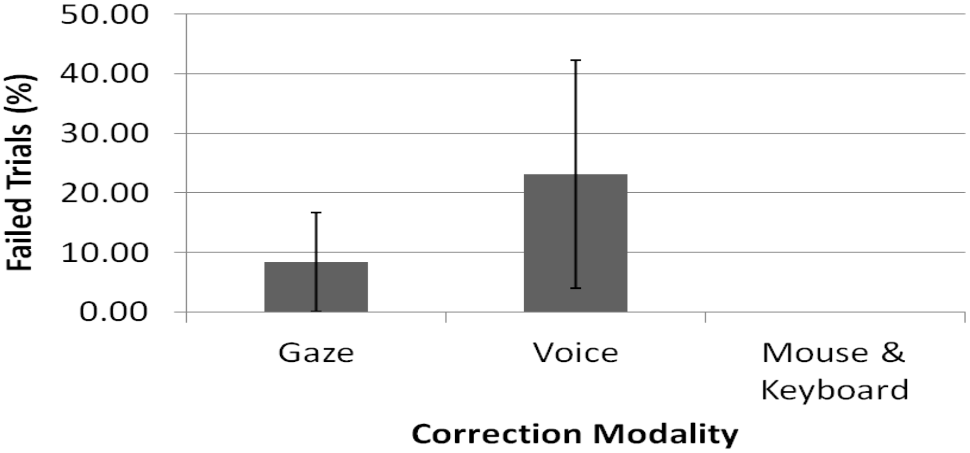


Figure 6: **Number of Failed Trials.** The figure displays the average number of trials in which the user was unable to reach the given target sentence within the allotted time of 60 seconds.

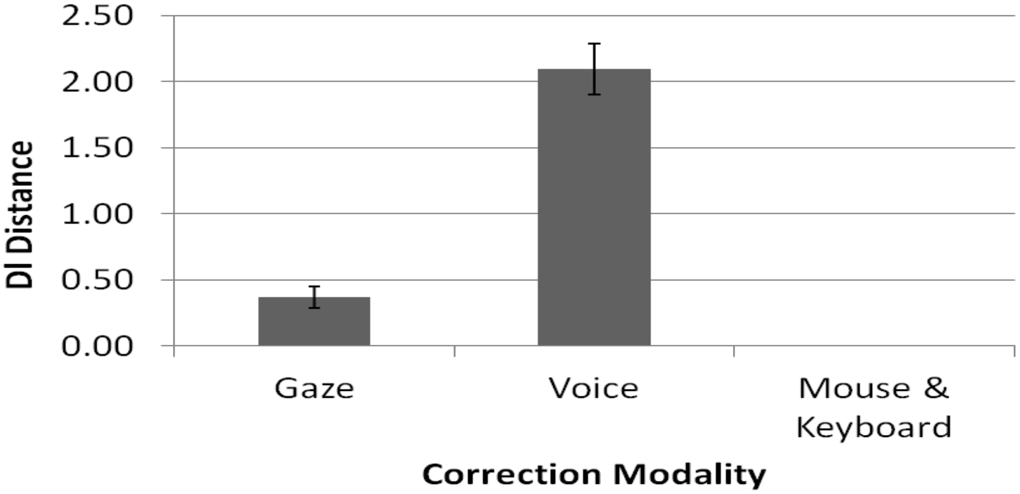


Figure 7: **Damerau-Levenshtein Distance.** This figure shows the average Damerau-Levenshtein distance between the target sentence and the generated sentence.

6 Discussion

The results of the user study showed that the gaze enhanced correction modality for a speech recognition task is not as fast as using the mouse and the keyboard for correcting misrecognized words. Yet, the gaze modality is significantly faster that using voice alone for correction and it is truly hand-free as opposed to the Mouse&Keyboard modality that required the usage of the hands for correction of misrecognized words. The time performance of the gaze enhanced speech recognition can be improved in future researches by implementing more sophisticated selection methods, replacing the 2 seconds fixation used in this work, by for instance dynamic and continuous selection of words just by looking at them. This alone would most likely improve the speed performance of the gaze correction modality.

The amount of mouse movement displacement required to achieve the target sentence was obviously non-existent for the gaze and voice correction modalities due to their true hands-free nature. The same can be say of the amount of keystrokes presses required for correction. The gaze and voice

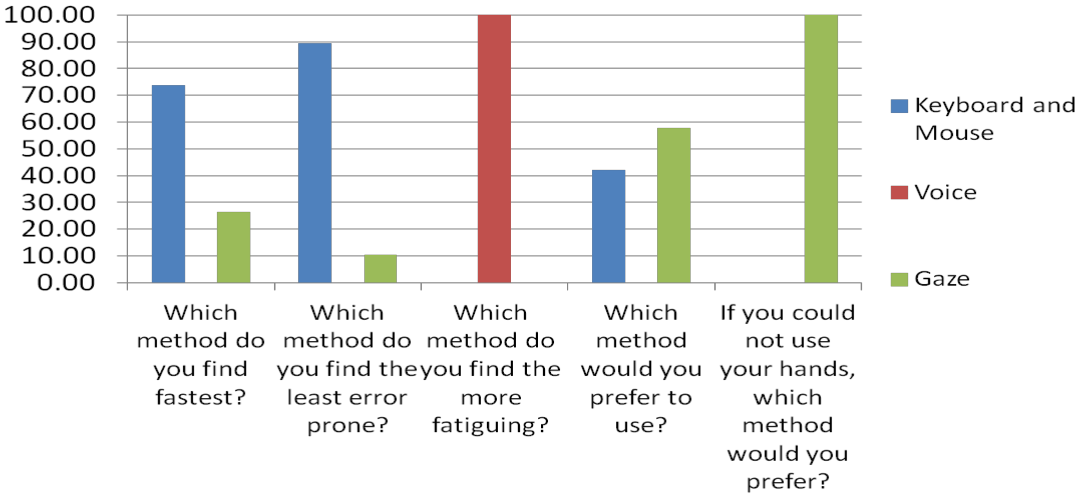


Figure 8: **User questionnaire.** Subjective opinions expressed by the subjects that participated in the user study about their perceptions of the different modalities being compared to correct misrecognized words.

correction modalities are obviously advantageous in this regard, since they completely eliminate the need of the physical effort associated to mouse movements and keystrokes presses. This is particularly beneficial for users with limited or no movement control of the hands, limited dexterity, users suffering from repetitive strain injury (RSI) and for scenarios that prohibit or discourage hand-based interaction. Voice correction of misrecognized words however, can easily lead to voice strain. This is due to the accuracy limitations of speech recognition algorithms that make it necessary for the users to repeat several times the same word until it is correctly recognized. Gaze-based correction, on the other hand, does not suffer from any of the previously mentioned drawbacks of the alternative modalities.

In terms of the number of failures registered during the experimental trials to achieve the target sentence, the gaze modality was significantly better at reaching the target sentence than the voice modality but worse than the mouse & keyboard modality. This was also evident in the Damerau-Levenshtein distance between the target sentence and the achieved sentence during the experimental trials. Here, it is important to emphasize that the most of the subjects involved in the user study had never been exposed to gaze tracking technology before. Hence, they did not have time to properly familiarize themselves with the technology. Given enough time to go through the learning curve of using gaze interaction would most likely allow learning effects to take place and improve the performance of these measurements.

The main limitation of the proposed modality is due to the constraints in gaze accuracy that any gaze estimation algorithm has. Using the proposed system with the typical font size that average users employ for editing texts would be challenging since it would be difficult for the system to discern the particular word at which the user is gazing at. Hence, in our user studies we employed relatively large font sizes. 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.

In 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 & keyboard modality but it possesses the advantage of being truly hand-free with obvious implications for handicapped computer users, users suffering from RSI syndrome and for scenarios where the usage of the hands is not possible (surgery room for instance). Moreover, the gaze modality significantly outperforms the other hand-free modality to correct misrecognized words, using voice, in all the variable being monitored in the user study. Furthermore, the gaze based correction modality also prevents the appearance of voice strain for the correction of misrecognized words since it prevents considerably the amount of 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 speech recognition with gaze tracking and gaze interaction to create a truly hands-free 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] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181 – 200, 2010.

[2] B. Fritzell, "Voice disorders and occupations," *Logopedics Phoniatrics Vocology*, vol. 21, no. 1, pp. 7–12, 1996.

[3] D. Rozado, F. B. Rodriguez, and P. Varona, "Low cost remote gaze gesture recognition in real time," *Applied Soft Computing*, vol. 12, pp. 2072–2084, Aug. 2012.

[4] A. Hyrskykari, *Eyes in attentive interfaces: Experiences from creating iDict, a gaze-aware reading aid.* PhD thesis, University of Tampere, Department of Computer Sciences, 2006.

[5] L. Kozma, A. Klami, and S. Kaski, "GaZIR," in *Proceedings of the 2009 international conference on Multimodal interfaces - ICMI-MLMI '09*, (New York, New York, USA), p. 305, ACM Press, Nov. 2009.

[6] G. D. Tourassi, M. A. Mazurowski, B. P. Harrawood, and E. A. Krupinski, "Exploring the potential of context-sensitive CADe in screening mammography," *Medical Physics*, vol. 37, p. 5728, Oct. 2010.

[7] H. Koesling, A. Kenny, A. Finke, H. Ritter, S. McLoone, and T. Ward, "Towards intelligent user interfaces," in *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications - NGCA '11*, (New York, New York, USA), pp. 1–8, ACM Press, May 2011.

[8] R. Biedert, G. Buscher, S. Schwarz, J. Hees, and A. Dengel, "Text 2.0," in *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems - CHI EA '10*, (New York, New York, USA), p. 4003, ACM Press, Apr. 2010.

[9] S. Hillaire, A. Lecuyer, R. Cozot, and G. Casiez, "Using an Eye-Tracking System to Improve Camera Motions and Depth-of-Field Blur Effects in Virtual Environments," in *2008 IEEE Virtual Reality Conference*, pp. 47–50, IEEE, 2008.

[10] S. Rahardja, F. Farbiz, C. Manders, H. Zhiyong, J. N. S. Ling, I. R. Khan, O. E. Ping, and S. Peng, "Eye HDR," in *ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies: Adaptation on - SIGGRAPH ASIA '09*, (New York, New York, USA), p. 68, ACM Press, Dec. 2009.

[11] D. Rozado, F. B. Rodriguez, and P. Varona, "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*, pp. 1–8, Springer Berlin / Heidelberg, 2011.

[12] D. J. Ward and D. J. C. Mackay, "Fast Hands-free writing by Gaze Direction," *Nature*, vol. 418, no. 6900, 2002.

[13] L. Pomarjanschi, M. Dorr, and E. Barth, "Gaze guidance reduces the number of collisions with pedestrians in a driving simulator," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, pp. 1–14, Jan. 2012.

[14] W. Ghai and N. Singh, "Article: Literature review on automatic speech recognition," *International Journal of Computer Applications*, vol. 41, pp. 42–50, March 2012. Published by Foundation of Computer Science, New York, USA.

[15] S. Young, "Statistical modelling in continuous speech recognition (csr)," in *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, UAI'01, (San Francisco, CA, USA), pp. 562–571, Morgan Kaufmann Publishers Inc., 2001.

[16] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, no. 3, pp. 261 – 291, 1995.

[17] A. Chan, K. Englehart, B. Hudgins, and D. Lovely, "Myo-electric signals to augment speech recognition," *Medical and Biological Engineering and Computing*, vol. 39, pp. 500–504, 2001.

[18] L. W. Kat and P. Fung, "Fast accent identification and accented speech recognition," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 1, pp. 221–224 vol.1, Mar.

[19] M. T. Vo and C. Wood, "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, pp. 3545–3548 vol. 6, May.

[20] E. Tse, S. Greenberg, and C. Shen, "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, (New York, NY, USA), pp. 76–83, ACM, 2006.

[21] B. Velichkovsky and J. P. Hansen, "New Technological Windows into Mind: There is More in Eyes and Brains for Human-Computer Interaction.," in *CHI'96*, pp. 496–503, 1996.

[22] J. San Agustin, H. Skovsgaard, E. Mollenbach, M. Barret, M. Tall, D. W. Hansen, and J. P. Hansen, "Evaluation of a low-cost open-source gaze tracker," in *ETRA '10: Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*, (New York, NY, USA), pp. 77–80, ACM, 2010.

[23] D. Rozado, J. S. Agustin, F. B. Rodriguez, and P. Varona, "Gliding and saccadic gaze gesture recognition in real time," *ACM Transactions on Interactive Intelligent Systems*, vol. 1, pp. 1–27, Jan. 2012.

[24] D. O'Shaughnessy, "Invited paper: Automatic speech recognition: History, methods and challenges," *Pattern Recognition*, vol. 41, no. 10, pp. 2965 – 2979, 2008.

[25] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Commun. ACM*, vol. 7, pp. 171–176, Mar. 1964.