Leveraging Error Correction in Voice-based Text Entry by Talk-and-Gaze

Korok Sengupta
University of Koblenz-Landau
Koblenz, Germany
koroksengupta@uni-koblenz.de

Sabin Bhattarai
University of Koblenz-Landau
Koblenz, Germany
sbhattarai@uni-koblenz.de

Sayan Sarcar
University of Tsukuba
Tsukuba, Japan
sayans@slis.tsukuba.ac.jp

I Scott MacKenzie
York University
Toronto, Canada
mack@cse.yorku.ca

Steffen Staab
Universität Stuttgart
Stuttgart, Germany
&
University of Southampton,
Southampton, UK
s.r.staab@soton.ac.uk

ABSTRACT

We present the design and evaluation of Talk-and-Gaze (TaG), a method for selecting and correcting errors with voice and gaze. TaG uses eye gaze to overcome the inability of voice-only systems to provide spatial information. The user's point of gaze is used to select an erroneous word either by dwelling on the word for 800 ms (D-TaG) or by uttering a "select" voice command (V-TaG). A user study with 12 participants compared D-TaG, V-TaG, and a voice-only method for selecting and correcting words. Corrections were performed more than 20% faster with D-TaG compared to the V-TaG or voice-only methods. As well, D-TaG was observed to require 24% less selection effort than V-TaG and 11% less selection effort than voice-only error correction. D-TaG was well received in a subjective assessment with 66% of users choosing it as their preferred choice for error correction in voice-based text entry.

Author Keywords

Text Entry; Voice; Eye Tracking; Multimodal; Usability; Interaction Design

CCS Concepts

•**Human-centered computing** → **User studies; Text input;** Pointing devices;

INTRODUCTION

Recent improvements in speech recognition systems [1, 35] have made voice input a popular modality for digital interaction. Voice input is now widely adopted, a key factor being the speed of input compared to typing on a keyboard [32].

For voice-based text entry, validation of the entered text is necessary, as recognition errors are inevitable. Recognition challenges include ambient noise (that drowns out the voice), multiple voices speaking simultaneously, and recognition errors due to homophones or diction [42, 31]. These challenges impact not only the entry of text but also the use of voice commands to navigate the text and correct errors.

Error correction forms a major part of the text entry process. It involves the complex task of identifying errors, navigating to the errors, and then applying corrective measures. Thus, voice-based text entry that also involves validating and correcting the formed sentences is a challenge [30]. Sears et al. [35] suggest that 66% of the interaction time is spent in correcting errors with only 33% of the time used in transcribing. Karat et al. [12] note that the assumed productivity gain for speech dictation systems depreciates when error correction is factored in.

One challenge for voice input is the inability to naturally provide spatial information. To correct an error, the first task is to navigate to the location of the error. But, navigation by voice is a challenge. Strategies include target-based navigation or direction-based navigation [7, 17, 22]. In both approaches, recalling and articulating the commands and applying corrective measures slows the overall speed of text entry. To overcome these challenges, research has investigated combining voice input with another modality [24, 25, 26, 10, 27].

Most approaches that involve an additional modality require physical input and this presents a challenge when the hands are used for an activity other than typing. Some users may lack the fine motor control required to accurately place a pen or similar device. Thus, the need for digital inclusion has

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376579

led researchers to investigate hands-free approaches for text entry and error correction. Gaze, a natural modality like voice, has been well investigated for web navigation [28, 20] and text entry and editing [13, 16, 37, 38]. Although gaze has the potential to complement the voice as an input modality, there is little research [29] that combines voice as the primary modality with gaze as the secondary modality.

Research Scope

We present a novel approach called "Talk-and-Gaze" or "TaG" that uses gaze as an additional modality for hands-free voice-based text entry. TaG facilitates error correction in a hands-free environment by utilizing the strengths of gaze and voice as input modalities. The identification of words to be edited comprises two interaction tasks: First, the spatial position in the text is defined by the gaze. Second, the position must be *selected* when the erroneous word is gazed at, but not when the gaze is used for reading and validating the text (to avoid the Midas-Touch problem [11]). We have implemented two versions of Talk-and-Gaze. D-TaG uses dwell-time selection: An erroneous word is selected if the user's gaze dwells on the word longer than a pre-defined time threshold. V-TaG uses voice command selection: An erroneous word is selected if the user utters a command to lock-in the word at the gaze location.

We address the following research questions:

RQ1: How can we naturally integrate gaze for error corrections in voice-based text entry?

RQ2: How do the D-TaG and V-TaG versions of TaG compare with conventional voice-based error correction?

To answer these questions, we performed a comparative evaluation of D-TaG, V-TaG, and Voice-Only error correction. We performed objective and subjective evaluations of the three edit methods for a *read and correct task*. This was followed by a subjective analysis of the image description task where users could freely form text based on what they perceived from the given images.

The contributions of this research are as follows:

- We present the design and implementation of TaG, two novel gaze-augmented voice-based error correction methods where voice and gaze work in parallel helping in selecting and correcting errors.
- We show that D-TaG performs better for most of the evaluation parameters against the V-TaG and Voice-only approaches.

RELATED WORK

Voice-based approaches

While speech-to-text has improved, recognition challenges and false transcription remain a hindrance for voice-based text input. Suhm et al. [40] describe error correction where the incorrect word is replaced by re-speaking it, perhaps multiple times. The drawback is that speech recognition errors may occur repeatedly for words where correction is required. This leads to lengthy attempts at correction which are time-consuming and yield a poor user experience. They also suggested selecting the correct word from a list, but they did not discuss the scenario where the list does not incorporate the desired word.

Substantial research is directed at voice-controlled navigation within a document [5, 17, 7, 22, 35, 33]. The goal is to efficiently locate and select the desired word through navigational voice commands. Navigation methods can be classified as continuous, direction-based, and target-based. However, each method comes with challenges. Continuous navigation techniques [17, 7, 22] lack the ability to fluidly and continuously generate movements without repeating the command. De Mauro et al. [7] describe the design of a voice-controlled mouse for direction-based commands. They shortened the commands by mapping them to simple vowels. For example, uttering 'A' continuously moves the mouse cursor left. Each vowel mapped to a command for mouse movement. Whilst this reduces the time to say lengthy commands, users must learn the mappings. Complexity in constructing valid direction-based commands alongside longer navigation sequences complicates the interaction. Sears et al. [35] proposed another direction based approach for error correction using spoken commands to navigate to the location of an erroneous word (e.g., *move up, move down, move left, move right*). There are several challenges including the length of the voice commands, the non-recognition of commands by the speech engine, and the subsequent inconvenience and fatigue in saying commands multiple times. Target-based navigation is efficient compared to direction-based navigation; however, recognition errors when executing commands is an issue. Hands-Free-Chrome[1], a voice-controlled plugin for Google Chrome, uses target-based navigational approach, efficiently handling the navigation of URLs on a web page through the command "map". The command assigns a unique integer to all the available URLs, with the user uttering the desired number after the "map" command. While accurate in selecting the desired links, it struggles when homonymic numbers are not understood by the system.

Current commercial examples of voice-based text entry like Zoho Docs[2], Google Docs[3], and Microsoft Office 365[4] also suffer from similar challenges and limitations.

Integration of an Additional Modality

Oviatt et al. [23, 26] tried to overcome the limitations of voice input by combining voice with pen-based gestures. They studied different GUI-based interfaces and reported that the task completion time improved for a multimodal approach compared to unimodal approach. Similar experiments by Mantravadi [18] combined voice and gaze for menu selections and showed improved accuracy and less ambiguity with a multimodal approach. Kumar et al. [15] combined gaze and keyboard with "look-press-look-release" interaction for web navigation. Sengupta et al. [36] combined voice and gaze for hands-free usage of a Web browser and found a 70% improvement in link selection using a multimodal approach compared to a unimodal approach. Castellina et al. [4] also

[1] https://www.handsfreechrome.com
[2] https://www.zoho.com/docs/
[3] https://www.google.co.uk/docs/about/
[4] https://www.office.com/

found improved performance in a hands-free multimodal environment.

Integration of Additional Modalities for Error Correction

McNair et al. [19] combined voice and mouse for error correction by re-speaking the incorrect word or by selecting an alternate word from a list. The latter approach has at least two problems: The correct word may be far down the list or the correct word may not appear in the list. Danis et al. [6] used a similar approach where the voice was the primary modality for text entry and the mouse was a secondary modality providing spatial information for locating and selecting the incorrect word. Sindhwani et al. [39] investigated error corrections in a multimodal environment by combining gaze with traditional keyboards.

Using Voice and Gaze for Error Correction

Beelders et al. [2] showed an approach to interacting with the GUI of Microsoft Word through voice and gaze. Although erroneous words were located through eye movement and fixation, corrections were done with the help of an on-screen keyboard where the keys were selected by a combination of gaze input and voice commands. However, the two modalities were not used simultaneously to achieve any intended task.

To the best of our knowledge, the sole contribution that combines voice and gaze for multimodal error correction in text entry is by Portela et al. [29]. They present a method that uses gaze (with a 2 s dwell) both to select an erroneous word and to select the correct word from a list of alternatives. This was compared to a voice method where the list of alternatives was numbered. Speaking the number selected the alternate word. However, if the correct word was not in the list, the user had to re-speak the word to alter the prediction list. This repeated approach leads to frustration if the correct word is not present or speech recognition errors persist.

PILOT STUDY: DESIGN INVESTIGATION

A popular use case for voice-based text entry is the Google Speech API for converting speech to text on Google Docs. This widely used system has built-in functions for error correction if the Speech API transcribes spoken words incorrectly.

We conducted a pilot study to investigate design challenges in using voice control in Google Docs. The aim was to collect user feedback on the advantages and disadvantages of such a system in a hands-free condition.

Five university students (4 male, 1 female, ages 22-29) volunteered for the study. All had prior knowledge of speech-based commands on hand-held devices; however, none had experience using voice in Google Docs. The study was divided into three parts.

First, the background and motivation for the study were explained. Participants were shown how voice commands work on Google Docs and how to correct errors. Second, each participant was asked to fix errors in five sentences without any additional help. Finally, participants read a passage and corrected erroneous words. They had to remember the voice commands and make corrections. They were then asked to share their experience and think of voice-based commands that are intuitive for them. This qualitative feedback was provided to us in writing.

The participants listed the following challenges which were taken into account when our voice-only approach was designed.

1. Remembering and recalling commands.
2. Inability to select the desired word when it occurs twice in a sentence. For example, if the sentence was *He had a big head, big teeth, a big nose, and a big attitude.* and the objective was to select the second *big*, the select command inadvertently selected the last *big* unless the cursor was explicitly positioned at the *big* in question.
3. Inability to promptly select a word that occurs multiple times across different paragraphs.
4. Effort to navigate across multiple incorrect words in passages.

VOICE-ONLY APPROACH

To overcome the challenges found in our pilot study, we designed the initial interaction for voice-only error correction using a "map" mode (see Section 2.1). For our implementation, the command "map" assigns a unique number to each erroneous word in a passage. The participant then utters the number to select a word. This eliminates the need to recall commands and allows the participant to directly select an incorrect word. It also eliminates the challenge when a word occurs twice in a sentence. The voice-only error correction system then offers a list of predictions along with three additional editing options (refer to Figure 1a, 1b):

1. *Delete* – delete the currently selected word.
2. *Spell* – substitute the currently selected word by a new word that is spelled. This mode is introduced as re-speaking the incorrect word often does not lead to correct recognition.
3. *Case Change* – toggle the case of a letter in word that has been accidentally capitalized or needs capitalization.

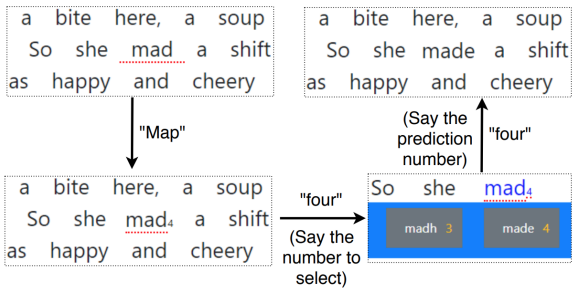
The map functionality also extends to "spell" mode where the participant performs letter-level correction for incorrect transcriptions caused by homophones, diction, or ambient noise. "Spell" mode gives an opportunity to spell the word in case recognition error occurs multiple times. The workflow of Voice-Only approach is seen in Figure 1.

PILOT STUDY II : DESIGN INVESTIGATION

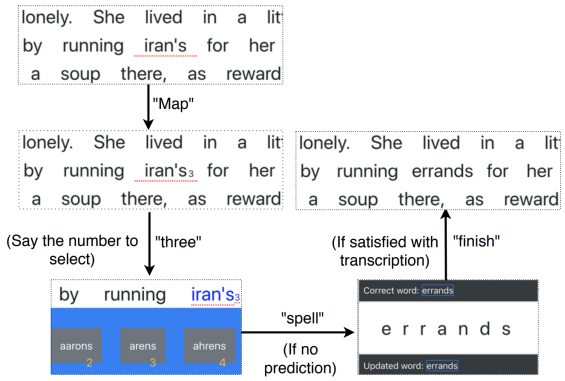
Based on the feedback from the first pilot study, the same participants were asked to use our voice-based approach and provide feedback. The investigation occurred in three parts and in the end participants were asked to share their experience.

Using our voice-based approach, participants noted the following:

1. Improved and quicker navigation style – they did not need to use long commands in comparison to the voice commands in Google Docs

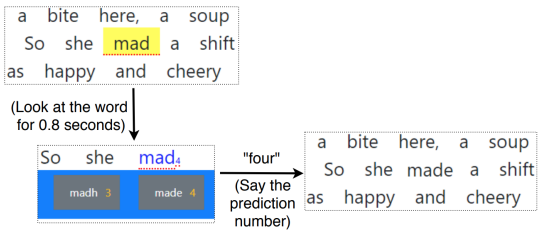


(a) Workflow of Voice-only approach using available predictions

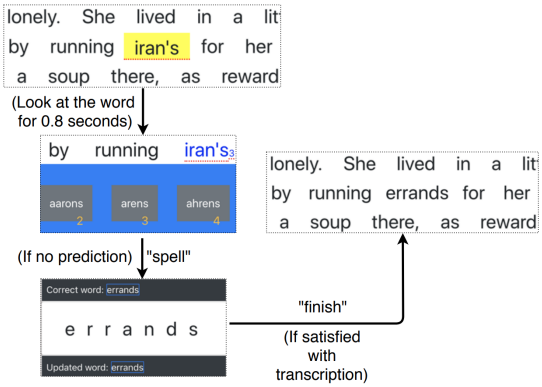


(b) Workflow of Voice-only approach using "SPELL" mode

Figure 1: Voice-only edit method using “map” functionality.



(a) Workflow of D-TaG approach using available predictions



(b) Workflow of D-TaG approach using “SPELL” mode

Figure 2: Dwell based D-TaG workflow depicting the “dwelling” approach which needs no verbal commands like “map” or “select” for selecting the erroneous word.

2. Predictions helped to quicken correction

3. Advantage of not adhering to one mode of error correction - Spell mode gives additional help.

4. Spell mode helped in distinguishing homonyms. Some words were homonymic because of the accents of non-native English speakers.

5. Repeated use of "map" command to select errors led to discomfort for some users.

TAG: AUGMENT VOICE-BASED TEXT INPUT WITH GAZE

From the feedback of the second pilot study, we understood that the map-based approach helps in minimizing navigational commands and ambiguity of word selection. However, it introduced an intermediate step in error correction. To reduce this for error correction, our design, TaG, augments voice with gaze to facilitate faster error selection followed by correction. A common challenge of gaze-based activation is Midas Touch [11]. This leads to incorrect triggering and eventual frustration. For our TaG method, we have examined two approaches to alleviate this:

1. D-TaG – Gazing and then dwelling on the incorrect word for 0.8 seconds selects the word and triggers the text predictions. While this minimizes the number of interaction steps, the risk of Midas Touch, or inadvertent triggering, remains. The dwell time was the average duration participants took between observing the incorrect transcription and calling out the mapped number in the second pilot study. The workflow is seen in Figure 2.

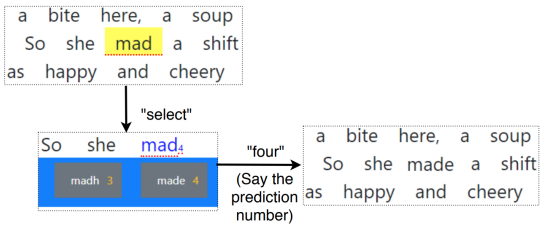
2. V-TaG – Focusing on the incorrect word and then saying “select” to select the erroneous word. While this avoids the Midas Touch problem, it also introduces an intermediate step in selecting the incorrect word. The workflow is depicted in Figure 3.

The selection of the predictions in both D-TaG and V-TaG used the voice to minimize recognition errors and the Midas Touch challenge. Uttering just the number associated with the correct prediction instead of the entire word also reduced the effort and recognition errors.

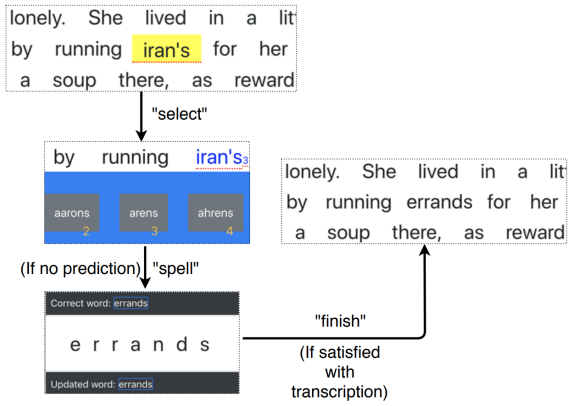
EXPERIMENT

Participants

Seventeen participants were recruited. All were well versed in English with B2 level proficiency and knew all the words in the sentence set. Most of the participants were university students with a background in computer science. While there was no problem in command recognition during our pilot study II, the recognition engine failed to understand the commands necessary for the selection of erroneous word for five of the participants during their training process. Non-recognition or mis-recognition of the keywords was due to the influence of heavy native language accent and this led to their exclusion at the onset of the training session. Ages ranged from 22 to 37 years (μ = 28.1, σ = 4.6). Seven participants were



(a) Workflow of V-TaG approach using available predictions



(b) Workflow of V-TaG approach using "SPELL" mode

Figure 3: V-TaG workflow showing the use of “select” to confirm the selection of an incorrect word highlighted by gaze.

male, five females. Five wore corrective devices for vision and five had prior experience in eye-tracking experiments. While some used voice commands on their smartphones, none had experience in voice-based typing or gaze-based typing. Participants were compensated 30 € for their time.

Apparatus

A Tobii EyeX[5]platform was used used to collect the gaze data. The eye tracker was attached below a 24-inch adjustable monitor. A stand-alone microphone was positioned beside the monitor on the desktop. Participants sat on a height-adjustable chair. See Figure 4. The experiments were conducted in an environment with controlled ambient light and sound. The software to evaluate the interactions was made on React Native[6] which recorded the participant’s performance. Data were stored in a *.csv* file for further evaluation.

Tasks

Evaluations of text entry and correction systems often employ a *copy task* where the participant copies text and then fixes errors if any occurred. Voice-based text entry evaluations frequently follow a similar protocol. The disadvantage is that the copying involves cognitive overhead; that is, reading then typing; this is atypical of most real-world situations.

[5]https://help.tobii.com/hc/en-us/categories/201185405-EyeX

[6]https://facebook.github.io/react-native/

Figure 4: Experimental setup showing the fixed display with the eye tracker, the stand-alone microphone, and a participant performing error corrections in "spell" mode.

Types of Error	Count
Missing Letter	37
Extra Letter	11
Double Letter	17
Mistakes	25

Table 1: Types of errors in the read and correct task. (For example: Missing - terible → terrible; Extra - hers → her; Double - upp → up; Mistake - want → went)

Therefore, our strategy was to let subjects perform a *read and correct task* and an *image description task* as described below.

Read and Correct Task. This task is motivated by situations when users encounter text they need to proofread and correct [39]. It allows for understanding the effort required in correcting erroneous text when already present. Since we wanted to investigate the interaction procedure, not the participants’ skill in finding errors, the errors were underlined in red (see Figure 2, 3). Underlining the error excluded visual search time from the interaction.

Image Description Task. Dunlop et al. [8] argue that evaluating text entry and editing requires free-form input that is not based on established transcription/copy tasks. They note that fixed-phrase copying provides internal consistency but lacks representativeness in natural text entry systems. Following their rationale, we adopted an image description task that they suggested. This setup is close to a realistic scenario of text creation and editing. We used the image dataset from Dunlop et al. [8].

Procedure

Participants first signed an informed consent form. This was followed by an explanation of the study. Then, they were shown how the system works by the experimenter (including the calibration procedure). After that the eye tracker was cal-

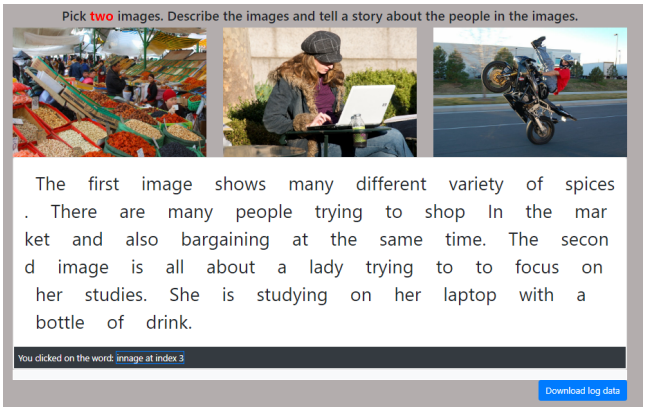


Figure 5: Image Description Task: Participants describe the images freely without any assistive visual marker to show errors.

ibrated to each participant using six calibration points. This was followed by a training block where they operated the system themselves. Once they were comfortable with the training process, the actual experiment started. Breaks were provided between sessions followed by participants recalibrating the eye tracker and continuing the test.

To offset order effects, participants were assigned in sequence to one of 3! = 6 orders for testing the three edit methods. After the experiment, participants completed the NASA TLX questionnaire, an SUS questionnaire, and an additional questionnaire. Testing took approximately 60 minutes per participant for each edit method. Participants were told that their gaze data will be recorded for evaluation purposes. Testing for each task included a screen recording for further analysis to understand the ease with which selection of erroneous words occurred.

For the *read and correct task*, the experiment consisted of a training block followed by five testing blocks. Each block included three passages, each with five errors that the user needed to correct. The passages were taken from American short stories[7]. Each passage was around 90 words which covered 50% of the screen space. The errors were chosen to include misspellings, incorrect letter entry, missing letters, and toggled order of letters. Table 1 summarizes the different types of errors and their count in the experiment.

For the *image description task*, there was a training block followed by three testing blocks. Participants were provided with three distinct images (as seen in Figure 5) for each block and were asked to describe any two. When they were satisfied with the transcription and corrections, they could go to the next image on uttering "next". However, the command only gets activated when "next" is mentioned after a pause. Each user performed three image description tasks where each set of images was different.

The procedure is illustrated in Figure 6.

[7] https://americanliterature.com/home

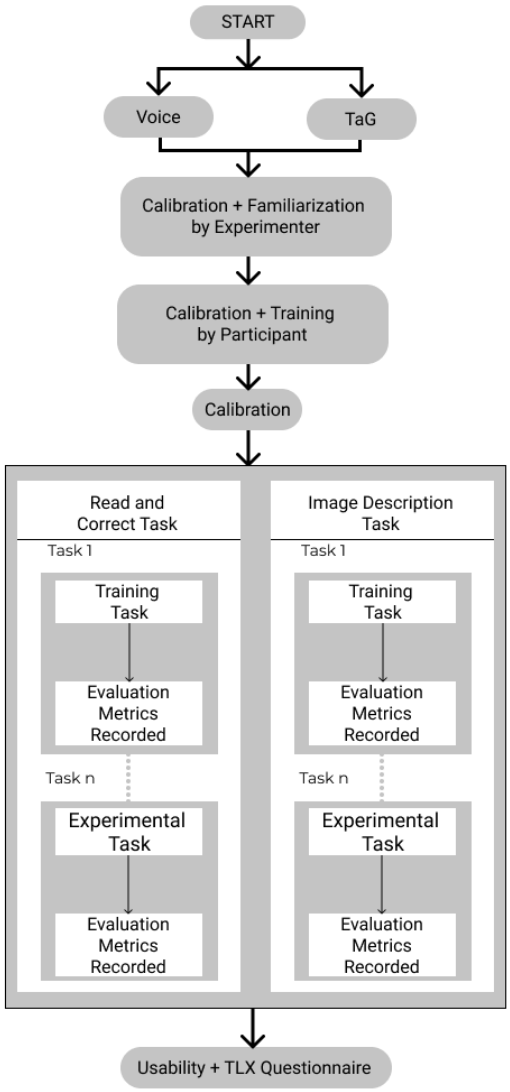


Figure 6: Experimental procedure for Voice-only, D-TaG and V-Tag edit methods

Design

The experiment was a 3 × 5 within-subjects design with the following independent variables and levels:

- Edit method (Voice-only, D-TaG, V-TaG)
- Block (1, 2, 3, 4, 5)

The dependent variables were block completion time (seconds) and selection effort (count). Block completion time was the time to correct all 15 errors in a block. Selection effort was a count of the number of events to select an erroneous word: the more selection events, the higher the assumed effort. By edit method, the events logged were non-recognition (Voice-only), a shift in focus or non-recognition of "select" (V-TaG), and selection miscues (D-TaG).

In summary, the total number of trials (corrections) was : 12 (participants) × 3 (edit methods) × 5 (blocks) × 3(passages per session) × 5 (error per passage) = 2700.

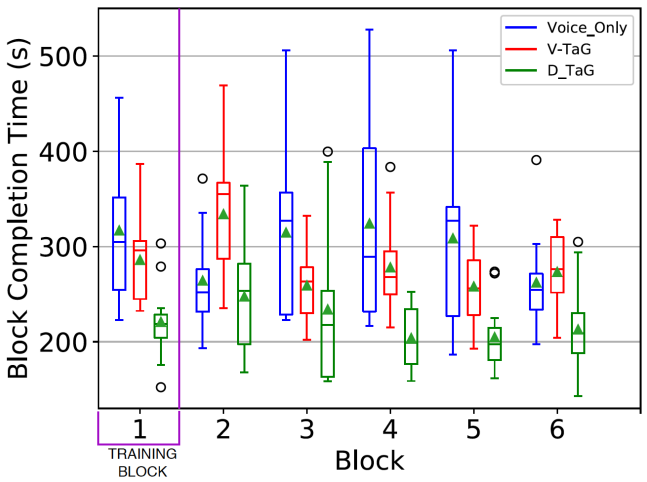


Figure 7: Block completion time (s) by edit method and block.

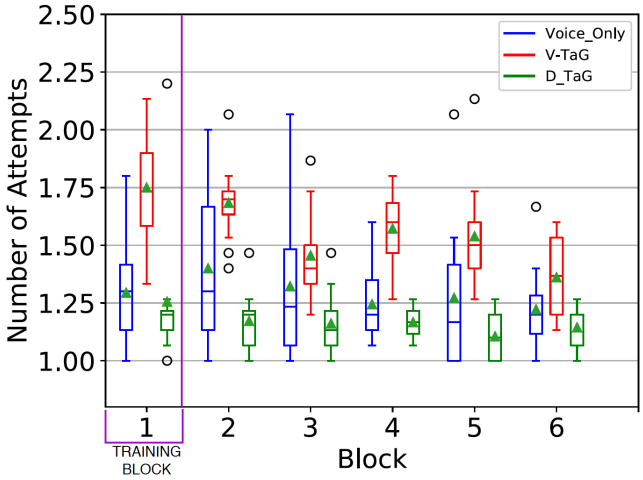


Figure 8: Selection effort (count) by edit method and block. (Selection effort is the number of attempts to select an erroneous word)

RESULTS

The detailed results of the Read and Correct task and Image Description task are described in the following subsections.

Read And Correct Task

Objective Measures

Block Completion Time. The grand mean for block completion time was 265.4 seconds. By edit method, the means were 294.8 s (Voice-only), 280.6 s (V-TaG), and 220.7 s (D-TaG). Thus, D-TaG was 21.4% faster than V-TaG and 25.1% faster than Voice-only. There was a slight improvement with practice with means of 282.0 s in block 2 and 249.5 s in block 6. (Block 1 was for training and was excluded from the data analysis.) See Figure 7. Using a repeated-measures ANOVA, the differences were deemed statistically significant for edit method ($F_{2,22}$ = 11.5, p = .0004) and block ($F_{4,44}$ = 2.67, p = .0447). The Voice-only edit method had the longest block completion time in 60% of the cases while D-TaG consistently was the fastest of the three approaches for error correction.

Selection Effort. The effort or the number of attempts to select an erroneous word was measured. The measure is a count per erroneous word, with a floor value of 1, implying a word was selected on the first attempt. To the extent selection effort was above 1, the measure reflects additional effort or frustration in selecting the erroneous word. As noted earlier, selecting erroneous words has been a challenge in most research and commercial applications for voice-based text entry.

The grand mean for selection effort was 1.32. By edit method, the means were 1.29 (Voice-only), 1.52 (V-TaG), and 1.15 (D-TaG). D-TaG, evidently, required 24.3% less selection effort than V-TaG and 10.9% less selection effort than Voice-only. There was improvement with practice with means of 1.42 in block 2 and falling to 1.24 in block 6. See Figure 8. The differences were statistically significant for edit method ($F_{2,22}$ = 19.1, p = .0001) and block ($F_{4,44}$ = 10.2, p = .0001). The V-TaG entry method had the highest selection effort in all blocks while D-TaG demonstrated the lowest selection effort in all blocks. The block-6 selection effort for D-TaG was 1.14, implying an additional selection about once for every seven erroneous words.

Subjective Feedback

A subjective feedback session was conducted to understand how participants perceived their interaction with the three edit methods. The goal was to understand the perceived task load using NASA TLX questionnaire [9] and the usability of the edit method using the System Usability Scale (SUS) [3]. We also included a custom questionnaire asking participants to subjectively rate the edit methods on accuracy, learnability, speed, and comfort.

The NASA TLX task load evaluation yielded means of 30.8 (Voice-only), 31.2 (D-TaG), and 50.9 (V-TaG). Although V-TaG had the highest score – indicating a higher task load compared to Voice-only and D-TaG – there were substantial differences among the participants with scores ranging from 29 to 75. A Friedman non-parametric test indicated the differences between the three edit methods were not statistically significant ($\chi^2(2) = 4.67, p = .097$). On interviewing participants, they mentioned that focusing on the erroneous word and then speaking "select" for triggering error correction was stressful.

Participants who did D-TaG before V-TaG were observed to wait and dwell on the error word. On asking why, most mentioned they forgot to give the "select" command as dwell selection was simple for them.

The System Usability Scale (SUS) evaluation was conducted to understand the overall usability of the edit methods. The scores were 81.0 (Voice-only), 80.2 (D-TaG), and 73.3 (V-TaG). The scores for Voice-only and D-TaG are quite good, placing them in the top 10% of SUS scores.[8] However, the differences were deemed not statistically significant using the Friedman test ($\chi^2(2) = 3.96, p = .138$). Participants expressed comfort in using D-TaG as they did not need to focus and say a command or say "map" to select an error.

The custom questionnaire was given to understand how participants perceived accuracy, learnability, speed, and comfort of

[8] https://measuringu.com/sus/

the edit methods. Responses were on a scale from 1 to 7, with higher scores preferred. Participants reported that the speed and accuracy of D-TaG made the experience of error correction simpler and easier than the other edit methods. Voice-only and D-TaG scored the same for learnability (6.6) and accuracy (5.7). D-TaG performed better on speed (6.0 vs. 5.3 vs. 4.7) and comfort (5.3 vs. 5 vs. 4.7). See Figure 9.

Image Description Task

A free text formation task was performed asking participants to describe images presented before them (Figure 5). A qualitative evaluation was performed based on their performance.

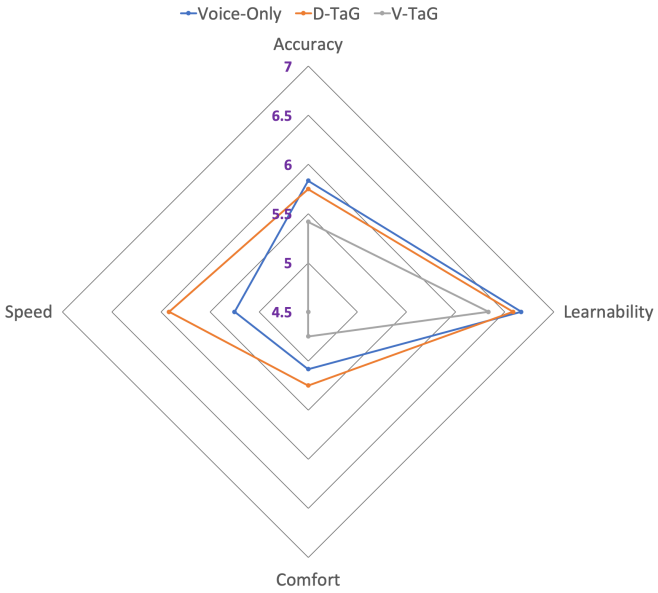


Figure 9: Read and Correct Task – average perceived performance on a 1-7 scale with higher scores preferred

Preference. At the end of the study, participants were asked to rank their preferences for the three edit methods. D-TaG emerged as the most preferred choice with 66.6% going in its favour. This was followed by Voice-only and finally V-TaG.

Comfort. As seen in Figure 10, Voice-only tops the list in comfort, followed by D-TaG and V-TaG. On asking participants the reason, they noted it was difficult to focus on an erroneous word while giving the command for selection with V-TaG. While all praised D-TaG, they raised issues with accidental selection of non-erroneous words. Voice trumps the list as it is precise even though the steps take longer.

Speed. D-TaG was perceived as the fastest edit method (see Figure 10) with most participants expressing comfort with the 0.8 second dwell time. However, when false triggering happened, they felt uncomfortable. One participant complained about the speed of erroneous word selection but proposed a hybrid approach that combined the voice and D-TaG approaches.

Accuracy. Selecting the erroneous word by gaze and confirming it by speaking "select" for the V-TaG edit method was difficult for some participants. This led V-TaG to the lowest perceived accuracy in comparison to the other edit methods. While they were comfortable with the selection in D-TaG, participants also appreciated the voice-only approach.

Usage intention. Most participants confirmed that they would like to use D-TaG over V-TaG. However, they also mentioned the possibility of using Voice-only in tandem with D-TaG.

Overall experience. Three participants expressed fatigue from using the "map" command with the Voice-only edit method. Some found it difficult to focus on the erroneous word while giving the "select" command. None reported fatigue with D-TaG even though some noted and did not like the Midas Touch issue.

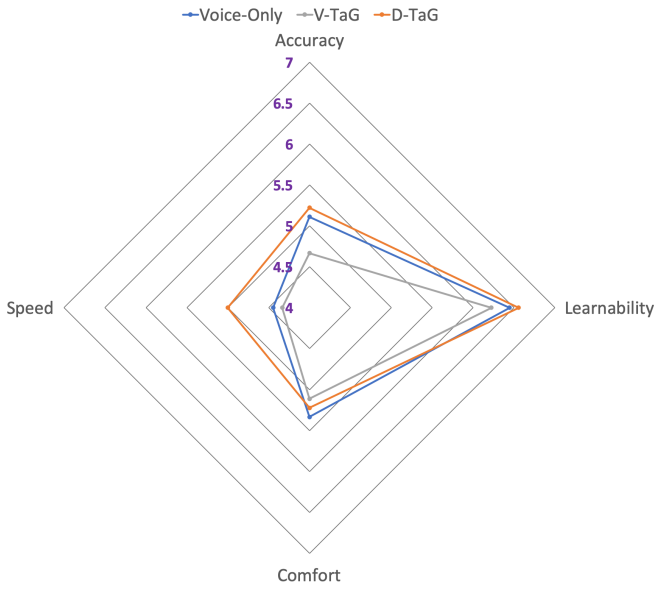


Figure 10: Image Description Task – perceived performance on a 1-7 scale with higher scores preferred

DISCUSSION

Discussions of the results below focus on (i) the use of gaze in a voice-controlled environment for improving selection and editing of text, (ii) the convenience of selecting erroneous words in presence of an additional modality, and (iii) use of a fall-back option when one modality fails to perform.

Objective measurements and subjective feedback favored the combination of voice and gaze for hands-free error correction. As observed in block completion time, D-TaG performed better than V-TaG. This is also supported by the perceived speed in the qualitative evaluation. One reason is the speed of gaze input for spatial exploration in the context of user interfaces. In this work, we also leveraged gaze in a voice-controlled hands-free environment to select textual errors in fewer attempts compared to other edit methods. This is seen in selection effort where D-TaG performed better than the V-TaG and Voice-only methods. Our approach overcomes the limitations of recognition errors in error selection, and thus also performed better in perceived comfort. These observations can be leveraged by designers for faster pointing and selection of items in hands-free applications. Applications for text entry and editing in head-mounted multimodal displays

or selection of interface elements in a multi-monitor system are areas where our approach has potential.

The questionaire responses show that the gaze-based D-TaG interface was considered comfortable compared to the V-TaG interface. The possible reason is the jitter in gaze movement while fixating on the incorrect word before uttering “select". Selection errors sometimes occurred and this forced the participant to repeat the selection process.

Interestingly, participants preferred the Voice-only approach over V-TaG. This is irrespective of the fact that Voice-only has more steps to error correction than V-TaG. This gives us insight for a fall-back mode which designers can leverage for a multimodal hands-free environment. Applications can primarily take advantage of the modalities available, but in case of eye strain or eye tracker drift, a voice-only approach is a fall-back method to complete the error correction task.

LIMITATIONS AND FUTURE WORK

The study presented here has some limitations. (1) Phonetically similar words were often incorrectly rendered (e.g., "little" vs. “Lidl”, "year" vs. “yeah”). Incorrect recognition increases the time for error correction and also creates a barrier for command recognition. For example, "select" was often recognized as “Sylhet” for three participants. For future work, this can be addressed by a self-learning approach where the system learns from mistakes corrected. (2) We used standalone eye trackers with traditional eye tracking challenges, such as calibration and drift. Some participants reported that they had to look slightly above or below the target word. This can be addressed by "on-the-fly" calibration [41, 34]. (3) Visual feedback for dwelling was not provided in this experiment. Future work would include visual feedback as used in many gaze-based selection methods [21, 14]. (4) Participants did not undergo extensive training. Future work could include more training to understand how far performance may improve beyond that shown in this evaluation. (5) The transcription API was not trained extensively to understand different phonetic variations of command used in our experiment by the participants. Future work could also focus on training the API for more robust recognition. (6) The text editing scenario only considered single-screen text. However, our approach supports editing of text that is longer than the height of the screen. Voice commands like “scroll up” or “scroll down” could be used along with gaze-based scrolling. This would help in jumping across pages to do error correction. (7) While the study focused on character-level error correction, complex text edits (grammatical errors, moving words or sentences) were not investigated. Future work could evaluate using voice and gaze for implementing such features.

This work provides directions to future applications involving voice and gaze for developers and designers. The subjective evaluation for the image description task was intended to understand if using voice and gaze for error correction could extend from a testing scenario to a more realistic scenario. A detailed evaluation of different use cases is planned for a near-future study.

CONCLUSION

Voice-based input offers a fast, hands-free approach for text insertion. We presented the design and evaluation of two versions of TaG (Talk-and-Gaze): D-TaG and V-TaG, two novel gaze-augmented voice-based text entry methods. Objective measures and subjective feedback for a *read and correct task* show D-TaG performed better than a Voice-only approach and V-TaG. Results also showed that D-TaG enables users to complete their task in the least number of attempts, thereby leading to lower cognitive load and higher usability scores. Our novel approach could be extended to different styles of text editing thereby expand the potential of voice and gaze for text-based interactions.

ACKNOWLEDGEMENT

This work has been performed as part of the MAMEM[9] project with funding from the European Union’s Horizon 2020 research and innovation program under grant agreement number 644780. We would like to thank Pooya Oladazimi and Tara Morovatdar (University of Koblenz-Landau) for their assistance during the evaluation phase and all the participants who took part in the evaluation studies and provided us with relevant feedback.

REFERENCES

[1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, and others. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *Proceedings of the 33rd International Conference on Machine Learning*. PMLR, New York, 173–182.

[2] T. R. Beelders and P. J. Blignaut. 2010. Using vision and voice to create a multimodal interface for Microsoft Word 2007. In *Proceedings of the 2010 Symposium on Eye Tracking Research & Applications (ETRA ’10)*. ACM, New York, 173–176. DOI: http://dx.doi.org/10.1145/1743666.1743709

[3] John Brooke and others. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[4] Emiliano Castellina, Fulvio Corno, and Paolo Pellegrino. 2008. Integrated speech and gaze control for realistic desktop environments. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications (ETRA ’08)*. ACM, New York, 79–82. DOI: http://dx.doi.org/10.1145/1344471.1344492

[5] Kevin Christian, Bill Kules, Ben Shneiderman, and Adel Youssef. 2000. *A comparison of voice controlled and mouse controlled web browsing*. Technical Report TR_2005-11. College Park, MD.

[6] Catalina Danis, Liam Comerford, Eric Janke, Ken Davies, Jackie De Vries, and Alex Bertrand. 1994. Storywriter: A speech oriented editor. In *Proceedings of the ACM SIGCHI Conference Companion on Human*

[9] http://www.mamem.eu

Factors in Computing Systems (CHI '94). ACM, New York, 277–278. DOI: http://dx.doi.org/10.1145/259963.260490

[7] C De Mauro, M Gori, M Maggini, and E Martinelli. 2001. *Easy access to graphical interfaces by voice mouse*. Technical Report. Università di Siena. Available from the author.

[8] Mark Dunlop, Emma Nicol, Andreas Komninos, Prima Dona, and Naveen Durga. 2016. Measuring inviscid text entry using image description tasks. In *CHI'16 Workshop on Inviscid Text Entry and Beyond*. ACM, New York. http://www.textentry.org/chi2016/9%20-%20Dunlop%20-%20Image%20Description%20Tasks.pdf

[9] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[10] Lode Hoste and Beat Signer. 2013. SpeeG2: A speech- and gesture-based interface for efficient controller-free text input. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*. Association for Computing Machinery, New York, 213–220. DOI: http://dx.doi.org/10.1145/2522848.2522861

[11] R Jakob. 1998. The use of eye movements in human-computer interaction techniques: What you look at is what you get. In *Readings in Intelligent User Interfaces*, W. Wahlster M. T. Maybury (Ed.). Morgan Kaufmann, San Francisco, 65–83.

[12] 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 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, New York, 568–575. DOI: http://dx.doi.org/10.1145/302979.303160

[13] Reo Kishi and Takahiro Hayashi. 2015. Effective gazewriting with support of text copy and paste. In *2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS)*. 125–130. DOI: http://dx.doi.org/10.1109/ICIS.2015.7166581

[14] Chandan Kumar, Raphael Menges, and Steffen Staab. 2016. Eye-controlled interfaces for multimedia interaction. In *IEEE MultiMedia*, Vol. 23. IEEE, New York, 6–13.

[15] Manu Kumar, Andreas Paepcke, Terry Winograd, and Terry Winograd. 2007. EyePoint: Practical pointing and selection using gaze and keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, 421–430. DOI: http://dx.doi.org/10.1145/1240624.1240692

[16] Päivi Majaranta. 2009. *Text entry by eye gaze*. PhD dissertation. http://urn.fi/urn:isbn:978-951-44-7787-4

[17] Bill Manaris and Alan Harkreader. 1998. SUITEKeys: A speech understanding interface for the motor-control challenged. In *Proceedings of the Third International ACM Conference on Assistive Technologies (Assets '98)*. ACM, New York, 108–115. DOI: http://dx.doi.org/10.1145/274497.274517

[18] Chandra Sekhar Mantravadi. 2009. *Adaptive multimodal integration of speech and gaze*. Ph.D. Dissertation. Rutgers University, New Brunswick, NJ.

[19] Arthur E McNair and Alex Waibel. 1994. Improving recognizer acceptance through robust, natural speech repair. In *Third International Conference on Spoken Language Processing (ICSLP '94)*. International Speech Communication Organization, Baixas, France, 1299–1302.

[20] Raphael Menges, Chandan Kumar, Daniel Müller, and Korok Sengupta. 2017. GazeTheWeb: A Gaze-Controlled Web Browser. In *Proceedings of the 14th Web for All Conference on The Future of Accessible Work (W4A '17)*. ACM, New York, NY, USA, Article 25, 2 pages. DOI: http://dx.doi.org/10.1145/3058555.3058582

[21] Raphael Menges, Chandan Kumar, and Steffen Staab. 2019. Improving User Experience of Eye Tracking-Based Interaction: Introspecting and Adapting Interfaces. *ACM Trans. Comput.-Hum. Interact.* 26, 6, Article Article 37 (Nov. 2019), 46 pages. DOI: http://dx.doi.org/10.1145/3338844

[22] Yoshiyuki Mihara, Etsuya Shibayama, and Shin Takahashi. 2005. The migratory cursor: Accurate speech-based cursor movement by moving multiple ghost cursors using non-verbal vocalizations. In *Proceedings of the 7th International ACM Conference on Computers and Accessibility (Assets '05)*. ACM, New York, 76–83. DOI: http://dx.doi.org/10.1145/1090785.1090801

[23] Sharon Oviatt. 1997. Multimodal interactive maps: Designing for human performance. *Human-Computer Interaction* 12, 1 (1997), 93–129.

[24] Sharon Oviatt. 2000. Taming recognition errors with a multimodal interface. *Commun. ACM* 43, 9 (2000), 45–45.

[25] Sharon Oviatt. 2003. Multimodal interfaces. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (2 ed.), J. A. Jacko A. Sears (Ed.). Vol. 14. Erlbaum, Mahwah, NJ, 286–304.

[26] Sharon Oviatt, Phil Cohen, Lizhong Wu, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, and others. 2000. Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions. *Human-Computer Interaction* 15, 4 (2000), 263–322. DOI:http://dx.doi.org/10.1207/S15327051HCI1504_1

[27] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98 – 125. DOI: http://dx.doi.org/https://doi.org/10.1016/j.inffus.2017.02.003

[28] Marco Porta and Alessia Ravelli. 2009. WeyeB, an eye-controlled Web browser for hands-free navigation. In *2009 2nd Conference on Human System Interactions*. 210–215. DOI: http://dx.doi.org/10.1109/HSI.2009.5090980

[29] Matheus Vieira Portela and David Rozado. 2014. Gaze enhanced speech recognition for truly hands-free and efficient text input during HCI. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design (OzCHI '14)*. ACM, New York, 426–429. DOI: http://dx.doi.org/10.1145/2686612.2686679

[30] Kari-Jouko Räihä and Saila Ovaska. 2012. An exploratory study of eye typing fundamentals: Dwell time, text entry rate, errors, and workload. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, 3001–3010. DOI: http://dx.doi.org/10.1145/2207676.2208711

[31] David B Roe, Jay G Wilpon, and others (Eds.). 1994. *Voice communication between humans and machines*. National Academies Press, Washington, DC.

[32] 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. In *Proceedings of the ACM Conference on Interactive, Mobile, Wearable and Ubiquitous Technologies*. ACM, New York, 159:1–159:23. DOI:http://dx.doi.org/10.1145/3161187

[33] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. "Your word is my command": Google search by voice: A case study. In *Advances in speech recognition*, A. Neustein (Ed.). Springer, Boston, 61–90. DOI: http://dx.doi.org/10.1007/978-1-4419-5951-5

[34] Simon Schenk, Marc Dreiser, Gerhard Rigoll, and Michael Dorr. 2017. GazeEverywhere: Enabling Gaze-only User Interaction on an Unmodified Desktop PC in Everyday Scenarios. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3034–3044. DOI: http://dx.doi.org/10.1145/3025453.3025455

[35] Andrew Sears, Jinhuan Feng, Kwesi Oseitutu, and Claire-Marie Karat. 2003. Hands-free, speech-based navigation during dictation: difficulties, consequences, and solutions. *Human-Computer Interaction* 18, 3 (2003), 229–257. DOI: http://dx.doi.org/10.1207/S15327051HCI1803_2

[36] Korok Sengupta, Min Ke, Raphael Menges, Chandan Kumar, and Steffen Staab. 2018. Hands-free web browsing: enriching the user experience with gaze and voice modality. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. ACM, New York, 88. DOI: http://dx.doi.org/10.1145/3204493.3208338

[37] Korok Sengupta, Raphael Menges, Chandan Kumar, and Steffen Staab. 2017. GazeTheKey: Interactive keys to integrate word predictions for gaze-based text entry. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces Companion (IUI '17 Companion)*. ACM, New York, 121–124. DOI: http://dx.doi.org/10.1145/3030024.3038259

[38] Korok Sengupta, Raphael Menges, Chandan Kumar, and Steffen Staab. 2019. Impact of Variable Positioning of Text Prediction in Gaze-Based Text Entry. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research Applications (ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article Article 74, 9 pages. DOI: http://dx.doi.org/10.1145/3317956.3318152

[39] Shyamli Sindhwani, Christof Lutteroth, and Gerald Weber. 2019. ReType: Quick text editing with keyboard and gaze. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, 203. DOI: http://dx.doi.org/10.1145/3290605.3300433

[40] Bernhard Suhm, Brad Myers, and Alex Waibel. 2001. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI)* 8, 1 (2001), 60–98. DOI: http://dx.doi.org/10.1145/371127.371166

[41] Oleg Špakov and Darius Miniotas. 2005. Gaze-based selection of standard-size menu items. In *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI '05)*. ACM, New York, 124–128. DOI: http://dx.doi.org/10.1145/1088463.1088486

[42] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in speech user interfaces. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*. ACM, New York, 369–376.