IMPROVING USABILITY OF GAZE AND VOICE BASED TEXT ENTRY SYSTEMS

Von der Fakultät Informatik, Elektrotechnik und Informationstechnik der Universität Stuttgart zur Erlangung der Würde eines Doktors der Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

Vorgelegt von

KOROK SENGUPTA

aus Bally, West Bengal, India

Hauptberichter:	Prof. Dr. Steffen Staab
Mitberichter:	Prof. Dr. Cosmin Munteanu

Tag der mündlichen Prüfung: 17.05.2023

Institut für Parallele und Verteilte Systeme (IPVS) der Universität Stuttgart

2023

DEDICATIONS

This thesis is dedicated to Late Mrs. Renu Sengupta, Late Mr. Sachin Vankhalas, and Amaan.

ABSTRACT

Text entry is an essential interaction in the digital environment. We use it to access content on the web, share thoughts and information, access navigational instructions, and use it for our interests. The conventional approach of inserting text with a keyboard and mouse has been with hands and fingers. This approach limits accessibility and often restricts interaction in circumstances of situational impairment. To overcome such challenges, we investigate alternative modalities like gaze and voice.

In the first part of the thesis, we investigate text entry by gaze and report our work on designing and implementing a new keyboard (GazeTheKey) that enhances the use of word predictions for improved efficiency. We further evaluate this design by comparing it with a traditional gaze-based keyboard and another design aiming to bring the word predictions closer to the visual fovea. This investigation and evaluation contribute to exploring an on-screen predictive keyboard design and present an enhanced layout for gaze-based text entry as a part of this thesis. We evaluate EEG signals while users use different keyboard designs to understand the cognitive load associated with each user experience. The result conclusively shows that our design performs better than the traditional gaze-based keyboards. This investigation and evaluation advance the text entry domain of research by contributing the findings that EEG signals are an efficient metric in understanding instantaneous cognitive load, which traditional approaches like NASA TLX or SUS fail to capture. Our novel design prominently demonstrates the schematic nature of word prediction selection, particularly when predictions are strategically positioned close to the keyboard's letter layout. However, the inherent challenges associated with gaze-only text entry remain, and to overcome that, we approach the direction of multimodal text entry - using voice and gaze together.

The second part of this thesis contributes to expanding the multimodal text entry paradigm by using voice and gaze in parallel and designing an approach called" Talk-and-Gaze(TAG)" for text revision. This approach reduces the limitations of the gaze-only text entry approach (selecting letters or words linearly). It delivers an improved performance of inserting text with the help of voice input for improved usability. Our novel multimodal interaction approach works parallel (voice and gaze input) and provides a fall-back mechanism when one modality fails to deliver.

ZUSAMMENFASSUNG

Die Texteingabe ist eine wesentliche Interaktionsform in der digitalen Welt. Wir nutzen sie, um auf Inhalte im Internet zuzugreifen, Gedanken und Informationen auszutauschen, Navigationsanweisungen abzurufen und auch für unsere persönlichen Interessen. Der herkömmliche Ansatz für die Eingabe von Text mit Tastatur und Maus ist die Eingabe mit Händen und Fingern. Dieser Ansatz limitiert die Zugänglichkeit und schränkt die Interaktion in Situationen, in denen eine Beeinträchtigung vorliegt, häufig ein. Um solche Herausforderungen zu überwinden, untersuchen wir alternative Modalitäten wie Blicke und Stimme.

Im ersten Teil der Arbeit untersuchen wir Texteingabe mit Hilfe des Blicks und stellen unsere Entwicklung und Implementierung einer neuen Tastatur (GazeTheKey) vor, die die Verwendung von Wortvorhersagen steigt, um eine verbesserte Benutzererfahrung zu ermöglichen. Wir evaluieren dieses Design, indem wir es mit einer traditionellen blickbasierten Tastatur und einem weiteren Design vergleichen, das darauf abzielt, die Wortvorhersagen näher an die visuelle Fovea zu bringen. Um die kognitive Belastung der jeweiligen Benutzererfahrungen zu verstehen, werten wir EEG-Signale aus, während die Benutzer die verschiedenen Tastaturdesigns verwenden. Das Ergebnis zeigt eindeutig, dass unser Design besser abschneidet als die traditionellen blickbasierten Tastaturen. Dieser Teil der Arbeit bringt die Forschung im Bereich der Texteingabe voran, indem er EEG-Signale als effiziente Metrik für das Verständnis der momentanen kognitiven Belastung etabliert, die mit traditionellen Ansätzen wie NASA TLX oder SUS nicht erfasst werden kann. Unser neuartiges Design zeigt auch, wie einfach es für Benutzer war, Wortvorhersagen auszuwählen, wenn ihre Präsenz erhöht und in der Nähe der Buchstaben der Tastatur platziert wurde. Die inherenten Herausforderungen, die bei der Texteingabe rein mit Hilfe des Blickes verbunden sind, bleiben jedoch bestehen, und um diese zu überwinden, gehen wir den Weg der multimodalen Texteingabe - mit Stimme und Blicken zusammen.

Der zweite Teil der Arbeit befasst sich mit dem Design und der Entwicklung einer multimodalen Interaktion aus Blick und Stimme (”Talk-and-Gaze: TaG”) für das Szenario der Textrevision. Dieser Ansatz reduziert die Einschränkungen der Texteingabe rein mit Hilfe des Blickes und bietet eine verbesserte Leistung und Benutzererfahrung. Unser neuartiger multimodaler Interaktionsansatz arbeitet parallel und bietet einen Rückfallmechanismus, wenn eine Modalität nicht funktioniert.

ACKNOWLEDGMENTS

The road to wanting something from life is always guided by people who believe in the call and help us achieve that. My research would not have seen fruition if the people at WeST, University of Koblenz-Landau and Analytic Computing, University of Stuttgart had not helped me in my journey.

I would primarily like to thank my supervisor, Prof. Dr. Steffen Staab, and my ex-colleague Dr. Chandan Kumar for their help, guidance, and support in training me on this journey. They have co-authored almost all my publications and enriched them with their inputs and feedback. I would also like to thank Dr. Sayan Sarcar for his immense help and guidance at different stages of our multiple publications.

I take this opportunity also to thank Dr. Raphael Menges and Dr. Jun Sun for their support and collaboration in different research works that I had the opportunity to collaborate with them. Furthermore, I would also thank Dominik Brosius for being a colleague with whom I could share all my insecurities and challenges.

A big shout out to all the fantastic researchers with whom I collaborated on this journey.

I want to thank Tara Morovatdar, Prantik Goswami, Sabin Bhattarai, Pooya Oladazimi, and Min Ke for helping me extend my research work through their contributions.

My research was funded mainly by the MAMEM project (European Union's Horizon 2020, grant agreement number 644780), and I would extend my acknowledgment to the entire team.

I would also like to thank Silke Werger and Sabine Hülstrunk for their support at every stage of my stay in Koblenz. Finally, I would like to thank my wife, Anindita Mandal, for her countless sacrifices during this journey and for being my constant pillar of support through thick and thin. Without her, this journey would not have taken place. My parents, sister and in-laws - life would never be the same without you and thank you for being a part of this journey.

CONTENTS

1

INTRODUCTION

Text entry is an essential interaction with digital systems. It is one of the most common tasks in ICT (Information and Communication Technology) devices like desktops, laptops, and even hand-held devices like cell phones or tablets. The task of text entry is prevalent in our professional interactions, personal and social ones. Variations in form factors like cell phones, smartwatches, or tablets are crucial in designing the user experience of text entry systems without compromising efficiency and effort. The text entry task is vast and primarily restricted to hand-based interactions via keyboard-mouse or by touching a virtual keyboard. Insertion of text is mainly required for logging in to a system (let us say, one's personal computer), searching for content (perhaps a particular file), creating text-based content (writing a report on a text editor like Microsoft Word), searching the Web (*Google Search*, *DuckDuckGo* and many more), registering on websites (let us say, on New York Times), composing emails or chat-based communications. Nevertheless, the fundamental principle of text entry is the same for all these processes. It involves the primary task of scanning and selecting letters and symbols from the keyboard to form text like words, passwords, proper nouns, etc. As this process continues, the user must read through the collected inputs to validate the correctness. However, the selection techniques can vary: from a traditional selection of keys on the keyboard by a single finger or multiple fingers to using an on-screen keyboard and selecting the letters on the traditional keyboard with a mouse, stylus [52], or joystick [186] or rotary controls [195].

The performance of text entry systems is often measured by how fast one can insert characters to form words. Speeds of 100 or more words per minute (WPM) on physical keyboards with multiple fingers [150, 184, 185] to 60 WPM on mobile tactile keyboards using two thumbs [29] to 30-40 WPM on virtual keyboards for touch screen mobile devices have been recorded [12]. WPM as low as 10-20 WPM have been recorded on smartphone watches [4, 86]. As the form factor (screen size here) decreases, the WPMs also decrease. However, investigating comfort and cognitive load associated with such devices when performing text entry tasks is limited.

Researchers altered different parameters and input methodologies to understand if that could lead to efficient text entry. Switches were used to overcome keyboard-mouse challenges for users suffering from motor control ailments. Ntoa et al. [125] list an exhaustive study of switches they investigated. Work on muscle contractions [46], eye blinks [8, 156] or eye movements [103] has also been undertaken for the text entry process. While input modality was widely investigated, language models [181] and word predictions [137] also played a crucial role in improving text entry. Efficient word predictions reduced keystrokes and helped users complete words faster. Keyboard designs were also investigated to understand if grouping keys or layouts could improve the text entry experience. The main goal has been to improve text entry speed in all these cases.

Amyotrophic lateral sclerosis: A neurodegenerative disease that leads to loss of motor control.

While the hands and fingers have mostly been the preferred mode of interaction on a keyboard for text input on digital systems, this predominant mode of text entry limits access to people suffering from motor control ailments or in scenarios that lead to situational impairment. Challenges like deteriorated finger dexterity, ALS , and muscular dystrophy often limit the motor controls of patients, thus leading them to digital seclusion since the predominant method of text entry requires hands and fingers to work well for digital interaction and communication [80]. Older adults with limited speed and agility may also find the keyboard's efficient use cumbersome and enter text letter-by-letter after searching those keys on the keyboard. This often leads to a higher cognitive load and poor user experience, pushing this population to reject the use of technology [180].

Gaze is the intersection of the line of sight and the screen

Alternative modalities like gaze and voice are investigated widely to understand their potential for regular usage as an additional communication channel beyond the traditional keyboard-mouse combination or touch. Projects like MAMEM[1], GazeTheWeb[2], GazeMining[3], Euphonia[4] focus on investigating how alternative modalities can be used for interaction on digital content. Findings from such projects and the growing market demands for alternative modalities (see Figure 1.1) motivated us to conduct our investigation – even for able-bodied individuals. While these projects focus on accessibility issues, the interaction and technology of such projects can also be used to overcome situational impairments in interaction - for example, when hands are busy performing some action, voice control can take over. Gaze and voice have also been investigated to be used together as multimodal inputs. Beelders et al. [16] brought gaze and voice together for Microsoft Word 97 for transcribing content. Projects like Microsoft's

1 https://www.mamem.eu/
2 https://west.uni-koblenz.de/research/gazetheweb
3 http://gazemining.de/
4 https://sites.research.google/euphonia/about/

HoloLens[5] is one example where all-natural modalities come together for enhanced interaction.

One of the most natural and intuitive approaches to understanding user attention has been to observe the human gaze. Initially, it was used to comprehend where users were looking and how design could be enhanced. The advancement and proliferation of eye trackers have transformed the gaze into more than just an input for understanding user attention. Today, gaze signals are utilized as input methods for text input and web navigation, aiding individuals accessing information previously digitally excluded due to motor control challenges. The growth of the eye-tracking industry also raises optimistic expectations regarding integrating such hardware into mainstream ICT devices (source: gaming.tobii.com/products/laptops). This would facilitate not only able-bodied individuals but also those seeking improved accessibility in digital design for greater inclusivity.

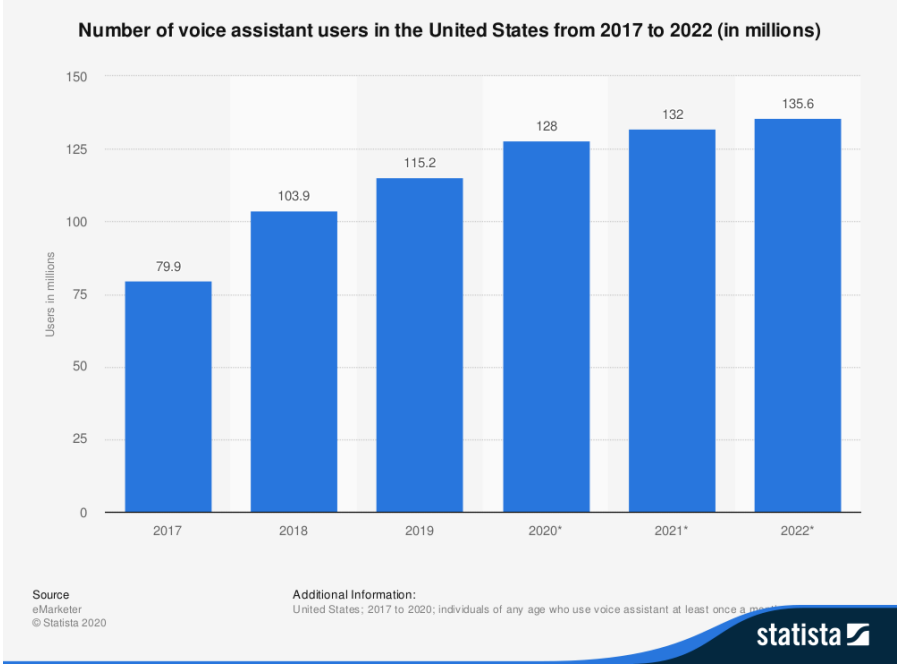
Voice, another natural modality like gaze, is rapidly accepted as a command-based input for information exchange via intelligent voice-enabled assistants like Google Home, Siri, Alexa, or Cortana. One of the main reasons for the acceptability of voice input is the inherent nature of the human voice to express intentions explicitly. In addition, voice is used in verbal and non-verbal interactions to access information from the web—the advancements in natural language processing help transcribe text with fewer errors than before.

In this thesis, we investigate and evaluate an on-screen gaze-based text entry system. Our investigation to understand the stress associated with gaze-based text entry led to the need for measuring and understanding instantaneous cognitive load. The second part of the research sheds light on multimodal interaction for text entry systems, where we built a voice and gaze-based interaction approach for text revision scenarios. This investigation aimed to understand the efficacy of the additional modality that increased the communication bandwidth.

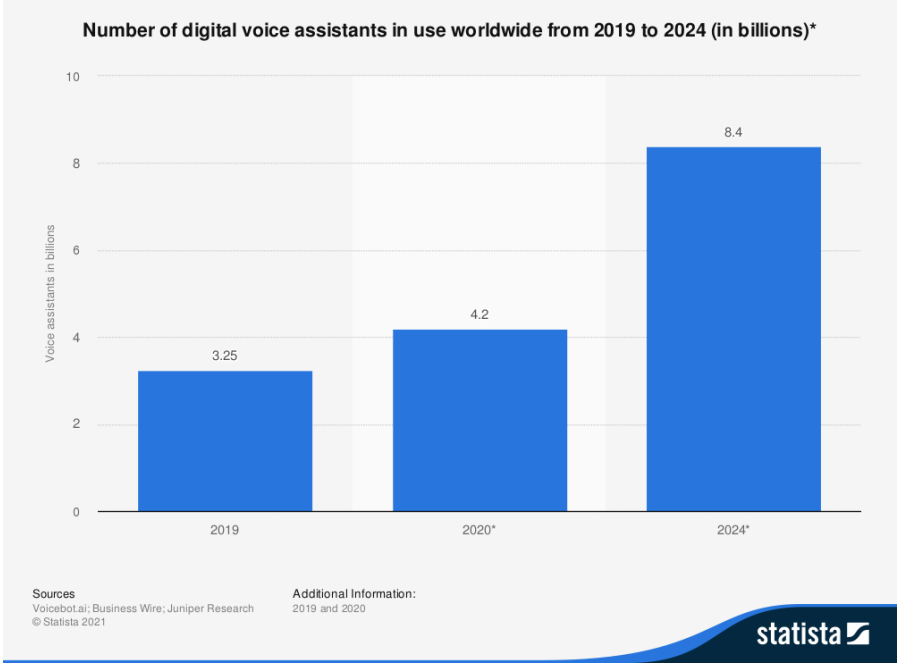
1.1 UNDERSTANDING TEXT ENTRY PROCESS

Text entry is an iterative process involving perceptive, cognitive, and motor skills to form words to build the desired sentence. The design space for methods of text entry is expansive. It includes decisions about the variations in interfaces (touch screens, size, and shape of buttons, position of interaction elements), interaction techniques for text input (gestures, feedback, modalities), use of intelligent methods

5 https://www.microsoft.com/en-us/hololens



(a)



(b)

Figure 1.1: Bar charts representing the growth and adoption of voice-based interaction devices (a)The bar chart shows the growth of the use of voice assistant users in the United States from 2017 - 2022, signifying the growth of adoption of such alternative modalities in our households and other places.; (b)The bar char represents the number of voice assistants sold worldwide with a projection of 8.4 billion units in 2024.

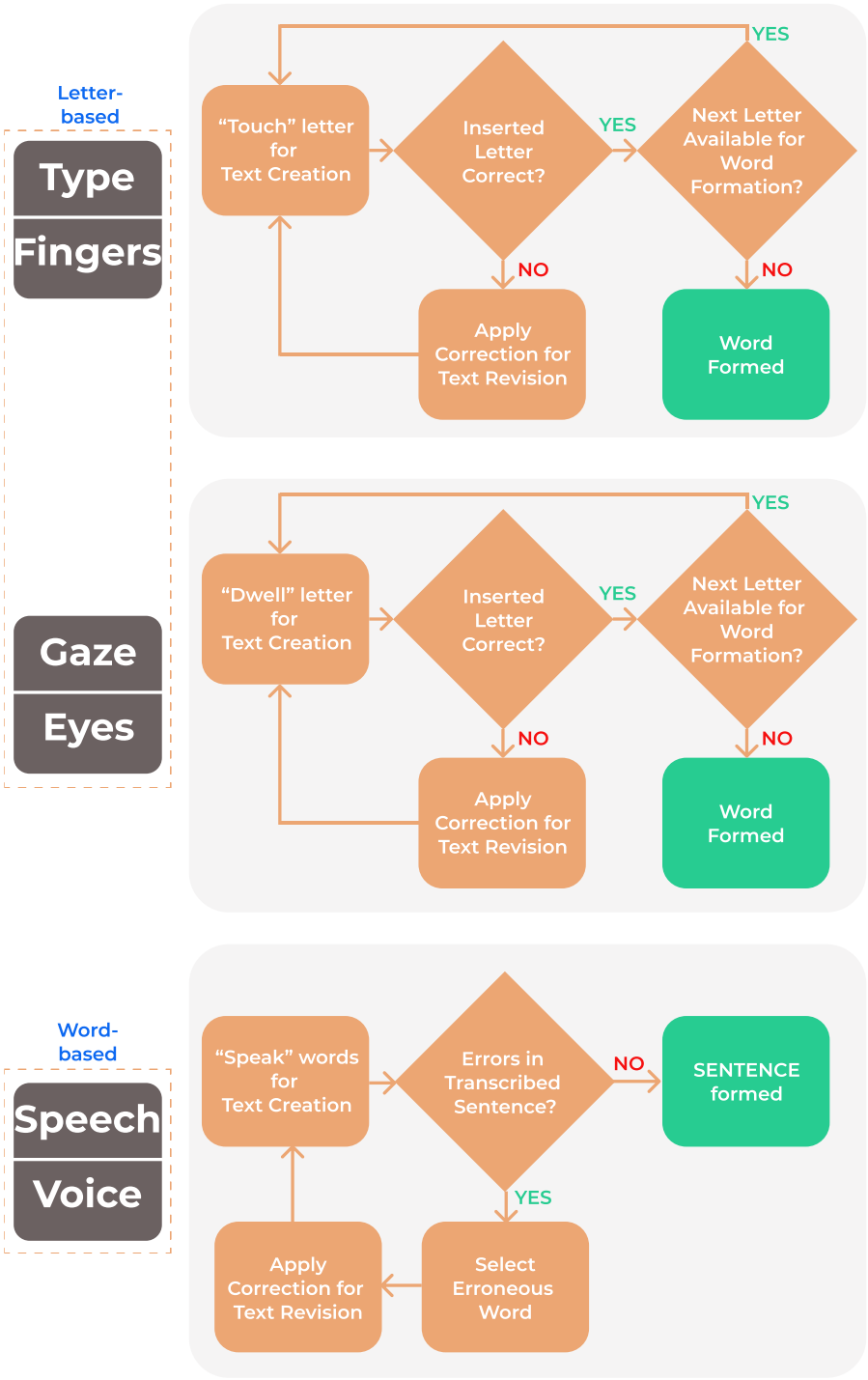


Figure 1.2: Conceptual diagram representing the process of text entry via fingers, eyes, and voice. Text entry by fingers and gaze is mostly letter-based (however, the addition of the word prediction changes the classification and makes them hybrid.) On the other hand, voice input is word-based, where the speech-to-text engine transcribes the spoken words into sentences. We see when the text creation and revision occur in the three distinct blocks.

(intelligent word prediction and auto-correction), and many more. Thus, it is important to understand the process of text entry. We have split the process into two parts for our research: (i) text creation and (ii) text revision. The text creation process is split into two parts: (i) letter-based creation and (ii) word-based creation. *Figure 1.2* gives us a conceptual representation of the process of text entry for the traditional hand-based approach along with the two alternative modalities in the discussion here: gaze and voice.

While text creation is in progress, error(s) can occur. It is mainly taken care of by developing the text revision mechanism where error correction occurs. Most error correction mechanisms involve the following steps: (i) navigating to the error location, (ii) removing the error, and finally (iii) inserting the correct entry. Regular error correction mechanisms often add to the cost of text entry, thus minimizing the savings in time and effort for an efficient text creation mechanism.

1.1.1 *Text Creation*

Text creation is a complex process because the person conceptualizes the sentences to be formed and deconstructs them into words and further into letters to produce an error-free entry. While this process is in progress, the fingers navigate to the exact keys on the keyboard (virtual or digital) that the mind has generated to record the letter entries, thus forming words and sentences. If an error is detected, an error correction process is initiated, which will be discussed in the *Text Revision* subsection. The creation process becomes comparatively simpler when one talks and technology takes over and helps transcribe the speech to text. However, the errors generated from this speech-to-text transcription must also be corrected. Thus, the text revision section again plays a vital role in understanding the complete text entry process.

The traditional method of using one's hands with a writing instrument such as pen and paper employs a similar approach. Nevertheless, to address particular challenges encountered by users within a digital context, researchers have delved into various aspects, including keyboard designs, as evidenced in Panwar's work [132], enhancements in word prediction capabilities [50], and the exploration of diverse input modalities [43, 113, 142]. In instances where the utility of the keyboard-mouse combination was found to be obstructive, researchers have concentrated on investigating alternative modalities [78], the development of optimized keyboard layouts [132], and the efficient utilization of word prediction technologies [98].

The central focus revolved around how text creation was accomplished in all these instances. Here, we delve deeper into the limitations associated with these approaches, including a steeper learning curve compared to the conventional keyboard-mouse design and interaction. Additional challenges encompass incorrect word predictions, inadvertent selection of erroneous predictions, instances akin to the "Midas Touch" phenomenon [95], or recognition errors stemming from voice-based systems. These issues lead to heightened task overload and diminish the efficiency these approaches aimed to introduce into efficient and user-friendly text entry systems. Research has demonstrated that deploying multimodal systems can overcome some of these challenges and enhance the overall usability of such systems.

1.1.2 *Text Revision*

While text creation is in progress, users scan and check the letters/-words for error correction. If an error is detected, the text revision process involves the following steps: (i) navigate to the erroneous word/letter, (ii) select the position (iii) delete/replace the letters or words. The schematic presentation of text revision can be seen in Figure 1.3.

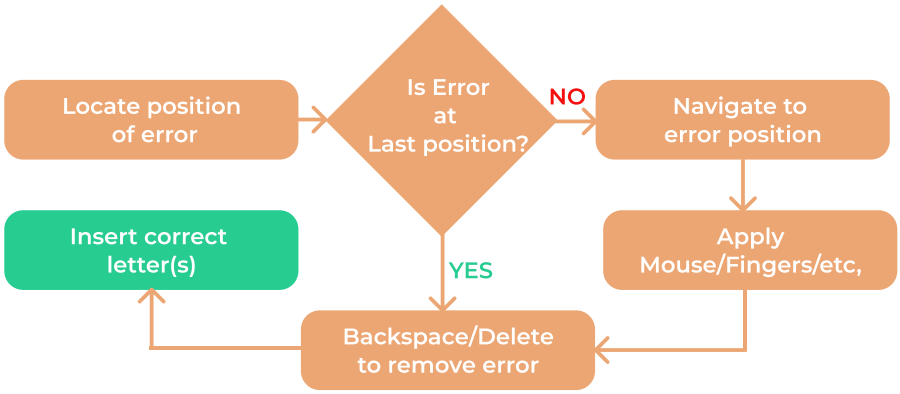


Figure 1.3: Schematic diagram representing the process of text revision. The generic process starts with the observation of the error and its location. Once we know there is an error, we can apply corrective actions immediately if it is currently at the last position. However, navigating to the error changes slightly when the error is somewhere between the sentence construction and the written paragraph. Then, navigating to the error forms a major point of interaction. This is where gaze as an input modality is fast while voice is not.

To further improve the process of text revision, prediction engines often help replace entire words based on the prediction logic. Re-

prediction engines are responsible for predicting the most suitable word based on the letter we type and/or the word(s) typed before.

searchers have also investigated the selection and replacement of erroneous words for voice-based systems by re-speaking them [176].

1.2 RESEARCH METHODS AND GOALS

This thesis explores input modality combinations to enhance the digital text entry experience. Building upon prior research in text entry, this thesis seeks to gain insights into the potential enhancements that hands-free modalities, such as voice and gaze, can offer to the text entry process. We have conducted a comprehensive investigation into the design space, on-screen keyboard functionality, input modalities, and their various combinations, all aimed at elevating the quality of the text entry experience.

This exploratory study has given rise to two fundamental research questions (RQ), each accompanied by three distinct research subquestions.

RQ1: How to improve the usability of a word prediction enabled gaze-based keyboard?

Usability primarily refers to the ease with which users can interact with a system, encompassing elements such as learnability, efficiency, memorability, error frequency, and satisfaction [122]. Performance is more quantitatively oriented, focusing on how effectively and swiftly users can complete tasks using the system [23]. The overlap between these two constructs becomes evident when considering that enhanced usability often leads to improved performance. However, it is crucial to recognize that high performance does not always equate to high usability. We need performance-centric measurements and user qualitative feedback to understand usability from a broader perspective. Therefore, in HCI research and practice, it is imperative to comprehensively evaluate both usability and performance, acknowledging their interdependence while appreciating their contributions to the overall user experience. Thus, for our first research question, we wanted to know how to improve the usability of a predictive on-screen keyboard for a Gaze-based text entry system.

Further exploration of this research question has given rise to three distinct research sub-questions, each focusing on issues impacting system usability, strategies to surmount these challenges, and the

evaluation process aimed at comprehending system usability and user experience.

RQ1.1: *What impacts the usability of an on-screen gaze-based keyboard?*
To improve the usability of a gaze-based system, we first have to investigate and understand the challenges of text entry in accessing the information. This question helps us understand existing problems like Midas Touch, drift, and fatigue, develop the possible solutions, and overcome them for an improved usability experience for on-screen gaze-based text entry systems.

RQ1.2: *How can we improve the usability issues of gaze-based text entry systems?*
To overcome the challenges discussed above, what are the different approaches we can take to improve the usability of on-screen keyboard text entry in a gaze-based system are discussed here. We explored the dwell-time-based method and introduced an improved approach to take advantage of the design of gaze-based keyboards for improved usability.

RQ1.3: *How to evaluate and understand the factors impacting usability of on-screen gaze-based keyboards?*
The success of any approach to overcome challenges can be understood by evaluating the system and taking user feedback. It is essential to understand that objective and subjective measures often align but not always. We explored an innovative metric to understand stress when using gaze-based keyboards.

RQ2: How to improve the usability of hands-free text entry with voice by integrating gaze?

Voice was another natural input modality that expressed intentions, and advancement in S2T (speech-to-text) technology has improved its recognition capacity. However, voice input comes with challenges, leading us to our RQ2, where we investigate voice and gaze for improved text entry.

We explored RQ2 further, leading to three more sub-questions for RQ2. These challenges focus on current problems, our approaches to circumvent them, and an evaluation to understand how effective our approach has been.

RQ2.1: *What impacts users in integrating voice as a primary modality for text entry?*

While voice offers a fast solution for hands-free text entry, it comes with challenges and inherent limitations, like incorrect recognition. Voice input works as a solitary modality and can often be combined with others in a multimodal setup for enhanced interaction. We investigated the users' challenges using voice-based text entry and developed solutions to overcome them.

RQ2.2: *How can we improve the usability issues of voice-based text entry systems in multimodal context?*
Voice modality faces the limitation of incorrect recognition and navigation in space. This section explores the different approaches for voice as an input modality in text entry and revision scenarios by exploring the multimodal setup of combining voice with gaze. We use voice input for text entry, but we use gaze to navigate to an erroneous word. This overcomes the challenge of repeatedly using voice commands to reach a specific location and takes advantage of fast eye movements.

RQ2.3: *How to evaluate and understand the factors impacting the usability of voice-based text entry systems?*
Literature and previous work provide a pathway for understanding how voice-based systems could be evaluated. However, our approach to overcoming the challenges uses a modality combination. Thus, objective and subjective evaluation is crucial in understanding the system's usability.

1.3 RESEARCH CONTRIBUTIONS

This thesis explores the combination of input modalities in text entry and revision. In line with the research questions stated above, the primary outcomes of this work are summarized in the following points:

C1 This thesis contributes to exploring an on-screen predictive keyboard design and presents an enhanced layout for gaze-based text entry.

To improve the text creation process, several works have focused on improving the design of the on-screen keyboard and, thus, created layouts that were heavily deviating from the traditional layout. We were inspired by these works to design a keyboard layout that retained the traditional layout pattern (thus reducing the learning curve) but utilized word

predictions' impact on such on-screen keyboards. This improvement allowed us to add more predictions, thus making it easier for the user to interact with the predicted word with minimized scanning distance. Contribution *C1 maps to the RQ1*, which aimed to improve the usability of word prediction enabled gaze-keyboard.

C2 The thesis contributes to the introduction of measuring brain signals for understanding instantaneous cognitive load and presenting it as a metric for understanding user experience.

The challenge of usability analysis has always focused on understanding how people feel when using the system. This information can often be biased based on performance, association, and other factors that affect the individual. For a scenario like text entry, such factors often play a key role in understanding improvements. NASA TLX often measures cognitive load [57]. However, we introduced the idea of measuring actual brain signals while the experiment is in progress to understand temporal cognitive load. This approach gave us an additional layer of information that had not been accessed before. Contribution *C2 also maps to R1*, as through this process, we can understand our designs' impact from a cognitive load perspective.

C3 The thesis expands the multimodal text entry paradigm by using voice and gaze in parallel and designing an approach called "Talk-and-Gaze(TAG)."

We successfully utilized the strength of voice and gaze in parallel for an efficient text revision scenario. We utilized the strength of speech-to-text models for faster and more efficient text entry and also used the super-fast movement of gaze for pointing to the designated location for text revision. Our approach used both the strengths of gaze and voice for an improved hands-free text editing scenario. Contribution *C3 maps to RQ2* where we investigate the potential of voice to work with gaze as a multimodal system for an efficient text entry process.

1.4 SUPPORTING PUBLICATIONS

The following publications support this thesis:

P1 **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). Association for Computing Machinery, New York, NY, USA, 121–124. DOI: https://doi.org/10.1145/3030024.3038259

P2 **Korok Sengupta**, Jun Sun, Raphael Menges, Chandan Kumar and Steffen Staab. "Analyzing the Impact of Cognitive Load in Evaluating Gaze-Based Typing," 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS '17), Thessaloniki, 2017, pp. 787-792, DOI: https://doi.org/10.1109/CBMS.2017.134.

P3 **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 74, 1–9. DOI: https://doi.org/10.1145/3317956.3318152

P4 **Korok Sengupta**, Sabin Bhattarai, Sayan Sarcar, I. Scott MacKenzie, and Steffen Staab. 2020. "Leveraging Error Correction in Voice-based Text Entry by Talk-and-Gaze". In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–11. DOI: https://doi.org/10.1145/3313831.3376579

1.5 THESIS OUTLINE

The work presented in this thesis is based on the publications as listed in the previous section. The thesis as a whole constitutes understanding text entry from the point of hands-free interaction and enabling readers to understand generic domain-specific challenges that need to be overcome for improved usability and user experience. The remainder of the document is structured as follows:

- Chapter 2 provides a historical account of text entry using keyboards and how it evolved from typewriters to modern-day keyboards for text entry. This chapter highlights the three key components of any text entry research document: Keyboard layout, chosen input modality, and finally, evaluation metrics.

- Chapter 3 extends the modality perspective and describes the alternative modalities investigated for text entry scenarios. This chapter primarily focuses on the plethora of work done for exploring gaze and voice modalities for text entry purposes. However, other modality sources like BCI and switches have also been described.

- Chapter 4 is one of the core chapters of this thesis that showcases the design investigation and challenges associated with on-screen gaze-based keyboards. It documents our approach to designing and developing a new and improved on-screen keyboard for gaze-based text entry: *GazeTheKey* [P1]. It also explores the impact of positioning word predictions for on-screen gaze-based keyboards [P3]. The chapter answers all the research questions associated with RQ1.

- Chapter 5 further explores the findings of Chapter 5 and aims to expand the evaluation metrics for gaze-based text entry systems. It presents a novel evaluation metric [**P2**] previously used only as subjective feedback. This chapter answers RQ1.3 from RQ1.

- Chapter 6 expands the envelope of hands-free text entry further by addressing the challenges of gaze-based text entry by adding voice as the primary modality of interaction and keeping gaze as the secondary for navigation [**P4**]. The chapter answers all the research questions associated with RQ2.

- Chapter 7 concludes this thesis with a conclusion that highlights the contributions of this thesis to the scientific community. It also discusses the limitations and future work planned beyond this thesis's thresholds.

2

FOUNDATIONS

In this chapter, we introduce the essential concepts of text entry that form the basis of our study. We begin by exploring the history of text entry methods, tracing their evolution over time in Section 2.1;. This sets the stage for a detailed examination of text entry via keyboard in Section 2.2, where we dissect the various parameters influencing this process. Subsections include an analysis of keyboard layouts in 2.2.1, an overview of different input modalities in 2.2.2, and a discussion on the performance metrics critical to text entry efficiency in 2.2.3.

2.1 HISTORY OF TEXT ENTRY

Communication via a textual medium is an integral part of our lives. From scribbles to penning down thoughts, the "*writing*" process has ensured socio-cultural development and storage of ideas on a tangible medium. Writing enabled us to express our emotions, concerns, and thoughts and be stored and transferred across long distances.

The earliest signs of writing and storing information can be found around 3300 BC [1] in the form of *cuneiforms*. That later evolved into language based on sounds by the Sumerians. In all these cases, writing acted as a tool for storing information and passing it on to the next generation. It took nearly 160 years, from 1714 to 1874, for the typewriter to gain significant recognition as an effective writing tool. It took another 100 years for such tools to gain acceptance and popularity in increasing the text entry speed for storing and sending information. Scientists and engineers developed and improved the writing tools' shape, mechanics, and layout to increase writing or text entry efficiency and usability.

Beyond the 1980s, the rise of personal computers slowly fazed out the use of typewriters. The computers had additional features that made text entry and information storage much easier than typewriters. The availability of commercial internet service also increased the usage of

1 https://www.newscientist.com/article/mg23230990-700-in-search-of-the-very-first-coded-symbols/

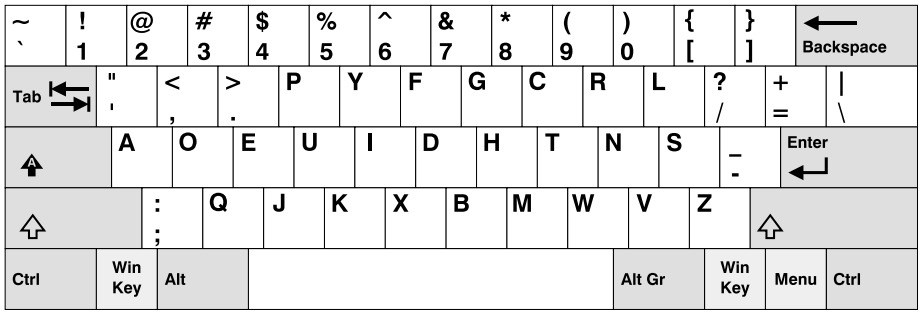


Figure 2.1: The Dvorak Keyboard, designed by Dvorak and Dealy in 1936. The layout was designed after careful investigation of hand motion since the objective was to design an accurate, faster layout and create less stress than QWERTY. Unfortunately, the cost of efficiency that Dvorak layout provides is not high enough, thus impeding its growth and switch with QWERTY.

personal computers. However, the keys to insert the letters remained almost the same as was in the Sholes-Glidden typewriter[2]. Attempts were made to optimize the arrangement, as seen in the Dvorak layout (see Figure 2.1). However, the Qwerty layout from Sholes-Glidden still stays prominent in most countries.

Text entry plays a crucial role in our lives, which often gets unnoticed. Several works investigated different domains of text entry to understand what could enhance the usability and performance of the system for an improved user experience. The following sections list the different parameters of text entry and how various research works contributed to understanding it.

2.2 PARAMETERS OF TEXT ENTRY VIA KEYBOARD

As discussed earlier, the process of text entry starts at the cognitive level, which translates to the movement of actuating parts of the human body on either a tangible, tactile surface or on a digital, similarly (in most cases) laid out interface. The following subsections discuss the prior research conducted in understanding the different directions of text entry like keyboard layout, input modalities, and measuring performance.

2 https://www.antikeychop.com/sholesgliddentypewriter

2.2.1 *Layout*

Layout of text entry systems is a crucial parameter that impacts the text entry process and user retention on that design. From physical keyboards to digital keyboards on-screen, layout plays a crucial role in how fast the user adopts the design and improves their text entry speed while maintaining accuracy. In the following sub-sections, we list the different keyboard layouts (physical and digital interactions), which show the amount of exploration that keyboard design went through to improve the usability and user experience of the system. Our work in Chapters 4 and 5 also highlights different on-screen keyboard designs for gaze-based text entry systems where we explore different layout parameters.

However, it was not user experience and other usability principles that guided Scholes to design the layout of his typewriter. Instead, the mechanical limitation of the machines guided the design that became so popular that we use an adaptation of that design even today. The layout design of keys on a keyboard was to improve the text entry rate. Researchers investigated the optimal configuration of the keys by understanding the frequency of words and letters in a language corpus. Thus, the dependency on the language in which the layout is designed became a common property for optimal designs.

The following points discuss the physical/tangible keyboard interface and the rise of the Digital/on-screen keyboard interfaces for interaction with hand or other input modalities like a stylus or pen.

1. Physical: The early design alterations in the typewriters were seen in the *index typewriters*[3] (see Figure 2.2) that consisted of a single wheel for letter selection followed by a key that confirmed the selection for the letter to be imprinted on the paper. The popularity of index typewriters grew since they were affordable and portable, unlike the prevalent Sholes's design.

   While the original Sholes-Glidden typewriter (1874) revolutionized writing, it still lacked the basic feature of the two-letter case - the uppercase and the lowercase of normal typefaces. To overcome this challenge, *double keyboard typewriters*[4] came into place (see Figure 2.3).

   While these designs impacted the text entry scenario when typewriters came into play, the modern-day hand-held mobile phones

3 https://www.contextualternate.com/exhibition01#ex01-about
4 https://www.antiquetypewriters.com/typewriter/caligraph-2-typewriter/

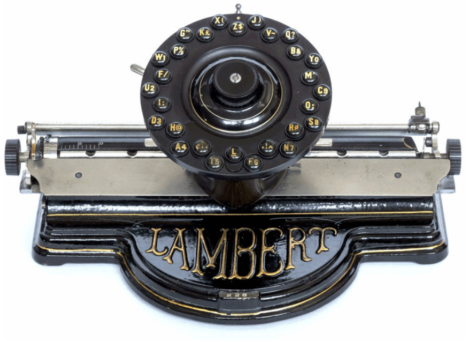


Figure 2.2: The LAMBERT index type writer of 1884, invented by Frank Lambert. The small portable typewriter had a circular anticlockwise layout with a central selection button to imprint the letter on the paper.

Figure 2.3: The "CALIGRAPH 2" typewriter of 1882.The keyboard had a unique layout of the two letter cases. The lowercase is in white while the upper case is in black. The image is resourced from *Quin, Liam R. E.: "Typewriters from the Martin Howard Collection"* (2008)

had an alternate layout. It solves the challenge of adapting 26 characters into a 12-key setup for an efficient text entry (Figure 2.4). A *multitap* method was designed wherein each key would provide a piece of different information when tapped more than once. So if one had to type the word"sky", on the mobile phone keypad, the keystrokes would be *7-7-7-7-5-5-9-9-9*. Nine keystrokes would be required for a word with only three letters (considering no errors were made). This increase in keystrokes naturally reduced the typing speed and efficiency of the system.

Predictive text entry came into play to overcome this challenge, where a dictionary would generate matching words from the single interaction sequence on the keys. One of the most well-

known and used predictive systems for such a use case scenario was *text on nine (9) keys*, better known as *T9*, developed by Tegic[5]. The predictive algorithm enabled users to arrive at the desired word from the combination of keys on the keyboard. Thus, to type"sky", one needed to press *7-5-9*. This reduced the cognitive load associated with multitap, enhanced text entry speed, and improved keystroke ratio.

Figure 2.4: The multitap keyboard in the earlier generation mobile phones before the touchscreen era.Each key had 3 letters except number 7 and 9. The 0 was used for entering a space.

2. Digital: On-screen digital keyboards came much later into existence than the traditional typewriters. While most of them tried to emulate the same layout style, some researchers opted to investigate optimal layouts by delving into the language model and understanding which letters are used the most. Some examples are:

    - *Fitaly layout*[6] (Figure 2.5): A unique one-finger accessible layout with two keys for entering space and letters so placed to minimize the travel distance between keys to form words (Figure 2.5). It has six rows compared to the three rows in a traditional keyboard layout and roughly six alphabets in each row. This ensured a compact design suitable for single-hand use cases. The design also enabled users to reduce hand movement when inserting text. It was primarily designed for text entry using a stylus on a touch screen. Later it was modified for left-handed, right-handed, and mouse-driven input modalities.

    - *Opti Layout* [96] (Figure 2.6): This optimized keyboard layout was designed keeping in mind the Fitt's law [48], character and diagram frequencies in the English language, and trial and error method of text entry. The design ra-

5 https://en.wikipedia.org/wiki/T9_(predictive_text)
6 https://textware.com/fitaly/fitaly.htm

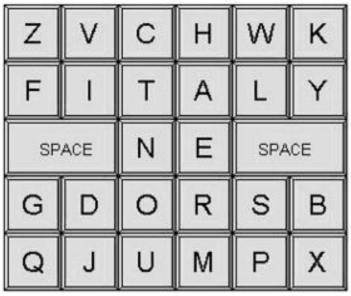


Figure 2.5: Fitaly Layout: A commercial one-finger typing layout that aimed to minimize the travel distance between the keys to form words

tionale behind this layout ensured that there should be no dead space between keys, where no action is assigned; no limit on how many sizes or shapes can be used; shape should be rectangular to fill in a typical application window. According to Mackenzie and Zhang [96], this layout performed 35% faster than the traditional QWERTY and 5% faster than Fitaly.

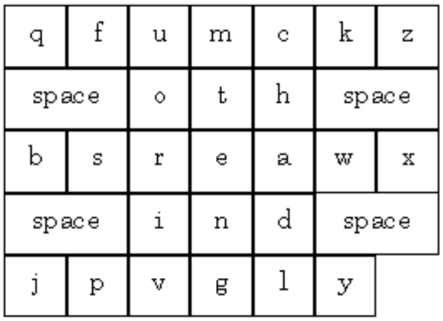


Figure 2.6: The Opti Layout: Designed by Mackenzie and Zhang, the core objective of the layout was to improve the writing speed using Fitt's law. The layout has four space bar keys considering the heavy usage of space between words to form a sentence.

- *Cirrin Layout* [108] (Figure 2.7): Inspired by short hand and unistroke gesture, this circular layout was designed for stylus-based input to speed up the text entry process. The word gets formed as the stylus traverses through the letters inside the ring. A challenge with this design layout was the visual feedback of a letter selection was at the center of the screen which was hindered when user moved the stylus across the characters on the ring.

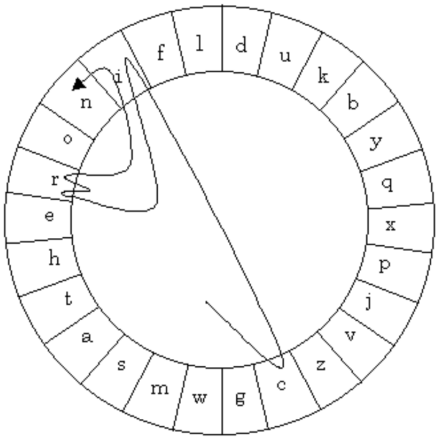


Figure 2.7: Cirrin Layout: Designed by Mankoff and Abowd, this design facilitated the use of a stylus for an efficient text entry process without lifting the device from the skin. This is efficient from using a stylus on a normal keyboard and emulating touching every letter with a finger.

2.2.2 *Input Modalities*

In HCI, modality is referred to as the independent channel of input/output between a computer and human. A system with more than one channel of input/output is multimodal.

From the onset of this chapter, the focus has been on writing or typing as a means of text entry where the human hand plays a crucial role. While the hand remains dominant for the keyboard-mouse combination, other modalities like voice, gaze and switches are now used for text entry, as described in this subsection.

1. Mechanical: The keyboard-mouse modality combination is still dominant when it comes to text entry on the computer. While it serves the purpose of input of text for most cases, there are some use cases where this modality fails to deliver. People with limited mobility or accessibility issues often find the keyboard-mouse combination difficult to use. Also, for situational impairment, when our hands might be busy doing something else, and we still need tools for text entry, researchers have investigated and designed other mechanical ways.

   Switches have been used and investigated for text entry [125]. They vary from mechanical to pneumatic switches controlled by breathing or even bite switches. Since switches are generally a binary form of input, other modalities (as will be discussed) have been used as switches like eye blinks [9, 156], muscle contractions [47] and non-verbal interactions [138]. Song et al. [169] studied using a mechanical joystick for text entry with word predictions for people suffering from motor-control challenges. The MDITIM (Minimal Device Independent Text Input Method)

method [63] was also investigated to be controlled by a joystick. Trackballs [63, 190] have been used in several text entry scenarios for people with motor-control issues (an example image can be seen in Figure 2.8). Head tracking[7] for quadriplegics have also been studied. A noticeable challenge with such approaches

Figure 2.8: An image from a user study performed by Wobbrock et al. [191] where the participant prefers the use of a Stingray trackball.

is the limitation of entering multiple letters with speed, as we do when typing with hands. In most of the alternative approaches, the process of letter entering is fairly linear - one at a time.

Dwell time is the duration of a user's fixation exceeding a predefined threshold, resulting in a selection trigger.

2. Gaze[8]: One of the most natural means of interaction is the human gaze. We always *see* something and then react to it. For text entry, gaze plays a crucial role as it helps in scanning letters but also assists us in navigating to the point of an error to apply the error correction mechanism. Gaze-based text entry has been either dwell-based or dwell-free.

   Dwell times are adjusted to make the key selection faster. Gaze-based text entry performance was analyzed extensively by Räihä and Ovaska [143]. Text predictions were found to play a crucial role in improving text creation. Several works [98, 162] investigated the position of text predictions on gaze-based keyboards to understand their impact in improving the performance and usability of the system. Several works have attempted to solve the problem by adding another modality in the system to overcome the challenges of gaze-based inputs. Gaze has been combined with voice [159], touch [134], and even switches [78] to improve the text entry and interaction experience.

3. Voice: Like gaze, another natural modality is voice. While gaze input can be used as a modality for pointing specific locations, voice commands can be used to define the task at hand clearly.

7 https://www.naturalpoint.com/smartnav/

8 entry rate, error rate terminology is explained in detail in the subsequent sections

Figure 2.9: A photograph from Menges [110], showing MAMEM trials where participants suffering from Parkinson disease writing emails with the help of gaze input.

With the advancement of the Automatic Speech Recognition systems (ASR), voice input has been used to enter text by dictation [124] or by spelling out all the letters [106, 123]. While it is assumed to help people suffering from motor control challenges, voice has its limitation of incorrect recognition [193]. People suffering from speech impairments are limited from using such modality. However, research now investigates the use of non-verbal vocal signals [173] for text entry [59]. In this case, sounds like humming, hissing, and whistling are used in conjunction with a specially designed interface layout [59] or with another modality [159].

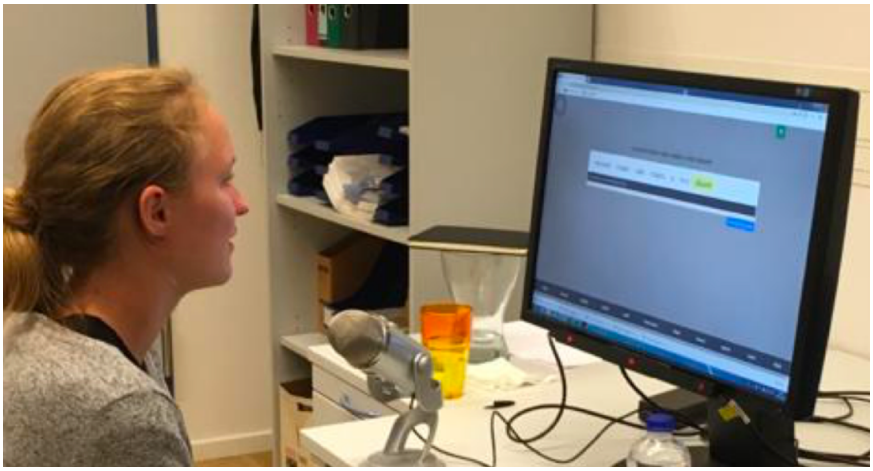


Figure 2.10: In this image, a participant is seen entering a text with voice input with the help of the external microphone placed in front of her.

4. Biosignals: Another alternative approach to text entry has been to use the biosignals from the human body and create a mapping

system that translates to a selection of letters. Considering non-invasive approaches, Electromyography for intentional muscle contraction [45] and Electroencephalogram for brain signal [42] (Figure 2.11)are two modalities that have been investigated chiefly for text entry scenarios. In both these conditions, the nature of the input signal is extremely noisy. It needs sufficient signal processing to turn muscle twitch or brain signal into a viable communication input method. This input modality is beneficial for communication with people completely paralyzed or who can perform some facial muscle movement.

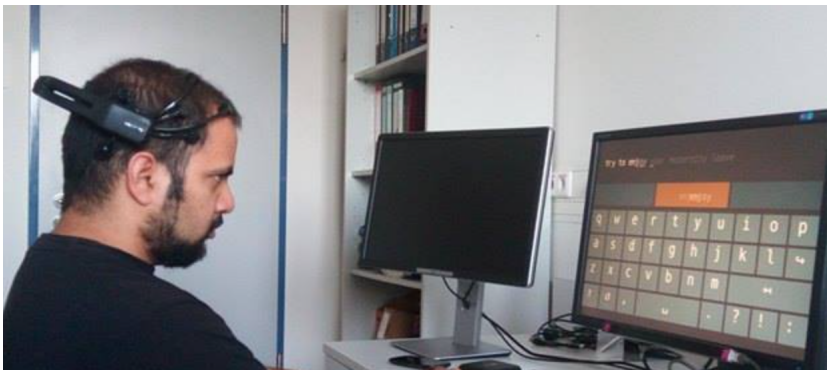


Figure 2.11: A photograph from Sengupta et al. [163] showing the measurement of EEG signals for text entry-related purposes.

For this thesis, we investigate alternative modalities like gaze and voice as primary channels of text input and revision. Bio-signals like EEG were used as a performance metric to understand the cognitive load associated with interacting alternative modality approaches for text entry and revision.

2.2.3 *Performance*

The evaluation of any system is often performance-centric. The system's performance is often a staple question whenever something new is designed and investigated. Often, the measurement of speed for text entry systems is the primary concern. While speed is an important feature, other metrics should be examined to understand the system's usability. This section presents some of the empirical measures of text entry performance. While many of these metrics are *method-agnostic*, this section intends to highlight some of the key measures without which a text entry system's objective and subjective evaluation remain incomplete.

2.2.3.1 *Objective Measures*

1. Entry Rate: This point discusses some of the measures associated with text entry speed. While the discussion can extend from character-level entry to word-level transcription, we present two of the most common metrics associated with how fast text can be entered into a system.

    - *Words per Minute (WPM)*: The most popular and widely reported metric for recording speed of text entry is WPM [187]. WPM considers only the length of the final transcribed string and how long it took to produce it. It does not consider the number of keystrokes or gestures made during the text entry process. The formula for computing WPM :

    $$\text{WPM} = \frac{|T|-1}{S} \times 60 \times \frac{1}{5}$$

    In the above equation, T represents the total length of the transcribed string by the user. T contains all the alphanumeric entries from the user apart from the backspace entries. S represents the time taken by the user to transcribe T. The *60* in the equation represents time in seconds and $\frac{1}{5}$ represents the assumption that words are mostly 5 characters long [192]. Since the S represents the time from the first character to the last, the *-1* is present in the numerator.

    pack my box with five dozen liquor jugs $\rightarrow$ *39 characters*
    $\wedge$ $\wedge$
    t= 0 sec *t= 10 sec*

    Thus, WPM = $\frac{|39|-1}{10} \times 60 \times \frac{1}{5} = 45.6$

    A challenge with such a metric not considering the backspace and ignoring the keystrokes associated with the backspace or correction procedure.

    - *Keystrokes per second (KSPS)*: As mentioned, WPM does not consider the process of text entry. It merely focuses on the final results. KSPS helps us with that information to understand the entire text entry process.

    $$\text{KSPS} = \frac{|IS|-1}{S}$$

IS is the input stream of characters that were executed to form the sentence. this includes not just alpha-numeric insertions but also corrective procedure in case of any error. *S* represents the time taken by the user to transcribe *IS*.

The red letters indicate incorrect entries. The < indicates the use of backspace keys

The example for KSPS shows how different it is from WPM.

t= 0 sec
$\vee$
pac*l<*k my bo*c<*x with fi*bw<<*ve doze*m<*n liqu*e<*or jugg*<*s $\rightarrow$ *45 keystrokes*
$\wedge$
t= 20 sec

Thus, KSPS = $\frac{|45|-1}{20}$ = 2.2

2. Error Rate: While speed plays a part in understanding how fast the text entry can be possible on a system, Error Rates provide us with the information of how accurate the intended system usage is. The fewer errors made by the user during text entry, the more accurate the system evaluates to the user's intention. Error rates can be evaluated from the system perspective or the user perspective. If the user makes a spelling mistake, then the user creates the error. However, if an intelligent speech-to-text transcription system fails to transcribe an intended word, that becomes a system error. In both cases, if the user recognizes the error and applies a correction mechanism, then the cost of text entry increases as the total time for task completion rises.

   We discuss the common error metrics:

   - *Keystrokes per Character (KSPC)*: This simple ratio of the number of characters entered (*IS*)to the total length of the string (*T*) was formulated by Soukoreff and Mackenzie in 2001 [170]. The lower the value of KSPC, the better the performance. A value of 1 signifies perfect text entry.

     $$\text{KSPC} = \frac{|IS|}{|T|}$$

     Thus for the example of KSPS, KSPC would be $\frac{|45|}{|39|}$ = 1.153.

   - *Uncorrected Error rate*: As the name suggests, this corresponds to the number of errors that were left uncorrected upon completion of the sentence.

3. Correction: Error correction plays a crucial role in understanding the system's robustness. The cost of correction often impacts the task completion time. The more efficient the correction process, the faster the task completion. The process of error correction varies based on the different modalities used. However, a baseline evaluation can be performed to come up with a ratio based on the number of words or characters (*IF*) that were initially erroneous and then was fixed(*F*). Mackenzie and Soukoref introduced this approach for character-level assessment as *Correction Efficiency*(CE) [171].

$$\text{CE} = \frac{|IF|}{|F|}$$

CE is applicable for single character-based interaction for error correction. However, there are other approaches for error correction that minimize the cost. One common approach is to replace the wrong word with a correct word from the predictions. Another approach that deals with the removal of multiple characters was introduced as Fisch in-stroke word completion technique by Wobbrock et al. [188].

The CE also varies when we consider error correction in voice-based systems. There, word selection and word re-utterance play a crucial role.

In the pursuit of measuring both speed and accuracy, objective metrics frequently fail to address essential questions, such as:

- Despite its prominence as a primary benchmark, is the user content with the system's speed?
- Does system usage induce fatigue?
- Was it straightforward for users to acquire proficiency with and adapt to the system?

2.2.3.2 *Subjective Measures*

To overcome the gap in understanding the perspective of system usage from the user, subjective measures are used. They help us in understanding the performance from the user's perspective along with how they feel when using the system. The subjective measures involved in testing a text entry system are similar to any other new system in design. While the following metrics do not fall under text entry metrics, they have been used to understand the system's usability and the perceptual-cognitive load.

1. System Usability Scale (SUS): Designed originally in the 1980s by John Brook to test the usability of electronic equipment, this questionnaire was used and modified later to test the usability of all kinds of systems - from websites to mobile phones to even machinery. While other questionnaires like SUS are also available, the success of SUS lies in its nature of being agnostic to technology and being small and quick for users to complete the procedure.

   The SUS questionnaire comprises of 10 questions (Q) with the possibility of five responses (R): (i)Strongly Disagree, (ii)Disagree, (iii) Neutral, (iv)Agree, (v)Strongly Agree.

   - I think that I would like to use this website frequently.
   - I found the website unnecessarily complex.
   - I thought the website was easy to use.
   - I think that I would need the support of a technical person to be able to use this website.
   - I found the various functions in this website were well integrated.
   - I thought there was too much inconsistency in this website.
   - I would imagine that most people would learn to use this website very quickly.
   - I found the website very cumbersome to use.
   - I felt very confident using the website.
   - I needed to learn a lot of things before I could get going with this system.

   For the responses R, 1 is deducted from the value if the question is odd, and 5 if even. Then the sum is calculated followed by 2.5 times the value to reach the SUS score

   The questions Q and its responses R can be represented as: $Q = (q_i, r_i)|r_i \in R = \{1,2,3,4,5\}, i \in \{1,2,3,4,....,10\}$

   In order to find out the SUS score:

$$f(q_i) = \begin{cases} r_i - 1 & i \in \{1,3,5,7,9\} \\ r_i - 5 & i \in \{2,4,6,8,10\} \end{cases} \tag{2.1}$$

   Thus the SUS Score is:

$$SUS = 2.5 \times \sum_{i=1}^{10} f(q_i) \tag{2.2}$$

2. NASA Task Load Index (TLX): Another subjective measure like SUS, but designed to understand the *task load*. Developed originally at NASA's Human Performance Group, this subjective evaluation metric is prevalent in task-based evaluations. The objective of this survey was to understand different sources of

the workload associated with the task. This subjective measure takes into consideration the (i)Mental Demand, (ii) Physical Demand, (iii) Temporal Demand, (iv) Performance, (v) Effort, and (vi) Frustration and asks the following questions:

- How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching)? Was the task easy or demanding, simple or complex, exacting or forgiving?
- How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
- How much time pressure did you feel due to the rate of pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
- How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?
- How hard did you have to work (mentally and physically) to accomplish your level of performance?
- How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

The initial evaluation is a 10-point scale with a 0.5-point interval, following which the user is asked to provide importance to which task load he/she deemed more in comparison to another. The number of times each load is given prominence over the other generates a weighted score. Using these values, the final score is generated, which depends on the scale scores of each load. Researchers often use the "Raw TLX" score to avoid this confusion.

For our experiments, we used the online tool[9] to collect data and generate the final NASA TLX score.

3. Custom Questionnaire: Custom heuristic questionnaires are often developed to better understand the system's usability. Questions are often adapted from sources like Nielsen's heuristic[10] checklist that provides a guideline for understanding the usability of the system (as can be seen in Chapter 4). In our experiment, we also used a subjective custom questionnaire that focused on

9 https://www.keithv.com/software/nasatlx/
10 https://www.nngroup.com/articles/ten-usability-heuristics/

understanding the perception of Accuracy, Learnability, Speed, and Comfort, and we asked users to rate it on a seven-point Lickert scale.

3

GAZE AND VOICE MODALITIES FOR TEXT ENTRY SYSTEM

This chapter investigates two powerful input methods, human gaze and voice, and their integration into text entry systems. Section 3.1 delves into various eye movements and their application in selection and pointing interactions. Subsection 3.1.1 further narrows the focus to gaze input for text entry, discussing mechanisms like dwell time, dwell-free, saccade-based, and smooth pursuit interactions. Section 3.2 studies speech as an input modality, tracing its evolution from verbal to non-verbal modes and how these advancements enrich text entry and interaction.

3.1 UNDERSTANDING GAZE INPUT

This section provides the background of eye tracking and its use for interaction and text entry. We start our section by understanding how eye movements are used in eye tracking for interaction. Finally, we move on to how the eye movements are used for text entry - as a standalone input modality or in combination with another modality.

The human eye plays a crucial role in our daily lives in inspecting, perceiving, and understanding the environment around us from a sensory perspective. Gaze forms one of the most natural forms of communication amongst humans [71] and forms a bridge of our actions with the environment around us. The visual information around us is sent to our brains for processing via the photoreceptor cells in our retina. These cells get stimulated when light reflections from the environment reach our eyes.

While photoreceptor cells in the retina are active, the eye must maintain a ”fixation” position to process incoming visual information. Typically, the eyes move rapidly from one fixation point to another within a few milliseconds. These swift eye movements have been categorized by Robinson [145] as ”saccades” and ”smooth pursuits.” In the context of gaze-based interaction [121], researchers have explored

saccades, smooth pursuits, and fixations, encompassing various types of physiological nystagmus.

Nystagmus (ni-stag-muhs) is a condition in which eyes make rapid, repetitive, uncontrolled movements — such as up and down (vertical nystagmus), side to side (horizontal nystagmus) or in a circle (rotary nystagmus)

1. *Fixations*: It is during fixation that the eyes remain relatively still (while making short movements: physiological nystagmus). The word fixation in gaze-based literature is often used to mention the act of fixating on visual stimuli.

2. *Saccades*: The rapid eye movement from one fixation to another is referred to as a saccade. During a saccade, the eye can reach a velocity of 700 degrees per second [13]. Detection of saccades in gaze-based evaluation helps in fixation detection [126].

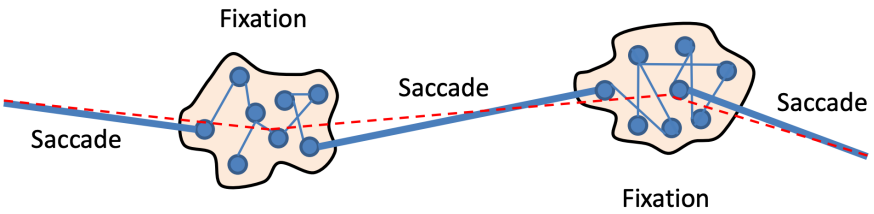


Figure 3.1: Fixation and Saccades in an eye movement [75]

3. *Smooth Pursuits*: Also know as "Fixation-in-Motion" [121], smooth pursuits are defined as eye movements when eyes track a moving object (For example, a car in motion.). Both saccades and smooth pursuits are related to the motion of the eyes. Saccades initially have a high velocity followed by deceleration, but for smooth pursuits, the speed depends on the object's movement it is tracking.

Using gaze signals for input has been investigated and has been found to be broadly classified under two categories: (i) selection and (ii) pointing.

The Midas touch problem [64] is a common challenge in using gaze signals as input. Jacob coined this term to describe a situation where the system cannot disambiguate if the gaze signal is used for selection or just exploration, thus leading to inadvertent selection. To overcome such challenges, different selection and pointing processes have been designed.

1. *Selection*: The voluntary/involuntary act of choosing an object, region, or point of intent from a set of stimuli can be defined as the selection process. In an eye-tracking context, the process involves the precise control and coordination of eye movements.

Selection plays a critical role in human-computer interaction and usability studies, as it determines what elements users attend to and interact with, influencing their cognitive processes and task performance.

One of the initial and intuitive interactions was using eye blinks as a gaze interaction selection technique. However, it suffered from the inherent challenge of disambiguating intentional blinks from unintentional blinks. In his article [85], Jacob considered blinking an infeasible interaction for selection for all these challenges.

While Jacob considered blinking as infeasible, he also mentioned the dwell-time selection method, where one needs to fixate on an intended selection element for a longer duration to trigger the selection process [64]. This process is simple and more effective than blinking. However, visual feedback becomes necessary for users to understand the ongoing dwelling process, leading to selection to adapt interfaces for dwell-time-based selection. Research on adaptive dwell-time [142, 198] also showed improvement in selection.

Another approach to selection has been eye-gestures [35, 140, 152]. Gestures gave the advantage of not being dependent on fixations, relying on movements, and detecting movement patterns.

2. *Pointing*: Considering gaze as an input, pointing is directing the gaze towards a target. It is typically defined as a saccade towards the target. The saccade's amplitude is typically used to measure the distance between the current gaze position and the target. Pointing with eye gaze is a fundamental aspect of eye-tracking technology, often used in applications like eye-controlled navigation, enhancing accessibility and usability of computer systems and assistive technologies.

   While the selection process was investigated in detail, the problem persisted with pinpointing where selection needs to happen accurately. One approach followed was magnification [6, 83, 199]. Magnification involved a multi-step approach that involved time. For a more straightforward pointing technique, Lutteroth et al. [90] color-coded interaction elements. While it improved the interaction, the colors occupied the screen area.

3.1.1 *Using gaze input for Text Entry*

As mentioned earlier, text entry is crucial for different use cases. We investigate how other researchers have used gaze input for text entry:

1. Dwell-Time: It is one of the most common approaches to gaze-based interaction. When used in gaze-based text entry systems, every key is designed to trigger with a dwell time of 200-1000ms [100, 104]. However, the lower the dwell time for the keys, the possibility of accidentally triggering a letter increases, thus increasing the overall cost of text entry. In most dwell-time-based keyboard designs for text entry, the alphabets, major punctuation marks, and predictions are activated by dwelling on them. While the process is convenient, completing a word's letters by dwelling is time-consuming unless a good word prediction engine reduces the task load. Two approaches were investigated to overcome the challenge of the static dwell time selection approach: (i) User-adjusted and (ii) Automatic.

    - *User-adjusted*: A longitudinal study of ten participants was conducted by Majranta et al. [100] to investigate the impact of user-adjusted dwell time. From a starting 1000ms dwell time, participants achieved an average of 282 ms after using the system for a prolonged time. While the time decreased, with frequent usage, the users improved the average text entry speed from 6.9 WPM to 19.9 WPM. Another study by Raiha et al. [142] achieved 20-24 WPM using the user-adjusted dwell time technique.

    - *Automatic*: Spakov et al. [198] investigated automatic dwell time adjustment technique with nine participants. Their approach involved adjusting the dwell time based on the time of selection between two letters and the gaze moved out from the key. With this approach, they achieved a rate of 12.1 WPM.

2. Dwell-free: An alternative to dwell-based gaze typing is dwell-free typing. Instead of fixating on a key for a predetermined time, dwell-free systems allow users to gaze briefly at their intended key before moving to the next. The system is then responsible for disambiguating the user's input [120].

   A common disadvantage of the dwell time technique depended on the dwelling process for every key that needed to be triggered

to form a word or apply corrective measures. Kristensson et al. [74] explored the concept of speed gain if the dwell time was reduced to zero, i.e., a dwell-free setup. However, since their results are based on simulations, actual human experimental results are yet to be seen.

EyeSwipe [81] represents a dwell-free approach to emulate the functionality of well-established shape-writing systems commonly utilized on touch-enabled mobile devices. In this system, users initiate word selection by employing a reverse crossing technique to designate the first and last characters of the intended word. Subsequently, the middle characters are chosen as the user directs their gaze towards them sequentially. The selection of words is based on an n-best list generated from the user's recorded gaze path. According to the findings reported by the authors, non-disabled participants achieved an average text entry rate of 11.7 words per minute (wpm) with an accompanying average Mean String Distance (MSD) error rate of 1.31%.

3. Saccade-based: Saccades have been used to activate selections, and a combination of them have been designed to represent gaze-based gestures programmed for different actions on the interface. Morimoto et al. [118] uses saccade-based text entry for designing a context-switching keyboard, as seen in Figure 3.2. The challenge with such an approach is the keyboard size that takes up the entirety of the screen.

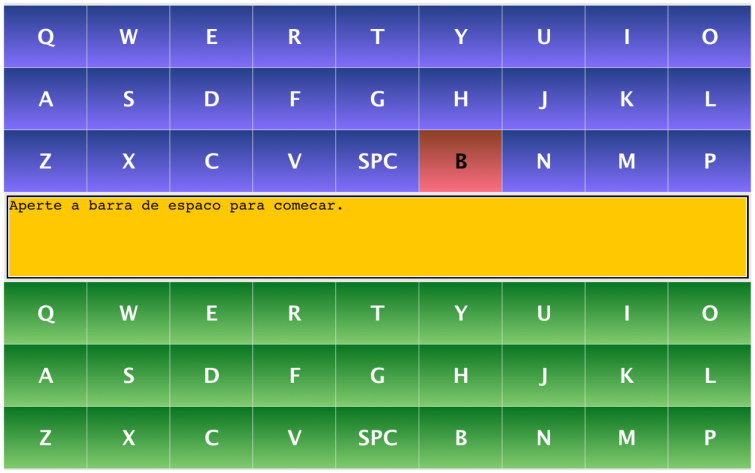


Figure 3.2: Context Switching Keyboard Design for saccade-based selection. The two separate regions of the keyboard (purple and green) represent two contexts. Focusing on the keys is done by short dwell times, and selection to type words is done by "changing" the context to the other keyboard and letter. Users can comfortably explore the whole content of a context without the effects of the Midas Touch problem.

Approaches like EyeWrite [189], pEYEwrite [61], Quikwriting [15] also attempt to solve the text entry challenge with saccade-based gaze typing.

4. Smooth pursuit: Zooming interfaces explored smooth pursuit as a means of interaction [183]. Since smooth pursuit is based on objects moving, text entry interfaces like Dasher [182] utilize this concept to provide a continuous zooming interaction for text entry. While the design does not adhere to the traditional keyboard-based approach, Dasher produced the fastest text entry speed in the gaze-based text entry realm.

   StarGazer [56], SMOOVS [91] also use smooth pursuits to interact with the elements on the screen.

3.2 UNDERSTANDING VOICE INPUT

Speech as an input modality is used as a communication tool in various ways to achieve natural means of digital communication. This section discusses how researchers have introduced this input modality for efficient hands-free interaction in text entry scenarios.

Our voice forms one of the earliest and natural modality for humans. Voice, initially in the form of simple sounds, helps us understand until the ambient language model takes over. We start moving from sounds to phonetically similar syllables and proceed to speech of the language we are surrounded with. However, even before progressing from sounds to words, we can communicate and express our intentions and emotions. This occurs via another modality like hand or hand gestures or elevation in the voice, leading to emotional signatures like happiness and frustration.

However, with the advancement of technology, Licklider, in the early 1960s, proposed the idea of spoken dialogue between man and machine [89]. In continuation with this work, Richard Bolt in 1980 [21] published his research "Put-That-There" augmenting voice command with gestures to produce an early version of the multimodal system that was able to create shapes on a large screen with simple voice commands like "create a yellow circle" and then spatially move it to another location by pointing in space and giving the command "put that there."

Human speech recognition has substantially improved the Automatic Speech Recognizers (ASR). This has led to their widespread adop-

tion in different domains and products. Voice-based conversational agents like Google Home, Siri, Alexa, and others started as simple conversational agents or VUI (Voice User Interfaces) and now perform various tasks based on the commands given (even like telling a joke or a story).

The advancements and adoption of voice-based interaction also come with specific challenges that are still being investigated to improve the overall user experience of VUIs. Some common challenges being worked on are language-dependent incorrect recognition, homophones, ambient noise, multi-speaker detection, and user mistakes [53, 168].

Voice input has been investigated as an eyes-free interaction [49, 135, 175, 178]. This led to voice becoming a go-to modality for accessibility challenge resolution. Zhong et al. [196] used voice for complete interaction of an android phone to facilitate such interaction for the blind population. Other such works that used voice input for device-based interactions are Capti-Speak [7], a non-visual web browser [5], etc. While it assisted accessibility issues, voice input has also been investigated for healthcare [92, 115, 117], tourism [72, 146], and even in the domain of education [10, 84]. Schulman et al. investigated voice as a tool for understanding attitudinal and behavioral cues [154, 155]. Not limited to these domains, voice input has also been used for gaming [25], assistance [88], and recommendation [85].

Research has focused on understanding voice input from the perspective of conversational agents and has investigated aspects like context [87], emotion [73], empathy [119], and even humor [37].

Using voice for text entry

As mentioned in Chapter 1 and discussed in Section 2.3.2 for gaze, text entry forms an inevitable and crucial interaction. Section 2.4.1 discusses the use of voice for different uses cases, showcasing the importance of hands-free interaction. Motivated by such findings, we investigate the affordance of voice input for text entry.

Voice as a natural modality for communication is faster than hand-based [148, 149] and even gaze-based text entry. Its inherent nature supports hands and eyes-free interaction for situations involving accessibility issues or when the hands are busy performing other tasks. However, even when quite advanced, voice recognition suffers from recognition challenges. These recognition challenges often form bar-

riers in designing and implementing voice-based user interfaces for improved user experience.

While we focus on the interaction and user experience for text entry, voice-based text entry is heavily dependent on the performance of Automatic Speech Recognizers (ASR) like Mozilla voice[1], Google's speech-to-text[2], Dragon speech from Nuance[3], etc. Recognition errors from ASRs and others limit the user experience, especially in the text entry scenario. This is particularly problematic since research shows that 80% of the transcription time is invested in correcting recognition errors [11] in voice-based text entry - thus adding overhead cost to an already efficient input modality.

While the challenges are being investigated, voice-based text entry can be classified into two categories: verbal and non-verbal.

1. Verbal: Verbal or traditional text entry with voice is simply the text entry process with the help of ASRs. As described earlier, this approach has been broadly investigated for different use cases and is efficient if recognition errors are disregarded. However, in reality, error corrections form an important interaction that impacts the overall text entry process.

2. Non-verbal: People suffering from speech impairment conditions like dysarthria [28] often cannot speak; thus, the verbal mode of text entry becomes a challenge for them. Non-verbal voice interactions like humming or whistling [18] have been investigated for such situations for text entry. This approach has been used for character-level text insertion based on humming patterns [27, 34], but is often slow and erroneous, leading to further frustrations. Multimodal combinations of humming or whistling with other modalities have been investigated for improving the non-verbal text entry [59, 174].

1 https://commonvoice.mozilla.org/en
2 https://cloud.google.com/speech-to-text
3 https://www.nuance.com/dragon.html

4

GAZETHEKEY: INTERACTIVE KEYS TO INTEGRATE WORD PREDICTIONS FOR GAZE-BASED TEXT ENTRY

In this chapter, we introduce "GazeTheKey" (GTK). This novel gaze-based keyboard design utilizes dwell time to optimize text prediction selection, enhancing the user experience in hands-free text entry. GTK's innovative approach focuses on bringing predictions closer to the visual fovea, effectively addressing the intrinsic limitations of gaze as a modality in text entry and providing a more effective, user-friendly solution. Section 4.1 provides a comprehensive overview of gaze-based text entry, exploring various gaze-based keyboards and identifying key factors contributing to word predictions' success and usability. In Section 4.2, we conduct a detailed design investigation, examining keyboard layouts, the utilization of word predictions, their placement, and various text entry techniques. Section 4.3 delves into understanding the usage and positioning of word predictions in virtual keyboards, highlighting their impact on user interaction and efficiency. Finally, Section 4.4 presents our design, "GazeTheKey," showcasing its unique features and demonstrating how it stands out in gaze-based text entry solutions. The following sections present the keyboard design's initial and final investigation against two other designs.

The chapter presents the understanding of key interaction components required for gaze-based text entry and a detailed design investigation. From the literature, we understand that word predictions play an essential role in improving the performance of gaze-based text entry systems. However, visual search, scanning, and selection of word predictions require a shift in the user's attention from the keyboard layout when selecting such predictions. Hence, the spatial positioning of predictions becomes a crucial aspect of the end-user experience. Thus, we also investigate the role of spatial positioning by comparing the performance of GTK.

We observe a high selection of word prediction when it is in the constant visual attention of users, often leading to incorrect selection, thus increasing the cost of error correction. Our evaluation and observation indicate that fast saccadic eye movements undermine spatial

distance optimization in prediction positioning. This is especially essential when understanding and developing new keyboard designs that target improved efficiency of text entry.

This chapter is adapted from two of our published papers, [161] and [162], published at IUI 2017[1] and ETRA[2] 2019.

4.1 GAZE-BASED TEXT ENTRY

Text entry in gaze-based systems is accomplished using on-screen keyboards facilitated by wearable or standalone eye trackers. In these scenarios, on-screen keyboards featuring various designs and text entry methods have constituted the primary interface for letter input and selecting available text predictions. In most cases, enlarging key sizes has been a prevalent strategy to mitigate drift-related inadvertent activations. There have been some designs and approaches where the dwell free [74], and smooth pursuit methodology [182] have also improved the text entry process.

Drift: the progressive displacement of fixation registrations that results from a gradual loss of eye-tracker calibration over time

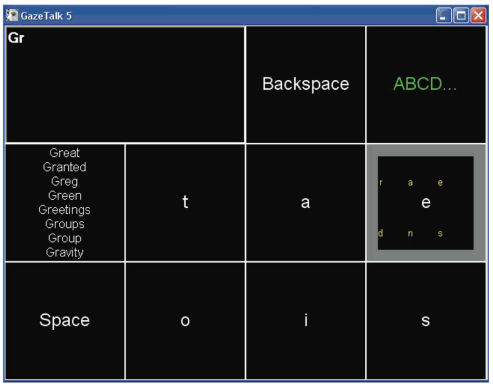
However, due to the larger keys, the on-screen keyboard takes up a considerable part of the screen. The large keys have a single purpose of just entering the letters, while the designated word prediction area displays the words. This design approach for dwell-based text entry reduces the efficiency of the system.

AugKey [38] (Figure 4.1a) presents the layout with word predictions placed on the right side of the keyboard augmenting keys with a prefix to allow continuous text inspection and suffixes to speed up typing with word prediction. This design attempted to utilize the space around the predictions to exploit the foveal region of visual perception. Johansen et al. [67] (Figure 4.1b)had a similar idea for each of the keys. However, their design split the predictions from the prefixes even though they utilized the space around the letters.

The success and usability of text predictions depend highly on the presentation and user interface parameters [51]. This includes (i) the number of suggestions to display (too few might miss relevant suggestions, and too many will add an extra delay of scanning a long list), (ii) layout of the presentation (horizontal, vertical, triangular), and most importantly (iii) the positioning of suggestion in the screen space of keyboard. Another design perspective that came out from the previous literature was to utilize the empty space of the large keys.

1 https://iui.acm.org/2017/index.html
2 https://etra.acm.org/

(a)



(b)

Figure 4.1: Gaze-based text entry keyboards with text predictions at different places. (a)Keyboard from Augkey by Diaz-Tula et al.; (b)Keyboard from GazeTalk by Johansen et al.

4.2 DESIGN INVESTIGATION

We systematically analyzed keyboard designs and input methodologies to understand the design and interaction space for gaze-based text entry systems. The aim was to understand the key components required for designing a text entry system and the existing pain points. We adopted the Zwicky box technique [197] and identified the following categories for gaze-based text entry:

Zwicky Box Technique: The process of breaking the problem down into categories, adding values to each category, and combining these values to create unique answers

1. Keyboard Layout: In Chapter 2, we discussed the different layouts and how each had the primary goal of improving the text entry speed. While many continued to use the Dvorak Layout with minor changes, others experimented with different placement of keys, space bar or punctuation to primarily increase the text insertion speed. Unfortunately, most new design layouts did not detail the training phase or the learning curve participants generated while using their system.

2. Usage of Word Predictions: The early investigations of gaze-based keyboards did not include a word prediction system. It primarily focused on letter-based text entry. Investigation into

the usage of word predictions led to them being integrated into the keyboard design - as was available in most hand-held on-screen keyboard designs.

3. Placement of Word Predictions: With the increased usage of word predictions, they were primarily placed above the keyboard as a row [98]. This was followed by moving the predictions around the keyboard, as shown in Figure 4.1. Some designs even delved deeper into letter-level predictions and their selection to enhance the text entry system.

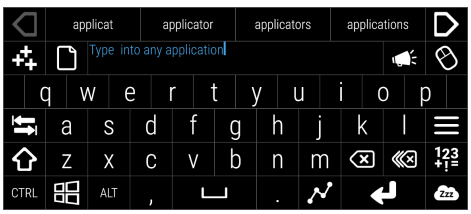
4. Text entry technique: The technique adapted to enter text (dwell, dwell-free, and saccade-based) has also been investigated, and each approach has been found to pose certain inherent challenges. While no other gaze-based text entry technique achieved as high a speed as Dasher, these techniques possess a steeper learning curve in practical scenarios.

Taking inspiration from Card et al. [24], we compartmentalized our categories in two dimensions:(i)*Keyboard Properties* - containing Keyboard Layout, Usage of Word Predictions and Placement of Word predictions; (ii) *Text Entry techniques* (as discussed in Chapter 3.1.1).

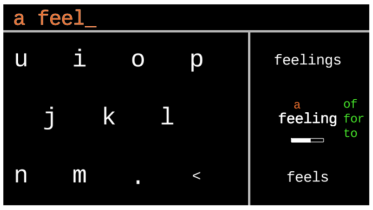
4.3 UNDERSTANDING USAGE AND POSITIONING OF WORD PREDICTIONS IN VIRTUAL KEYBOARDS

While investigating keyboard properties, we understood word predictions play a significant role in text entry. They are generated from a language corpus or a word frequency dictionary. Predictive algorithms help the user suggest words from the corpus most likely to occur after a particular sequence of user-selected characters. The research focused on predictive letter models like n-gram [66], and k-gram [116], which suggest the following terms of a given sentence based on the previous terms. Reflective text entry [151] improved the user experience of text entry as it considered abbreviated forms of words.

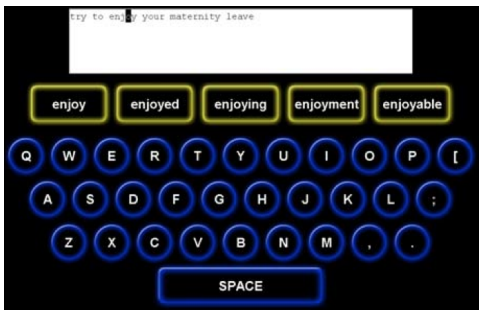
There have been several gaze-based text entry systems [38, 67, 98] that use word prediction as an essential feature in the virtual keyboard space. Prediction mechanisms are particularly valuable for text entry with virtual keyboards (for gaze-based as well as touch-based systems) [164, 172]. The success and usability of word predictions depend highly on the presentation and user interface parameters [50]. It includes (i) the number of word predictions to display (too few might miss relevant predictions, and too many will add an extra delay of



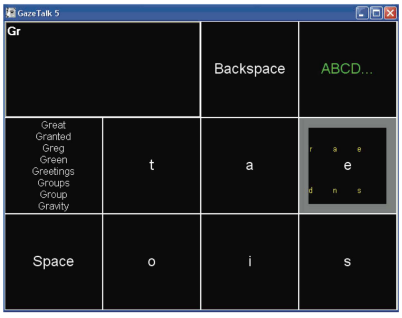
(a) Keyboard of OptiKey suite[3]



(b) AugKey layout [38]



(c) Eye Typing by MacKenzie & Zhang [98]



(d) GazeTalk Keyboard [67]

Figure 4.2: Gaze-based virtual keyboards with word predictions. Figure (**a**) shows us the onscreen keyboard from the Optikey suit. Figure (**b**) gives us the layout of the keyboard AugKey where the word prediction comes with prefixes around the prediction to exploit the foveal region of visual perception. Figure (**c**) was designed to involve word predictions with the next keystroke about to be hit. Figure (**d**) is from GazeTalk that included both word and letter predictions.

scanning long list), (ii) layout of the presentation (horizontal, vertical, triangular, etc.), and most importantly (iii) the positioning of the predictions in the screen space of keyboard. Positioning of word predictions is crucial since it deals with the user's visual attention while typing letters and relates to cognitive and perceptual influence.

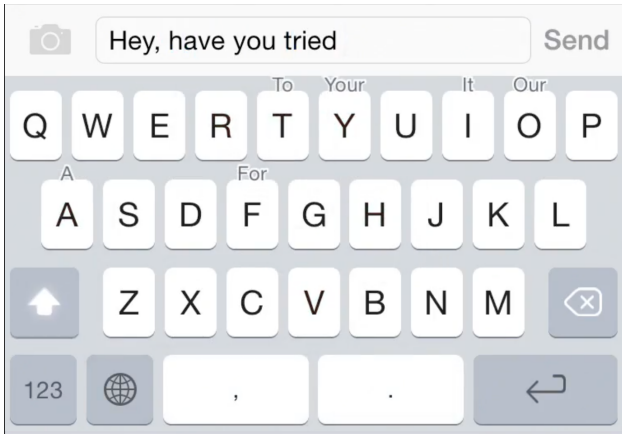
Figure 4.2b and Figure 4.2d showcases a few gaze-based text entry systems, signifying the variable positioning of predictions in different approaches. For most conventional designs, a predicted word list is placed on top of the keyboard layout near the text entry area. This can be seen in the interface (Figure 4.2a) of a popular open-sourced gaze-based interaction tool *OptiKey*. Figure 4.2c shows the eye typing approach with word and letter predictions by McKenzie and Zhang [98]. Their design, however, places word predictions below the text area. Figure 4.2b shows the *AugKey* approach [38] where word predictions are framed at the right side of the keyboard and also include prefixes around the key to exploiting the foveal region of visual perception. The *GazeTalk* system (Figure 4.2d) [67] provides both word and letter prediction features. The predicted word list is

3 https://github.com/OptiKey/OptiKey/wiki

on the keyboard layout's left side. The preview of the next character layout is also available within the currently selected cell.

In touch-based text entry on virtual keyboards (e.g., text entry on mobile displays), the representation and positioning of word predictions have received significant consideration. Some modern virtual keyboard layouts in the touch-based text entry domain present the predictions closer to the attention of the user by embedding them in the keypad as inter-spaced and in-letter dynamic predictions [54, 55]. Popular designs on Blackberry (Figure 4.3a[4]) and iOS keyboards (Figure 4.3b[5]). Cuaresma *et al.* [32] showed that bringing predictions closer to users' attention by in-letter predictions in mobile phone keyboards enhances their ability to interact with predictions and significantly improves the typing speed by touch interaction.

(a)



(b)

Figure 4.3: Virtual mobile keyboards that bring word prediction close to the keys. (a)Blackberry Keyboard; (b)Crimson Keyboard

4 https://www.donmckenzie.ca/portfolio/bb-virtual-keyboard/
5 http://ok.k3a.me

These approaches emphasize the role of word prediction in gaze-based text entry. However, it is unclear if the variable positioning of these predictions impacts the performance by reducing eye movements, visual search, or scanning time. Majranta *et al.* [99] argued that an increase in perceptual and cognitive load occurs due to the shift of focus from the keyboard to the word prediction list while scanning it. However, no concrete studies have investigated if the variable positioning of word predictions correlates with visual attention and could enhance the user experience while typing.

4.4 GAZETHEKEY

Considering the observations and approaches as described in the prior section, we aimed at understanding the individual keys of the keyboard where a lot of white space was not utilized, leading to these questions:

- How can the keys incorporate the word predictions at design level
- How does the user interact with word predictions

The interaction relies on the next word that the respective language model could predict based on the previously entered text by user [179]. Figure 4.4 shows the the keys containing the letter and the associated word suggestion at the bottom. The green framed letters on the key are the ones the user has already entered. The letters framed in yellow could be the next letter. The red frame signifies the prediction based on previous letters.

Figure 4.4: In GazeTheKey design, for each of the letters that will house word predictions, per key is estimated with previous letters entered by the user and letters associated with the key as input. In this image, the already entered letters to form words are in green, the letter on the key is yellow, and if the letter on the key is activated, then the predicted letters from the corpus to form the words are in red.

While Figure 4.4 presents the structure of word suggestion on keys, Figure 4.5 displays the user interaction methodology via different states of a key on eye gaze fixation. a) At first, the key is in its standard mode to trigger the input of the single letter displayed on it. b) This is implemented via a dwell time visualized through an orange circle. When the user gazes at the key, the circle grows from the center of the key until the key's area is filled. c) The input of the displayed letter is triggered, and visual feedback in the appearance of a black pulse is fed back to the user. d) If a suggestion is available by the prediction engine, the key now turns into suggestion mode e) If the prediction on the key is what the user intends to select, the user continues to dwell on the key and the prediction fills the key starting from the bottom and ending at the top. f) When the key is filled, the currently collected word is replaced by the given suggestion. The user can abort all key actions by looking at a different position on the screen.

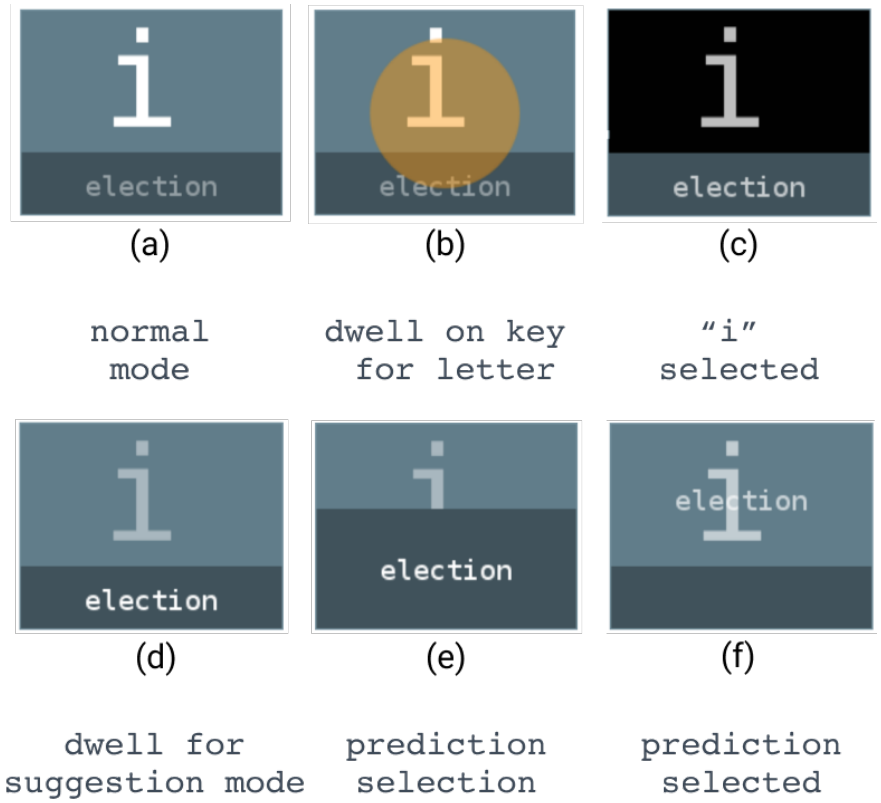


Figure 4.5: In GazeTheKey, we implement a double dwelling interaction that enables people to shift from letter to prediction selection without shifting their visual foeva. The approach to the placement of the predictions is presented in Figure 4.4

Figure 4.6 shows the complete design of GTK keyboard interface including above-mentioned functionality (detailed demonstration of GTK usage is available here[6]). The design includes the principles

6 https://youtu.be/-UDDTJHBPVA

of eye-controlled interfaces [76] (e.g., enlarged buttons and visual feedback to cope with eye-tracking accuracy); moreover, it follows the usability heuristics of Neilsen[7], keeping the design as close to conventional keyboard layout as possible with minor adjustment of key formation and addition of necessary keys for improving efficiency while typing.

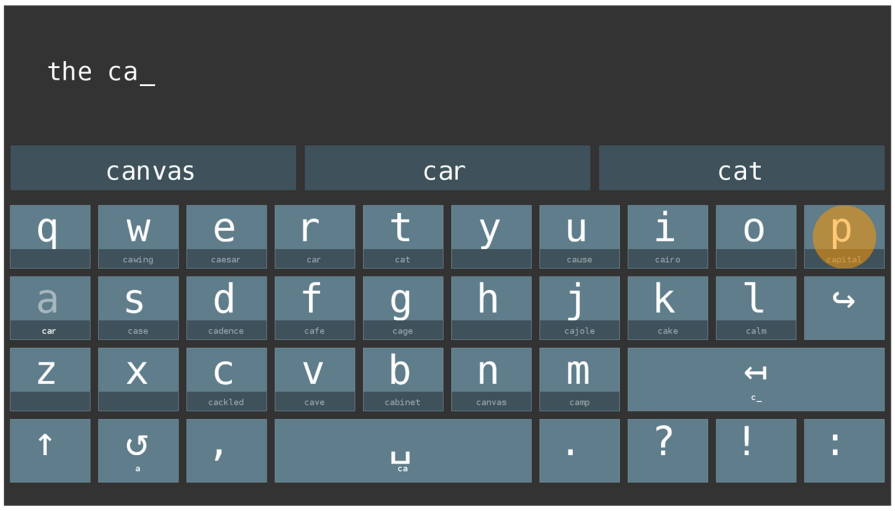


Figure 4.6: The complete GazeTheKey layout with predictions on every key based on the letter selection.

In the shown example in Figure 4.6, the user typed some letter of a word, and the current state of all keys shows relevant suggestions that can be selected via additional dwell time over the key (as explained in Figure 4.5). One limitation of such a two-step dwell time input is that the user couldn't dwell on a letter consecutively to type the letter multiple times. Hence, an extra *repeat key* has been added at the lower bottom of the environment. This key serves to trigger the input of the last selected letter. We have done further optimization to minimize the visual search for users, i.e., space and backspace key is used to present further information about the edited word. As shown in the example, the space button works as the confirmation of the typed word, and therefore, it is displayed on the bottom of the key. A preview of the edited word (after deletion by backspace) is displayed on the backspace key (essentially showing the usage of the respective key). These simple heuristics help the user stay with the gaze in the same position without checking the intended action on the edited word.

7 https://www.nngroup.com/articles/ten-usability-heuristics/

4.5 INITIAL EVALUATION

A preliminary investigation was conducted to understand the potential of our design.

4.5.1 *Apparatus*

An SMIREDn eye tracker running at 60Hz was attached to a 24-inch monitor that displayed 1280 x 800 pixels. The participants were asked to sit on an adjustable height chair before the calibration process to center the eyes at a distance of about 70 cm from the screen. Calibration of the SMI eye tracker was done by the SMI calibration tool. However, when participants reported too much drift, re-calibration was done.

4.5.2 *Participants*

Ten participants (five male, five female) contributed to our study; they were aged between 21 and 30 years (mean 24.8, SD 2.347) and had no prior experience with eye-controlled interfaces. However, all of them had adequate experience with computer usage, and all of them were familiar with the QWERTY layout of a keyboard. All the chosen participants were well-versed in English, although none were native English speakers.

4.5.3 *Procedure*

To test the performance of GTK, we performed an experimental evaluation to understand the system's efficacy. An experiment was designed that consisted of the 10 participants recruited to type sentences taken from the phrase set of Mackenzie and Soukoreff [93]. The experiment was built into five sessions, and each session had five sentences that the user needed to transcribe with the help of the newly designed keyboard. Each participant was provided with a training session, and before the actual onset of the experiment, their systems were reset so that the word prediction algorithm was not biased. A physical keyboard was placed in front of the participant whose *space bar* was used to go to the next sentence transcription. In summary, the design had:

10 participants ×
1 new keyboard design ×
5 sessions ×
5 sentences in each session
= 100 submissions in total.

4.5.4 *Results*

To understand the performance of our design, we performed subjective and objective evaluations. We performed the words per minute analysis for objective evaluations and the percentage usage of predictions.

1. Objective evaluation:

    - *Words per minute (WPM)*: Figure 4.7 shows us the performance of GTK across 5 sessions. The maximum WPM across five sessions for 10 participants was 11.17, with a mean of 9.34 (showing notable acceptance).

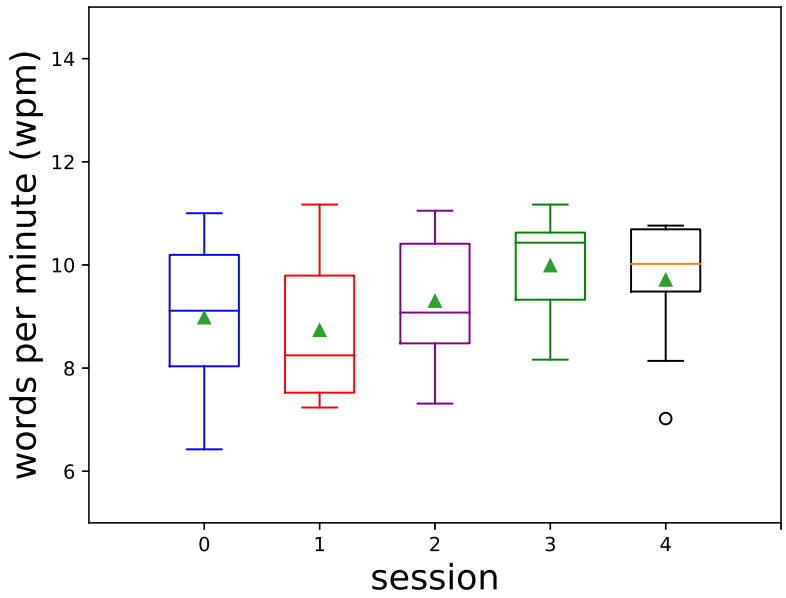


Figure 4.7: Usage of word predictions across different sessions for *GTK*

    - *Prediction Usage*: Since our design aimed to improve gaze-based text entry by integrating word predictions inside the keys, we evaluated the percentage usage of word predic-

tions when using GazeTheKey. Table 1 shows how participants (*P*) improved on the usage of on-key word predictions for their task completion after every session (*S*).

	S1		S2		S3		S4		S5	
	TR	On Key	TR	OK	TR	OK	TR	OK	TR	OK
P1	10.34	48.27	6.67	66.67	4.17	87.75	22.73	63.63	36.36	59.09
P2	15.38	69.23	18.75	62.5	30	63.33	44.44	48.14	53.57	32.14
P3	14.28	82.14	28.57	60.71	19.23	73.07	10.34	75.86	28.57	75
P4	22.22	44.44	22.22	59.25	40	54.44	3.7	92.5	3.8	89.65
P5	7.47	77.78	14.28	75	19.23	76.9	6.89	89.65	10	80
P6	82.3	0	85	0	82	5.47	85	0	80	0
P7	79.23	27.58	60.8	26.1	50.0	46.15	38.6	57.23	26.9	61.53
P8	45.83	41.67	62.5	37.5	45.83	45.83	30.76	61.53	17.24	62.06
P9	66.67	33.34	46.15	47.69	33.33	66.67	35.48	64.51	38	52
P10	48.27	37.93	57.14	35.71	36.27	60.5	41.67	58.33	44.4	55.25

Table 1: TR: Top Row; OK: On Key. Usage of predictions on key and on the top row (See Figure 4.6) for our design GazeTheKey. The data clearly indicates the growth of on key word predictions with the passage of time and sentences.

2. Subjective evaluation: To understand how the participants felt on using the system, we designed a custom heuristic questionnaire inspired from the heuristic principles [8] that Jacob Nielsen proposed:

    a) How good is the visibility of main interaction elements? (Interaction elements include keys, suggestions, typing area)
    b) How close did you feel to the features of this keyboard to a conventional keyboard that you are used to?
    c) How easy was it to control the keyboard?
    d) How easy was it to recover from errors made?
    e) How will you rate the design of the keyboard?
    f) How comfortable was to use eye tracking on this keyboard design?
    g) Was the design intuitive? (If there was no guidance, would you have figured it out easily?)

The average score from the heuristic evaluation was 8.05.

Our initial experiment demonstrates a noticeable increase in word prediction utilization when users can select desired words directly from the keys. This experimental study aimed to understand the efficacy of the design and understand if the performance of such a keyboard falls in the acceptable range of text entry. However, to comprehensively assess its performance relative to a traditional on-screen gaze-based keyboard, it is essential to compare various performance metrics. The upcoming sections will delve into the evaluation of GTK

8 https://www.nngroup.com/articles/ten-usability-heuristics/

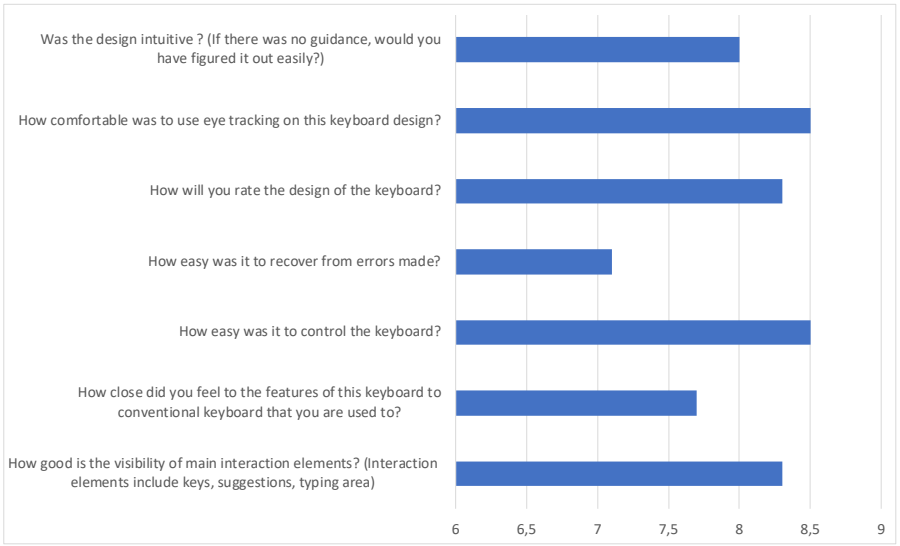


Figure 4.8: Heuristic evaluation of the usage of GazeTheKey.

and its effectiveness in positioning predictions closer to the visual fovea.

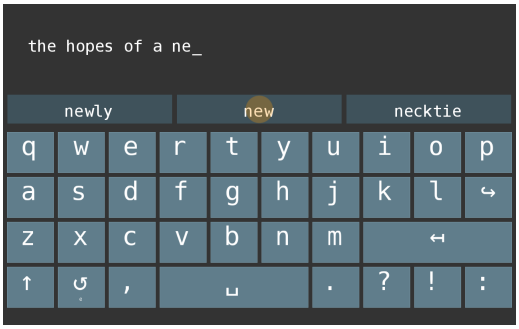
The initial evaluation sheds light on *RQ1* where we take the approach of designing a new predictive keyboard to overcome the limitations of existing gaze-based keyboards.

4.6 GAZE-BASED KEYBOARD DESIGN

For our investigation of understanding the impact of variable positioning of word predictions, we used the traditional keyboard layout (let us term it as Keyboard *A*), an inter-spaced keyboard with prediction position above each row (let us term it as Keyboard *B*) and finally the keyboard we designed, *GTK*.

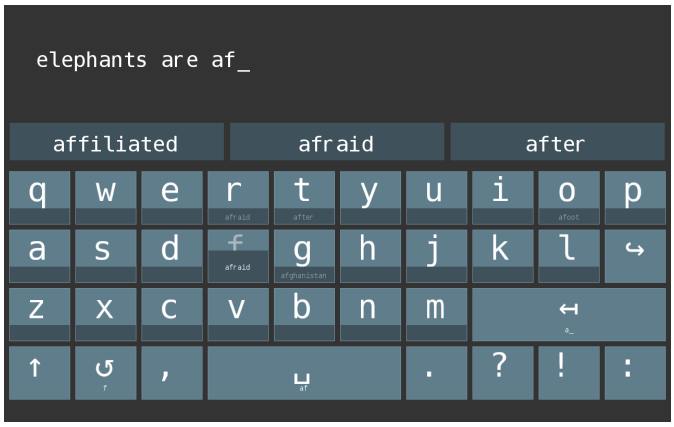
Keyboard A (Figure 4.9a) has a single line of word predictions on the top of the keyboard area. This design is adapted from the most conventional touch-based text entry keyboards design. It also represents the most prevalent design for gaze-based text entry keyboards. *Keyboard B* (Figure 4.9b), is an inter-spaced keyboard that has been designed to bring the word predictions inside the keyboard layout. The predictions are displayed as inter-spaced in the line over the last triggered letter to reduce the visual distance to the last area of fixation. The inter-spacing was inspired from keyboards as shown in Figure 4.3a and Figure 4.3b. This design was also created to investigate if the findings by Cuaresma *et al.* [32] for mobile phone keyboards also hold for gaze-based text entry systems. *GTK* (Figure 4.9c), embeds the prediction related to

the letter on the representative key. It brings the visual focus to the keys. It also has a single line of word prediction on the top of the keyboard area to ensure accessibility to the increased number of word predictions. In all the keyboards, the most relevant word prediction was placed in the *middle* followed by *left* and *right* for all the word prediction positions across the three keyboards.



(a) Keyboard A: Conventional Keyboard

(b) Keyboard B: Bringing word prediction inside Keyboard



(c) *GTK*: Bringing word prediction inside keys

Figure 4.9: Keyboard A, B designed to evaluate impact of variable word prediction position in connection to *GTK*

For Keyboard A and *GTK*, the complete keyboard layout along with the word prediction area took approximately 65% of the screen space. For Keyboard B, it was 77% of the screen space. The on-key word predictions for *GTK* occupied approximately 30% of the space of the key on which it was initially displayed.

The dwelling status can be seen clearly in Figure 4.9a and 4.9b and also in the step-by-step implementation of the double dwelling interaction approach in Figure 4.5

The keys and word predictions are the main responsive elements arranged in QWERTY order for the virtual keyboard interface. The QWERTY layout was modified to include the most used punctuation [31] for quicker access. The layout change with the dimensions mentioned earlier utilizes the available space and the eye tracker's accuracy. The font on these elements is rendered in white, while the

fore- and background is kept in shades of dark and unsaturated green to provide a clean and non-distracting experience.

Interaction is implemented via a dwell time of 1.0 seconds for key activation. The dwelling status is queried to the user with a transparent orange circle centered in the middle of the element and growing at fixation until filling the complete element. When the complete element is filled, the key or prediction is activated and the content added to the collected input.

Keys in *GTK* features a two-step dwell time approach. It requires a second dwell time for activation of the offered word predictions. Once the first step of dwelling selects the letter, the key switches to its prediction during the second dwelling step. The same duration of fixation dwell time is necessary to trigger the input of the displayed word prediction (Figure 4.5). The *space* and the *backspace* key in *GTK* include a preview of the currently edited word after activation of these keys. This integrates into the concept of locating the visual focus on the keys. All the keyboards include a particular key in the lower-left part to repeat the last letter for double-lettered words. This was necessary only for *GTK* as it does not allow for repetitive key activation. The offered prediction is activated during a second dwell time phase instead of the selected letter.

4.7 FINAL EVALUATION

The final evaluation involved five consecutive eye typing sessions on different days. The participants were asked to experiment with the keyboard layout allotted to them as per the Latin Square ordering. It was done to nullify the effect of bias. The experiment was conducted in a controlled lab environment with artificial illumination. The *dependent* (measured metrics: wpm, backspace usage, error rate, keystroke saved), *independent* (test conditions: keyboard layouts, prediction positioning, visual feedback) and *controlled* variables (ambient lighting, font size, font colour, key size, key colour, prediction size, visual feedback colour etc.) were noted for the proper execution of the experimental process. Before the actual experimental study, a pilot test was conducted with four participants to validate the experimental procedure. The participants were asked to enter each time a single sentence was presented in the text area in the upper region of the keyboard interface. At the first keystroke, the sentence disappeared, and the participant had to recall the sentence to continue. This procedure simulates free writing and prohibits the participants from comparing

the collected input with the desired result, which would influence the gaze data strongly [81].

4.7.1 *Participants*

The main experimental session consisted of 10 participants (five male and five female). The participants' age ranged between 21 to 30 years (mean = 24.8, SD = 2.348). Due to technical challenges [recording of gaze data had abruptly stopped], we considered 9 participants as the data recorded for 1 of the participants got corrupt and could not be recovered. While the findings offer valuable insights into this specific context, more participants would have ensured greater generalizability of the results to a broader population and enhanced statistical power, allowing for more robust and widely applicable conclusions. However, given the constraints and the specific research objectives, the smaller sample size was justified in this instance, providing a deeper understanding of the unique aspects of the study's subject matter.

70% of the participants wore spectacles, and none had prior experience with eye-tracking/typing environments. However, all of them had adequate experience with computer usage and were familiar with the QWERTY layout of a keyboard. All the chosen participants were well-versed in English, but none were native English speakers.

4.7.2 *Apparatus*

For this experimental study, we used the same apparatus setup as described in Section 4.5.1.

4.7.3 *Procedure*

Like Section 4.5.3, we followed the same approach here. The area of the collected text can be seen in Figure 4.9a, 4.9b and 4.9c. Each participant was introduced to a *training phase* which consisted of two sessions of 5 sentences each for the participants to get familiarized with the environment. The system was reset for every session so that the predictive engine would not bias word prediction. Participants were instructed to use the physical space bar on the physical keyboard in front of them to access the following sentence in the experiment. In summary, the design was:

9 participants ×
3 keyboard designs ×
5 sessions ×
5 sentences in each session (excluding practice phrases)
= 675 submissions in total.

4.7.4 *Results*

Standard metrics for text entry evaluation include [94, 97]: (i) Words Per Minute, (ii) Error, (iii) Keystrokes Saved. We have evaluated two other parameters to understand the usage of word prediction in a gaze typing scenario (iv) Backspace Key usage and (v) word prediction usage. The metrics below give a detailed direction to the findings. While typing speed performance indicates non-significant change, there is a high usage of predictions and backspace keys.

1. Words per minute (WPM): WPM of 9 participants across five

The explanation and details of Words Per Minute (WPM) can be found in Chapter 2, under the section titled "Performance" (2.2.3.1)

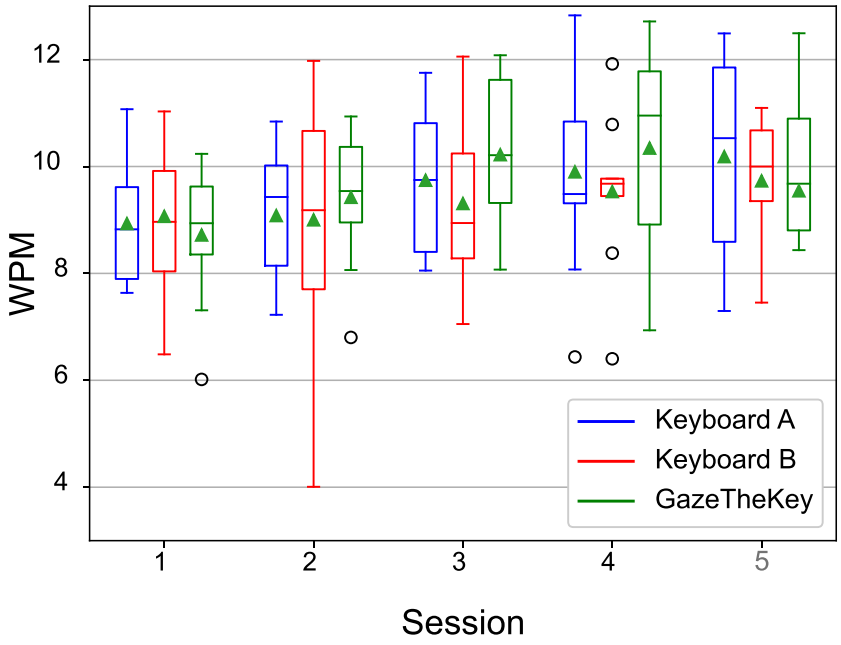


Figure 4.10: Words per Minute performance across different sessions for Keyboard A, B and *GTK*

sessions for three different keyboards designs can be seen in Figure 4.10. ANOVA on *WPM* across different sessions for the three different keyboards reveal a non-significant effect, $F_{2,12} = 0.420, p = 0.67(ns)$, with the grand mean of each of the keyboards being very close to one another: *9.57*, *9.36* and *9.65* wpm for *Keyboard A, B and GTK* respectively. The values lie

well within the range of 7-25 words per minute range reported in other setups [102, 183], indicating reasonable eye typing speed. More specifically, the noted text entry rate lies in the upper range for a dwell-based keyboard with no extensive training. For example, gaze-based text entry speeds using dwelling is about 10 wpm after about ten training sessions [105].

No significant learning effect was observed across the performance of the three keyboards.

Traditional metrics as mentioned in Chapter 2, under "Performance" (2.2.3.1) consider letter or word-based entry. For our case, this was slightly complicated since the user was at the freedom to choose letters or words.

2. Error: Uncorrected errors are missed or wrongly entered compared to the original sentence and not corrected. *Levenshtein Distance* is one measure of calculating the edit distance that measures the deviation of the input sentence to the original sentence. The grand mean of the uncorrected error for the three keyboards across different sessions are *0.56*, *1.36* and *0.88*. Figure 4.11 shows the errors left uncorrected by the participants across 5 sessions .

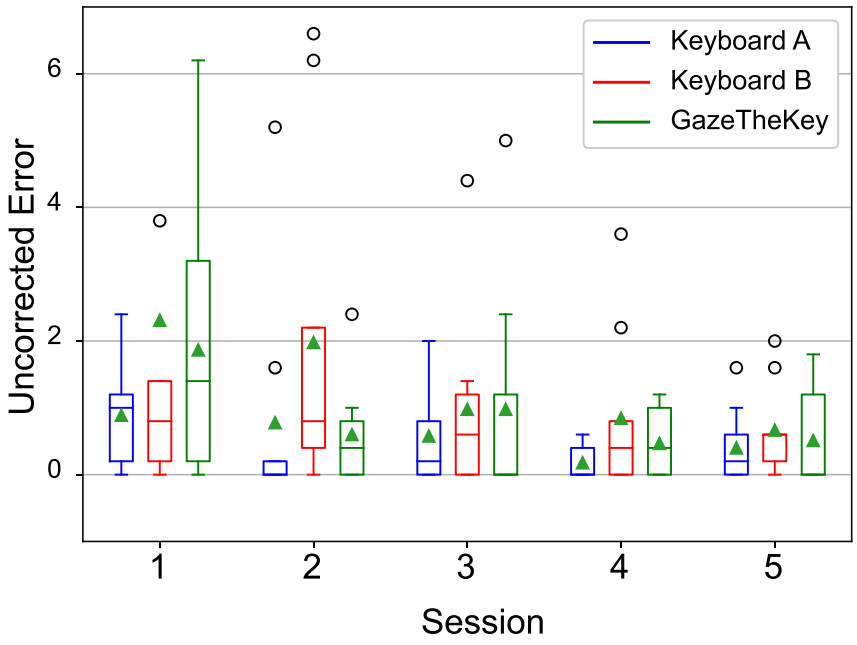


Figure 4.11: Uncorrected Error across different sessions for Keyboard A, Keyboard B and *GTK*

Shapiro Wilk Test[9] revealed the data to be not normally distributed. Hence we used a Friedman test, which gave a significant result with $p = 0.02$. Keyboard A had the least number of errors, followed by *GTK* and B.

9 https://www.sciencedirect.com/topics/psychology/shapiro-wilk-test

No learning effect for uncorrected error is observed across the performance on the three keyboards.

3. Keystrokes saved: Measurement of keystrokes is another crucial measure of performance in text entry systems. The use of word prediction reduces keystrokes, thus leading to faster text entry speed. In this experimental study, every keystroke was calculated and compared against the original count of letters for the sentence they were provided with. The percentage of keystrokes saved across different sessions for the three keyboards designs is shown in Figure 4.12. Grand mean of *35.48%*, *34.54%* and *28.16%* of saved keystrokes were recorded for the three keyboards across five sessions.

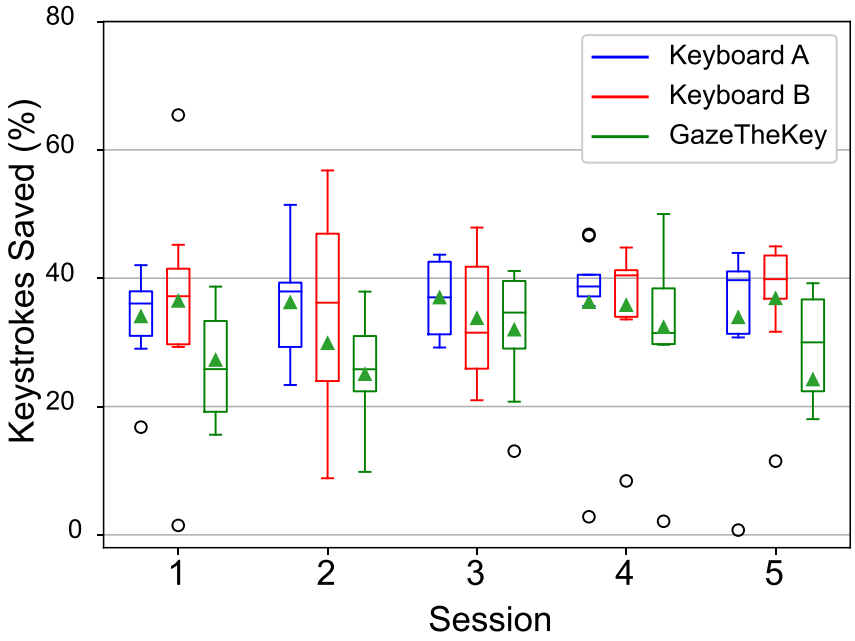


Figure 4.12: Percentage of Keystrokes saved across different sessions

ANOVA shows significant result with $F_{2,12} = 9.56; p = 0.003$ indicating the use of significantly fewer keystrokes to achieve complete sentences in *Keyboard A* than in Keyboard *B* and *GTK*.

No learning effect was observed for keystroke savings across the three keyboards.

4. Backspace: Backspace usage indicates the attempts made to correct the sentence/words before confirming. It also indicates the participants' corrections when they accidentally selected a wrong letter or a wrong word prediction from the list. Grand mean of *0.72*, *1.17* and *1.44* backspace hits for the three keyboards were recorded across five sessions. Figure 4.13 indicates the

efforts required to formulate a sentence were much higher for *GTK* and keyboard B through deleting the characters. Further investigation of the backspace usage revealed the high amount of backspaces were used for correcting/editing the picked predictions.

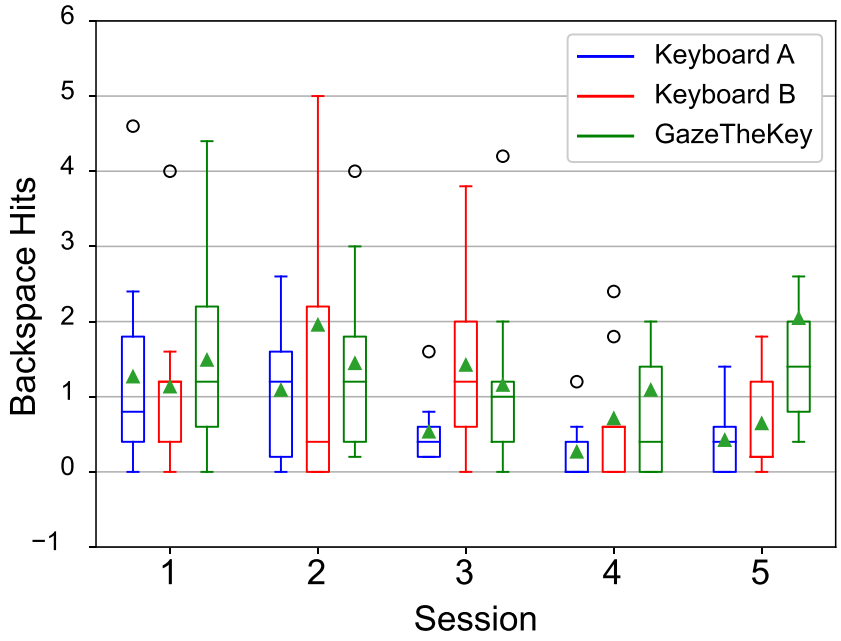


Figure 4.13: Backspace Key Usage across different sessions for Keyboard A, B and *GTK*

ANOVA shows a non-significant result with $F_{2,12} = 3.25; p = 0.07(ns)$.

5. Word prediction usage:This metric measures the usage of word predictions while formulating the sentence. It indicates how effective the predictions were and how easy it was to access them.

   For Keyboard A, with only a one-word prediction line at the top, the suggested usage was 90.12%. The inter-spaced Keyboard B had 91.11% usage and 93.21% in the *GTK* with predictions on the keys themselves. ANOVA gave a non-significant result $F_{2,12} = 0.08; p = 0.92(ns)$. Figure 4.14 shows us the session-based performance across the three keyboards.

   Further analysis shows that the participants accepted the layout's word predictions well. For keyboard B, 46.61% of the used predictions are chosen from the top, 33.88% from the center, and 19.51% from the bottom. In *GTK*, with prediction-enhanced

keys, 54.37% of the utilized suggestions are taken from the keys instead of the single-word prediction line on top.

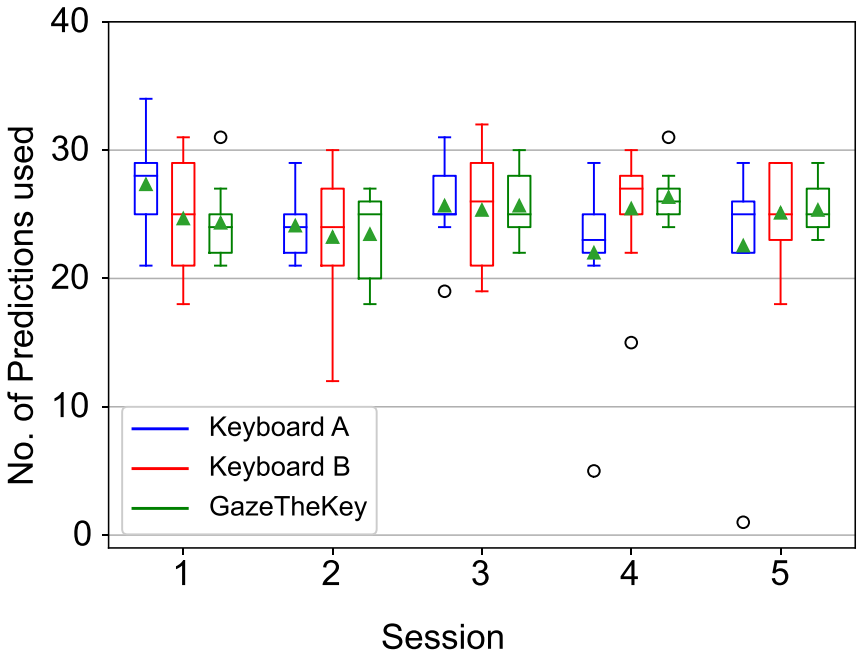


Figure 4.14: Usage of word predictions across different sessions for Keyboard A, B and *GTK*

Despite high text-prediction usage, there is no significant learning effect for any keyboards.

4.8 DISCUSSION

The experimental evaluation indicates that bringing the word predictions closer to the user's visual attention does not significantly impact text entry performance. Several implications of gaze-based interactions could arguably be the reason behind these findings.

One central observation is the inessential usage of word predictions by participants. Word predictions offer users the possibility to reduce effort by auto-completing the words. However, inter-spaced and in-letter predictions bring the word predictions in the constant visual attention of users, which might lead them to be overly reliant on predictions (as we can see with the increment of prediction usage for Keyboard B and *GTK*). We observed that the participants even picked partially relevant word predictions with additional suffixes, e.g., a participant selected the predicted word *organization* after typing *or*, and then edited the terms to write the desired word *organize*. Such instances require additional usage of backspace keys, and it

makes the actual benefit of predictions much smaller than anticipated, i.e., picking a word prediction does not necessarily correlate with fewer keystrokes to complete the desired word since it involves the editing task of the picked suggestion, which is a non-trivial task in eye typing.

Section 4.5.4 (Backspace) results confirm this assumption, as backspace usage is much higher for *GTK* and B than A. We calculated the number of backspace hits after selecting a word prediction to investigate this phenomenon further. Using backspace on selected word prediction exhibits the user picking partially relevant word predictions. Grand mean of 9 participants across 5 sessions was recorded as *2.56*, *3.67*, *5.56* for the three keyboard designs. This indicates that Keyboard B and *GTK* participants used partially correct predictions and applied more backspaces to correct the suggestions. It eventually aligns with the result on Keyboard B and *GTK* needing significantly more keystrokes than A (see Section 4.5.3), despite having similar text entry rates.

Dwelling on individual keys to compose a text is demanding and tedious. Hence, the user is keen on additional help from the system to ease the task. Word prediction helps the user in this aspect. However, like any other recommendation engine, predictions may not always be relevant and helpful for users. We can contemplate that bringing predicted words closer to user attention affects user cognition, as they become keener on picking the suggested options. However, this does not translate to the improved text entry performance.

Another reflection on performance is the rapid eye movements invalidating the effect of positioning benefit. The major variation in the design of Keyboard B and *GTK* was to bring the predictions closer to visual focus while selecting letters so that the user does not need additional time to switch attention to the external word prediction list. However, for gaze-based interaction, the fast eye movements might nullify this effect. It has been noted that eye movements are so fast that it provides an interaction medium potentially faster than the conventional mouse [166]. More specifically, eye saccades (movement between two consecutive fixations) are extremely fast movements that commonly takes 30 to 120 ms having an amplitude range between 1° and 40° (average 15° to 20°) [40]. The inspection and selection of prediction can be made quickly since it requires only one saccade or more saccades in the same direction. More specifically, for the individual keyboard designs, users retain the position information of the word list and hence can predetermine the path to reach the list. The user can *mark ahead* path [82] and hence the time can be significantly minimized. Furthermore, the variant position does not correlate with the scanning cost of word predictions. The user still has to scan for

relevant words to be picked from presented predictions irrespective of the positions, i.e., for both Keyboard A and B user has to look at all three predicted words to find out if the relevant predictions are present in the list. For Keyboard A user has to look at a distant top layout. However, the additional time required is not very significant due to fast eye movements.

Compared to touch-based text entry virtual keyboards, eyes always start moving toward the target before the hand. As eye movements are quite rapid, the eyes usually arrive at the target before the hand starts to move [1]. Touch-based input combines hand movement with eye movements since users need first to look and scan if the suggested word is relevant and then perform the selection by hand. Therefore, touch-based selections of word predictions require additional physical movement, which is not correlated with eye movements [19], i.e., hand movements need substantial time for interaction distinguished from eye movements. Hence, the keyboard designs to bring the predictions closer for touch-based inputs [55] invariably help reduce the effort of selecting predictions and improve the user experience and performance.

4.9 CONCLUSION

Word prediction is a valuable feature to enhance the typing experience. Relevant word prediction representation to end-users becomes essential for text entry with virtual keyboards. In this chapter, we assess the visual representation of word predictions by evaluating a newly designed keyboard that brings word predictions closer to the visual focus. To understand the usability of such a design, we further conducted a user study to evaluate the performance and user feedback by comparing our design with two similar dwell-time-based keyboards with the variable spatial positioning of word predictions.

The evaluation indicates that predictions near the visual fovea make users heavily dependent on the given predictions for gaze-based text entry. While this can be beneficial if the predictions are helpful, it leads to extensive usage of word predictions that could hamper usability.

The variant position does not correlate with the scanning cost of word predictions since the user must still scan for relevant words to be picked from the presented predictions. An interesting future direction would be to investigate this phenomenon in large-scale studies and understand how the scan time affects the typing process and how it can be minimized to improve performance.

This chapter sheds light on *RQ 1.1*, where we see the challenges of designing on-screen gaze-based keyboards. The detailed design investigation and experimental evaluation answers *RQ 1.2* and *RQ 1.3*, providing directions to the community on evaluating the performances of newly designed keyboards against traditional designs.

The upcoming chapter sheds light on how we can further investigate the usability and cognitive impact of new designs of on-screen gaze-based keyboards.

5

ANALYZING THE IMPACT OF COGNITIVE LOAD IN EVALUATING GAZE-BASED TYPING

In this chapter, we investigate the intricate nature of gaze-based text entry systems, specifically focusing on the cognitive load involved in using such interfaces. This chapter broadens the exploration to include cognitive aspects beyond traditional text entry metrics like words per minute, keystrokes per character, and backspace usage. Acknowledging the close relationship between gaze-based text entry, natural eye movements, and human brain cognition, we emphasize the significance of incorporating cognitive load as a key factor in evaluating the effectiveness of eye typing systems. Section 5.1 introduces the concept of cognitive load, outlining its characteristics and calculation methods, setting the stage for a comprehensive understanding of its role in gaze-based typing. Section 5.2 details our methodology, describing the experimental setup, the hardware used, participant selection, and procedures followed during the experiment. Section 5.3 presents the results of our study, offering crucial insights into the impact of cognitive load on gaze-based text entry. Finally, Section 6.9 concludes the chapter by summarizing our findings and reflecting on their implications for the future of gaze-based typing interfaces.

The EEG analysis offered insights into the cognitive variations of users during different typing phases and intervals. These findings underscore the importance of considering cognitive factors in improving the usability and efficiency of eye typing systems, paving the way for more user-centric design approaches in gaze-based text entry.

The contributions of this chapter are adapted from the full paper published at CBMS 2017[1] [163]

5.1 COGNITIVE LOAD

Different designs for gaze-based text entry have included input techniques like dwell time [101] and dwell free [133] based approaches. Another significant aspect is exploiting intelligent text prediction meth-

1 https://ieeexplore.ieee.org/xpl/conhome/8100282/proceeding

ods for more efficient text entry [98]. Moreover, the placement of word predictions [172] around the foveal region [38] has been investigated. However, how these designs impact user cognition is still being determined, i.e., if the mental effort required in the text entry process varies for different designs.

Cognitive effort could be measured by analyzing EEG signals. EEG signals have been used in different experiments [165] – along with gaze signals – to navigate different applications. Other directions are understanding artifacts caused by eye movements [136] or using EEG as event-related potentials [177]. However, EEG has rarely been used to analyze gaze-based typing, although it might provide helpful feedback about the cognitive demand of the user.

Antonenko et al. [3] define cognitive load as the load or the effort imposed on the memory by the cognitive processes involved in learning. Paas et al. [131] have extended this definition of mental effort as the cognitive capacity allocated to care for the demands imposed by a specific task. These research works – which focus on the cognitive architecture involving memory and time collectively – contribute towards a theory called *Cognitive Load Theory* (CLT) [131]. Text entry is a cognitively demanding task. While selecting letters is easy, forming words and checking their correctness involves the interaction of several information units, thus leading to a higher intrinsic load on the working memory.

The fluctuations of cognitive load across the task completion time provide us a detailed picture of where the system's usability suffered and caused the user challenges. Different techniques have been incorporated to measure this load; the NASA Task Load Index is one of the most common tests. However, when asked to be filled at the end of the experiment, these subjective ratings or scales do not shed light on the instantaneous intrinsic load. This is where physiological measurements like Electroencephalogram or Functional Magnetic Resonance Imaging play a key role in extracting such information.

Electroencephalogram (EEG) signals help record a continuous measure of cognitive load by picking up fluctuations in the signal when exposed to instantaneous load, thus providing us a better granularity in measurement. This granularity is missed if the overall cognitive load was recorded at the end of the experiment [3]. EEG signals have been used to investigate cognitive load and are researched well over years [14]. Apart from EEG, *Galvanic Skin Response* (GSR) has also been used to estimate the effect of cognitive load [30]. Compared to other available options to measure the cognitive state of an individual, such as PET, fMRI, fNRS, EEG has the advantage of both high

temporal resolution [20] and economic flexibility. The development of low-cost and light-weight EEG devices like Emotiv EPOC[2] allows researchers to investigate the domain of cognitive load easily [70]. This easy, non-invasive access to EEG signals motivated us to use such low-cost devices to study the cognitive reaction associated with gaze-based typing on virtual keyboards.

5.1.1 *EEG Signal Processing*

In this experimental investigation, we apply *Short-time Fourier Transform* (STFT) to the EEG signal time series to evaluate the cognitive load of each participant during the experiments [62]. Compared with simple *Fast Fourier Transform* (FFT), STFT can capture both time and frequency information in non-stationary signals – as in our case. We executed the following pipeline to extract cognitive load out of the raw EEG channels' data:

1. Preprocessing: We first divide each signal time series into multiple sliding windows of equal length (1024 samples, or 8 seconds) with a window slide unit of length 512 samples (4 seconds). This results in two neighboring windows sharing an overlap of 50% window length.

2. Fourier transform: In this step, a discrete Fourier transform of each windowed signal c with length N=1024 and sampling rate Fs=128Hz is performed (see Equation 5.1) [158], resulting in its spectrogram.

$$C_k = \sum_{j=0}^{N-1} c_j e^{2\pi i j k / N} \qquad k = 0, ..., N-1 \tag{5.1}$$

   This is subdivided into frequency bands [33]:

   - *Delta* (<4Hz): are the slowest of the EEG waves and can be detected during deep sleep.
   - *Theta* (≥4Hz and <8Hz): is observed during deep focus.
   - *Alpha* (≥8Hz and <14Hz): is observed when one is relaxed and awake but mostly eyes are closed.
   - *Beta* (≥14Hz): is generated during normal consciousness and active concentration.

2 https://www.emotiv.com/epoc 128Hz Sampling Rate, 14 Channels

3. *Computation of spectral power*. For a defined frequency band $[f_1, f_2]$, we can further estimate its *spectral power P* (see Equation 5.2) and the corresponding *spectral power ratio* (spectral power in a certain frequency band divided by total power in all bands).

$$P = \frac{1}{N} \sum_k |C_k|^2 \qquad k \in [\lfloor f_1 \cdot N/Fs \rfloor, \lfloor f_2 \cdot N/Fs \rfloor] \tag{5.2}$$

Studies have shown that for participants who are performing specific tasks with higher cognitive load (e.g., writing) – compared to relaxing – a higher percentage of high-frequency EEG waves (especially in the Beta band) can be observed [41]. Hence, we can compute the average value of the spectral power ratio of the Beta band in the EEG signal from all 14 channels within a time window. This serves as an indicator of the cognitive load *within this particular time window*.

5.2 METHODOLOGY

This setup was similar to the one we saw in Chapter 4.

Participants were asked to participate in three sessions on three days, each dedicated to one keyboard design. The experiment was executed in a controlled environment with artificial illumination. Latin square ordering was used for the counter-balanced setup of experimental session slots. The independent and control variables were carefully noted before the experimental process. Each participant was instructed on how the experimental process will be carried out in a short training session. They were specifically trained to read the sentence to type. The participants were also instructed on how the hinted sentence disappears on selecting the first letter. This behavior was chosen to ensure the simulation of free writing and to prohibit the participant from comparing the collected input letter-per-letter, which would influence eye gaze data [81].

5.2.1 *Participants*

The main experiment had five able-bodied male participants who were paid to participate, ages 22 to 26 years (mean = 24.2, SD = 2.17). None of the participants had prior eye typing experience and wore any corrective visual devices. Every participant was familiar with the QWERTY layout used in the designs.

5.2.2 *Apparatus*

The eye tracking setup for the experiment is similar to the setup described in Section 4.5.1 For the BCI device, Emotiv's 14-channel EPOC+ device was chosen to measure the brain signals at a sampling rate of 128Hz (Figure 5.1). The Premium SDK allowed us to extract the raw EEG data of each channel.

Figure 5.1: The image shows us the Emotiv EPOC+ headset with 14 electrode combination for capturing EEG data.

Recording and synchronization of keyboard event markers, eye tracking, and EEG data were achieved with LabStreamingLayer[3], which provided us with synchronized time stamps.

5.2.3 *Procedure*

The same procedure that is described in Section 4.5.3 has been followed here.

Each participant was requested to sign the informed consent form prior to their experimental session. They were then given an accurate description of the experiment and the devices being used. Prior to every session, the eye tracker was calibrated. In the training session, they were shown that in order to submit a typed sentence, they needed to hit the space bar on the computer's physical keyboard.

3 https://github.com/sccn/labstreaminglayer

5.3 RESULTS AND OBSERVATIONS

The experimental results provide an indication of the significant role of mental workload assessment while performing the high cognitive agility task of eye typing. We first provide the details on conventional performance metrics, present the experimental cognitive load results and discuss our findings.

5.3.1 *Performance*

Based on *words per minute* (WPM), the grand means of each of the keyboards are very close to one another: 9.20, 8.60, and 9.05 wpm for Keyboard A, B, and GTK, respectively. ANOVA for wpm values reveal a non-significant effect with $F(2,12)=0.403$, $p>0.05$ (ns).

Keystrokes per character (KPSC) [144] is another standard metric that is often used. We have adopted this concept to measure how many *keystrokes* were *saved* during a session. This reveals how text predictions positively influence typing effort, reducing the time required for typing. The average percentage of keystrokes saved was 39.0018, 35.4366, 33.4694 for the three keyboard setups. ANOVA, however, indicates a non-significant effect, with $F(2,12)=1.54$, $p>0.05$ (ns).

The *backspace key usage* is another indicative metric that hints about the number of mistakes rectified by the users while typing. Since eye typing is an exhaustive task, people often make mistakes. The average backspace usage for the three keyboards A, B, and GTK, were 2.92, 6.32, and 5.00 times. ANOVA for backspace key usage indicates a non-significant effect, with $F(2,12)=1.64$, $p>0.05$ (ns).

5.3.2 *Cognitive Load*

In this section, we compare the cognitive load of our participants in different experimental setups. As discussed in Section 5.1, we use the spectral power ratio of the Beta band of EEG signals to indicate the level of cognitive load.

Figure 5.2 shows the average cognitive effort required by participants for different keyboards. We can observe that *GTK* (with mean value 0.0824) has a lesser cognitive load compared to both Keyboard *A* (with

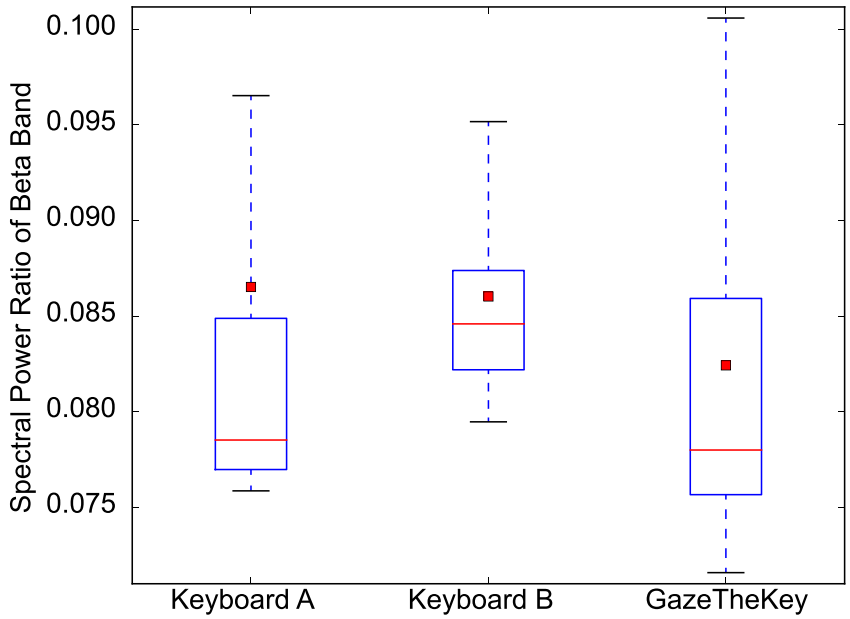


Figure 5.2: Comparison of overall cognitive load of participants using the three keyboards during the experiment. The X-axis marks the keyboard, while the Y-axis denotes the spectral power ratio of the Beta band of EEG signals, which indicates the level of cognitive load of the participant. Each data entry in the box-plot corresponds to the spectral power ratio value of one *time window* (see Section 5.1). The horizontal bar in the middle of the box shows the median value, while the red dot shows the mean value (same for all boxplots in this paper).

mean value 0.0865) and *B* (with mean value 0.0860). Shapiro-Wilk showed normal distribution of data. T-test shows that the differences are significant (Keyboard A and C with p=0.01542, N=150; Keyboard *B* and *GTK* with p=0.00047, N=150)[4]. *GTK* embeds individual suggestions on the letters itself, and hence users might have required less cognitive effort of scanning text prediction list. Keyboard *B* has similar predictions as *A* however, the dynamic appearance of the inter-spaced word list seems to confuse users, leading to high mental demand. Some participants in a parallel experiment also revealed similar observations, as they stated design B as frustrating; however, *GTK* design is more consistent.

The Shapiro-Wilk test is a hypothesis test that is applied to a sample with a null hypothesis that the sample has been generated from a normal distribution. If the p-value is low, we can reject such a null hypothesis and say that the sample has not been generated from a normal distribution.

These results do not have a direct correlation with the conventional performance metrics; however, all three metrics in Section 5.3.1 indicate a lower performance for Keyboard *B* (non-significant), and EEG analysis indicating significantly higher cognitive load, implying Keyboard *B* as a bad design choice for end-users.

4 N=5 participants * (1 training + 5 experimental sessions) * 5 sentences

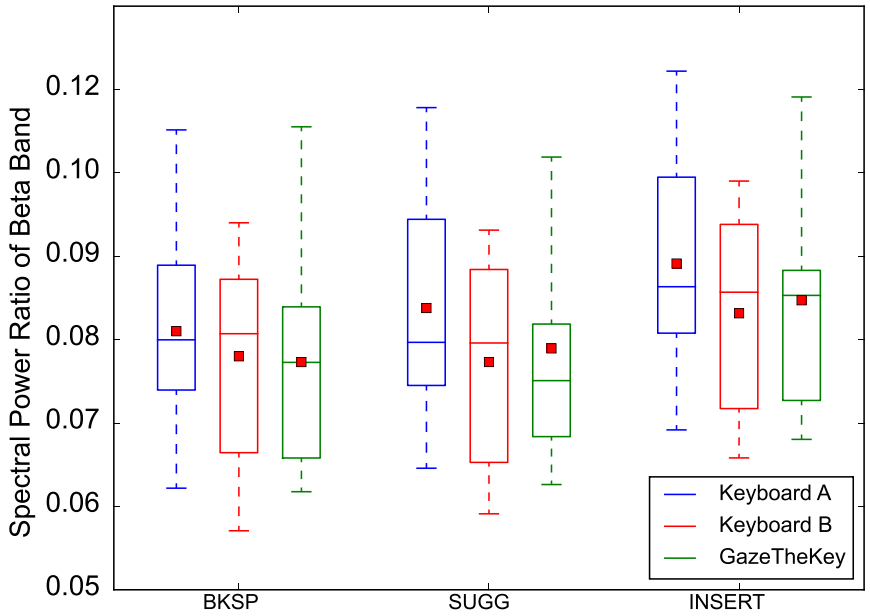


Figure 5.3: Comparison of cognitive load of participants in different typing modes using the three keyboards (shown with different colors) during the experiment. The X-axis labels different modes, while the Y-axis shows the spectral power ratio of the Beta band of EEG signals, which indicates the level of cognitive load of the participant. In each boxplot, we have 25 samples in total (outliers are omitted), corresponding to the 5 participants and five sessions for each participant.

We were also keen to investigate how different aspects of the text entry process impact user cognition. Hence we compared the cognitive load of participants in different typing *modes* when using the three keyboards:

- BKSP: the participant is deleting content by hitting the backspace key on the eye-tracking keyboard
- SUGG: the participant is selecting the suggestions provided by the eye-tracking keyboard
- INSERT: the participant is inserting single letters

Figure 5.3 reveals that the cognitive load is lower for all designs when the participants were deleting content (BKSP) or using suggestions on the keyboard (SUGG), than inserting content letter by letter (INSERT). This indicates a higher demand while selecting letters, as one needs to scan and process the information in the foveal region and then finalize which one to pick. However, when deleting letters because of a mistake, one must repeatedly fixate the backspace key. It could also

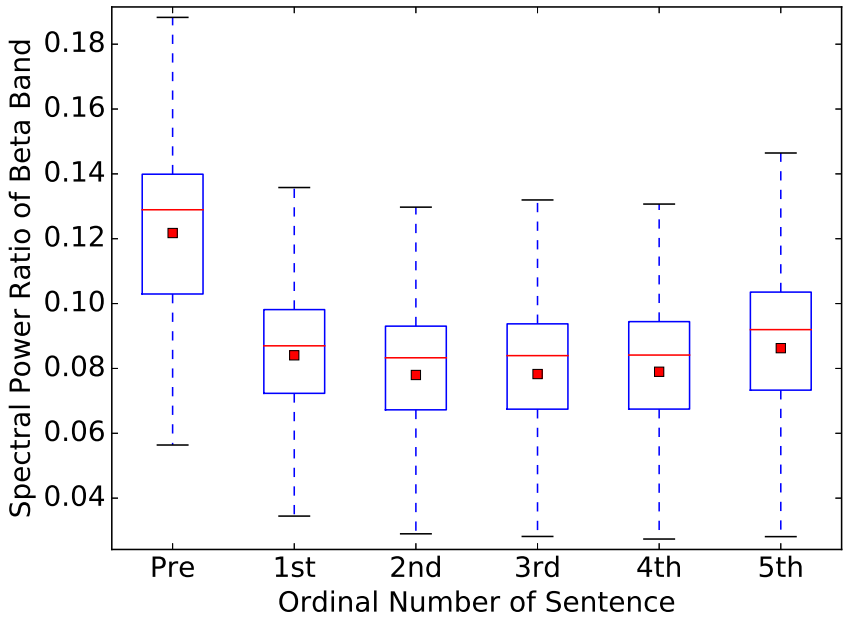


Figure 5.4: Comparison of cognitive load of participants when typing different sentences using Keyboards C during the experiment (the other two keyboards provide a similar pattern). The X-axis contains the ordinal number of the sentence in each session (Pre is the time before a keyboard is displayed and a participant asked to type). At the same time, the Y-axis shows the spectral power ratio of the Beta band of EEG signals, which indicates the level of cognitive load of the participant. In each boxplot, we have 25 samples in total (outliers are omitted), corresponding to the 5 participants and five sessions for each participant.

relate to why Keyboard *A* does not perform better despite having less backspace usage and errors. The effort required for error correction has no major impact.

Furthermore, we compared the cognitive load when the participants were in the pre-experiment phase (the time between each section starts and the first sentence is shown) to the task of typing the sentences. As shown in Figure 5.4, we do not observe a significant difference in the cognitive load among different sentences within a session. This could be explained by the fact that we randomized the order of different sentences, as some sentences might be more cognitively demanding than others. However, we observe that the cognitive load during the pre-experiment phase is higher than during actual typing. This observation indicates that users need time to adjust to the experimental gaze-based text entry environment. However, once they get used to the environment, the cognitive load is stable.

5.4 CONCLUSION

In this chapter, we conducted a small-scale experimental study to analyze the impact of cognitive load on gaze-based typing. However, a larger participant group and a longitudinal experimental design could have provided valuable insights into statistical evaluations, such as understanding the learning curve and other temporal patterns. These elements were not fully explored due to the limited sample size and the experimental design's shorter duration. We assessed virtual keyboards with variable positioning of word predictions. The results indicate the need to assess cognitive load impact in gaze-based typing scenarios. It provides a valuable direction to understand gaze-based keyboard designs from a cognitive load perspective. The alternation in word prediction positioning creates little or no effect on traditional performance metrics, but according to the EEG analysis, it is quite evident that cognitive load varies. In the future, we aim to improve the usability of gaze-based text entry by adapting the dwell time of virtual keyboards based on instantaneous cognitive load.

This chapter again sheds light on *RA 1.3* by expanding the traditional text entry system evaluation metrics. It investigates instantaneous cognitive load to discover the hidden nuances of user experience that are somehow lost in the traditional subjective evaluation process.

6

LEVERAGING ERROR CORRECTION IN VOICE-BASED TEXT ENTRY BY TALK-AND-GAZE

This chapter explores the innovative 'Talk-and-Gaze' (TaG) method, which uniquely combines voice input and eye gaze to select and correct text errors efficiently. This method capitalizes on the intuitive nature of voice input and the precision of gaze control, effectively addressing the limitations of using gaze as the sole input modality and the shortcomings of spatial guidance in voice-only systems. The TaG approach proves especially beneficial in text revision scenarios, offering a sophisticated and user-friendly enhancement to text entry processes.

Section 6.1 discusses the role of voice-based text entry, exploring the integration of voice/speech as an additional modality in interaction, its application in error corrections, and our specific research focus when using voice as an input. In Section 6.2, we describe a pilot study conducted to identify the limitations of existing systems. Section 6.3 presents findings from a small-scale experiment focused on a voice-only approach, highlighting its strengths and weaknesses. Section 6.4 introduces our design, TaG, and its unique features. Section 6.5 evaluates TaG against two other approaches through a comprehensive experiment with 12 participants. The results, detailed in Section 6.6, demonstrate significant acceptance and improvement with our design: corrections were performed more than 20% faster with dwelling than voice commands or voice-only methods; the dwelling approach required 24% less selection effort than the command approach and 11% less than voice-only error correction. Section 6.7 discusses these findings, and Section 6.8 concludes the chapter, summarizing our contributions and reflecting on their implications for the future of voice-based text entry.

The details of this chapter have been adapted from the full paper published at CHI[1] 2020 [159].

1 https://chi2020.acm.org/

6.1 VOICE-BASED TEXT ENTRY

Recent improvements in speech recognition systems [2, 157] 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 [149].

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 [147, 194]. These challenges impact the entry of text and 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 [142]. Sears et al. [157] suggest that 66% of the interaction time is spent in correcting errors with only 33% of the time used in transcribing. Karat et al. [68] note that the assumed productivity gain for speech dictation systems depreciates when error correction is factored in.

Target-based Navigation: The user identifies a target or destination and issues an appropriate command to trigger the target. For example: "Select Friday"; where Select is the command and Friday is the target word.

One challenge for voice input is the inability to provide spatial information naturally.

In order to correct an error, the first task is to navigate to the location of the error. However, navigation by voice is a challenge. Strategies include target-based navigation or direction-based navigation [36, 107, 114]. 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 [60, 128–130, 139].

Direction-based Navigation: The user specifies the direction and distance to navigate to a desired location. For example: "Move three words left"; where Move is the command, three words is the distance, and left is the direction.

6.1.1 *Integration of Additional Modality*

Most approaches involving an additional modality require physical input, which presents a challenge when the hands are used for an activity other than typing. Some users may lack the fine motor control required to place a pen or similar device accurately. Thus, the need for digital inclusion has 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 [111, 140] and text entry and editing [69, 99, 161, 162]. Although gaze has the potential to complement the voice as an input modality, there is little research [141] that combines voice as the primary modality with gaze as the secondary modality.

Oviatt et al. [127, 130] 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 a unimodal approach. Similar experiments by Mantravadi [109] combined voice and gaze for menu selections and showed improved accuracy and less ambiguity with a multimodal approach. Kumar et al. [79] combined gaze and keyboard with "look-press-look-release" interaction for web navigation. Sengupta et al. [160] 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. [26] also found improved performance in a hands-free multimodal environment.

6.1.2 *Using Voice and Gaze for Error Correction*

Beelders et al. [17] 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. [141]. 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.

6.1.3 *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 gaze defines the spatial position in the text. 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 [65]). 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.

In this chapter, we address the research question RQ2 and the subsequent sub questions associated with it (Chapter 1).

We performed a comparative evaluation of D-TaG, V-TaG, and Voice-Only error correction to answer these questions. 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.

6.2 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 considered 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.

6.3 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 [176].

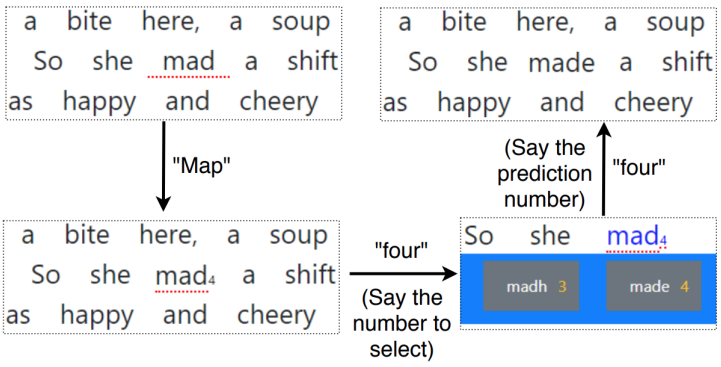
The command "map" assigns a unique number to each erroneous word in a passage for our implementation. The participant then utters the number to select a word. This eliminates the need to recall commands and allows the participant to select an incorrect word directly. It also eliminates the challenge 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 6.1):

1. *Delete* – delete the currently selected word.

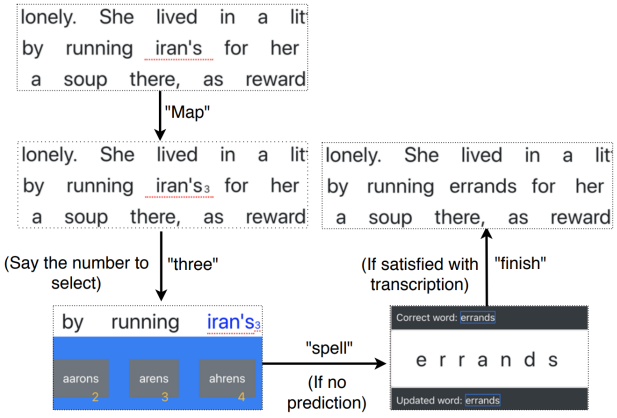
2. *Spell* – substitute the currently selected word by a new word that is spelled. This mode is introduced as re-speaking the incorrect word often does not lead to correct recognition.

3. *Case Change* – toggle the case of a letter in a word that has been accidentally capitalized or needs capitalization.

The map functionality also extends to the "spell" mode where the participant performs letter-level correction for incorrect transcriptions caused by homophones, diction, or ambient noise. "Spell" mode allows spelling the word in case recognition error occurs multiple times. The workflow of the Voice-Only approach is seen in Figure 6.1.



(a)



(b)

Figure 6.1: Voice-only edit method using "map" functionality. (a)Workflow of Voice-only approach using available predictions; (b)Workflow of Voice-only approach using "SPELL" mode

6.3.1 *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 experiences.

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

2. Predictions helped to quicken correction

3. Advantage of not adhering to one error correction mode - 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 the "map" command to select errors led to discomfort for some users.

6.4 TAG: AUGMENT VOICE-BASED TEXT INPUT WITH GAZE

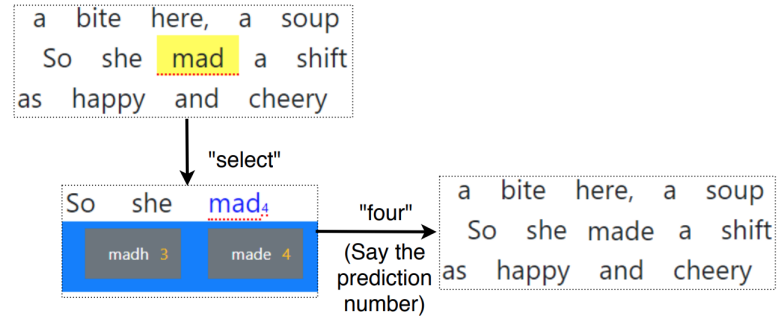
From the feedback of the second pilot study, we understood that the map-based approach helps in minimizing navigational commands and ambiguity of word selection. However, it introduced an intermediate step in error correction. Our design, TaG, augments voice with gaze to facilitate faster error selection followed by a correction to reduce this for error correction.

A common challenge of gaze-based activation is Midas Touch [65]. This leads to incorrect triggering and eventual frustration. For our TaG method, we have examined two approaches to alleviate this:

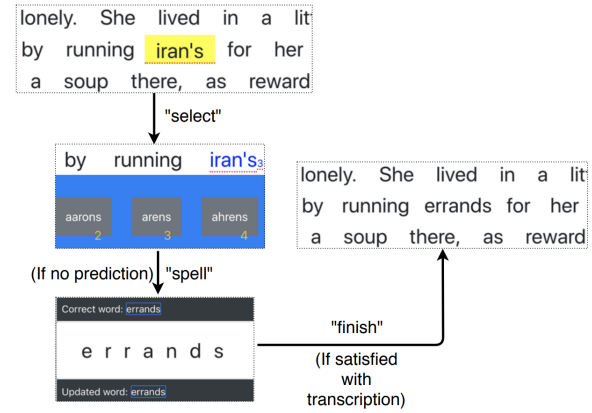
1. D-TaG – Gazing and then dwelling on the incorrect word for 0.8 seconds selects the word and triggers the text predictions. While this minimizes the number of interaction steps, the risk of Midas Touch, or inadvertent triggering, remains. The dwell time was the average duration participants took between observing the

incorrect transcription and calling out the mapped number in the second pilot study. The workflow is seen in Figure 6.3.

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 6.2.



(a)



(b)

Figure 6.2: Voice-only edit method using "map" functionality. (a)Workflow of V-TaG approach using available predictions (b)Workflow of V-TaG approach using "SPELL" mode

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.

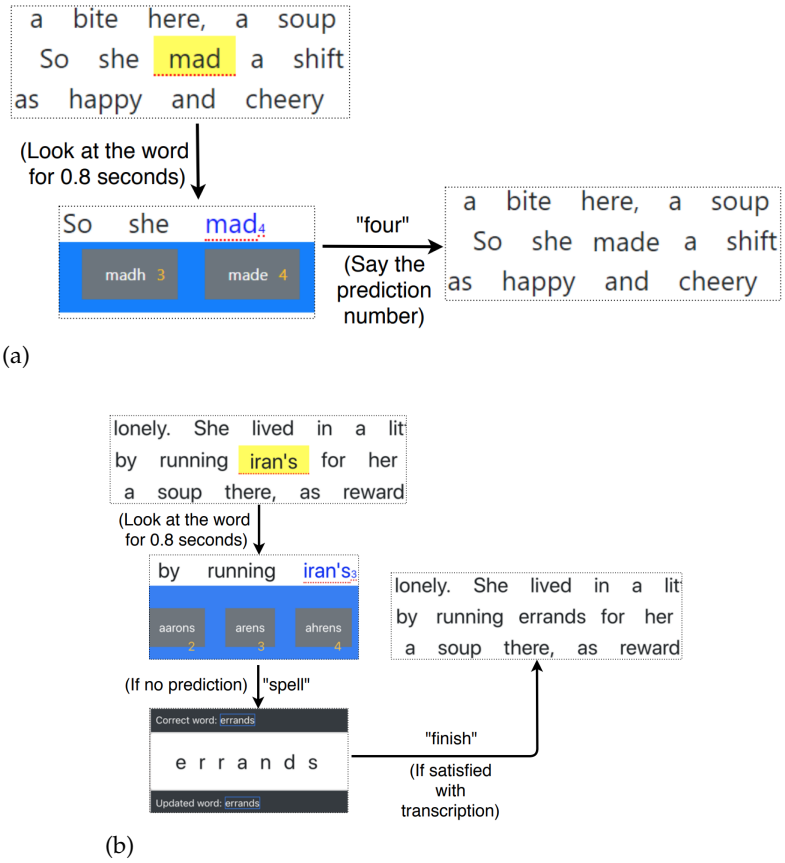


Figure 6.3: Voice-only edit method using "map" functionality. (a)Workflow of D-TaG approach using available predictions (b)Workflow of D-TaG approach using "SPELL" mode

6.5 EXPERIMENT

6.5.1 *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 to select erroneous words for five of the participants during their training process. Non-recognition or misrecognition of the keywords was due to the influence of heavy native language accents, which 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 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.

6.5.2 *Apparatus*

A Tobii EyeX[2]platform was 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 6.4. The experiments were conducted in an environment with controlled ambient light and sound. The software to evaluate the interactions was made on React Native[3] which recorded the participant's performance. Data were stored in a *.csv* file for further evaluation.

Figure 6.4: Experimental setup showing the fixed display with the eye tracker, the stand-alone microphone, and a participant performing error corrections in "spell" mode.

6.5.3 *Tasks*

Evaluations of text entry and correction systems often employ a *copy task* where the participant copies text and then fixes errors if any oc-

2 https://help.tobii.com/hc/en-us/categories/201185405-EyeX
3 https://facebook.github.io/react-native/

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

Table 2: Types of errors in the read and correct task. (For example: Missing - terible → terrible; Extra - hers → her; Double - upp → up; Mistake - want → went)

curred. Voice-based text entry evaluations frequently follow a similar protocol. The disadvantage is that copying involves cognitive overhead; that is, reading than typing; this is atypical of most real-world situations.

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 [167]. 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 6.3, 6.2). Underlining the error excluded visual search time from the interaction.

Image Description Task. Dunlop et al. [39] 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. [39].

6.5.4 *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). Afterward, the eye tracker was calibrated 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

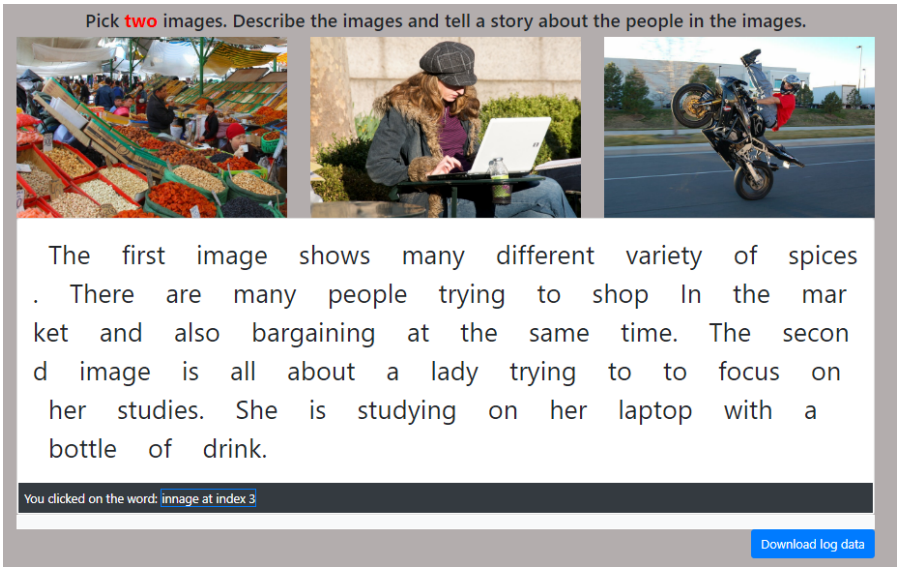


Figure 6.5: Image Description Task: Participants describe the images freely without any assistive visual marker to show errors.

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, a SUS questionnaire, and an additional questionnaire. Testing took approximately 60 minutes per participant for each edit method. Participants were told that their gaze data would be recorded for evaluation purposes. Testing for each task included a screen recording for further analysis to understand the ease of selecting erroneous words.

For the *read and correct task*, the experiment consisted of a training *block* followed by five testing blocks. The passages were taken from American short stories[4]. Each passage was around 90 words, which covered 50% of the screen space. The errors were chosen to include misspellings, incorrect letter entries, missing letters, and toggled order of letters. Table 2 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 6.5) for each *block* and were asked to de-

4 https://americanliterature.com/home

scribe any two. When they were satisfied with the transcription and corrections, they could go to the next image, 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 6.6.

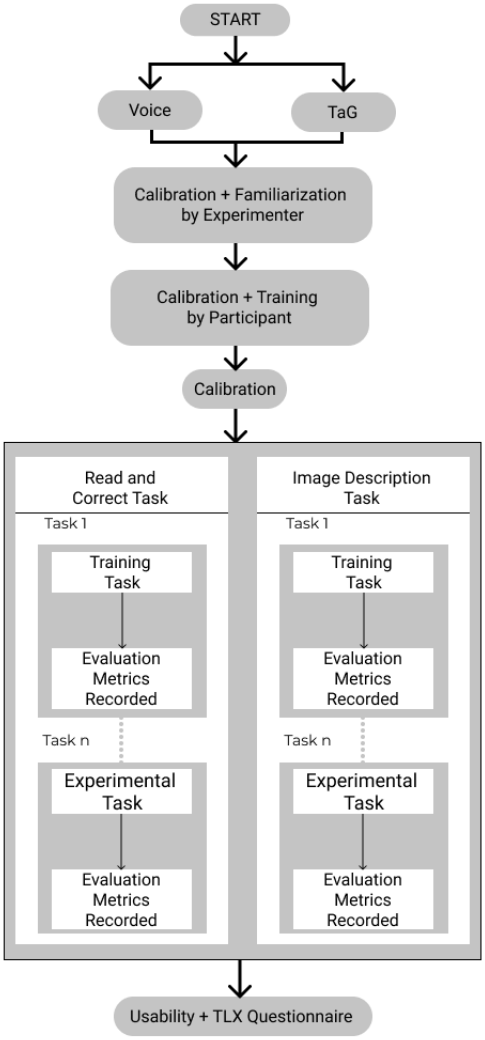


Figure 6.6: Experimental procedure for Voice-only, D-TaG and V-Tag edit methods

6.5.5 *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)

Each block included three passages, each with five errors for correction

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 the 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.

6.6 RESULTS

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

1. Read And Correct Task

    - *Objective Measure*

        – 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 6.7. Using a repeated-measures ANOVA, the differences were deemed statistically significant for edit method ($F_{2,22}$ = 11.5, p = .0004, η^2= 0.239) 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

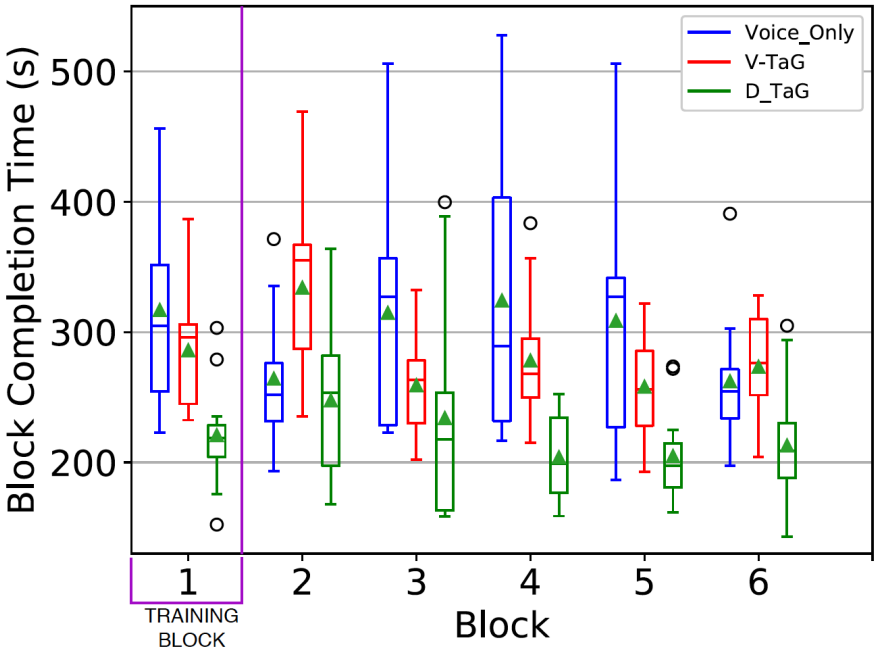


Figure 6.7: Block completion time (s) by edit method and block.

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 challenging 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 required 24.3% less selection effort than V-TaG and 10.9% less selection effort than Voice-only. There was an improvement with practice, with means of 1.42 in block 2 and falling to 1.24 in block 6. See Figure 6.8. 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 Measure*: 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 the NASA TLX

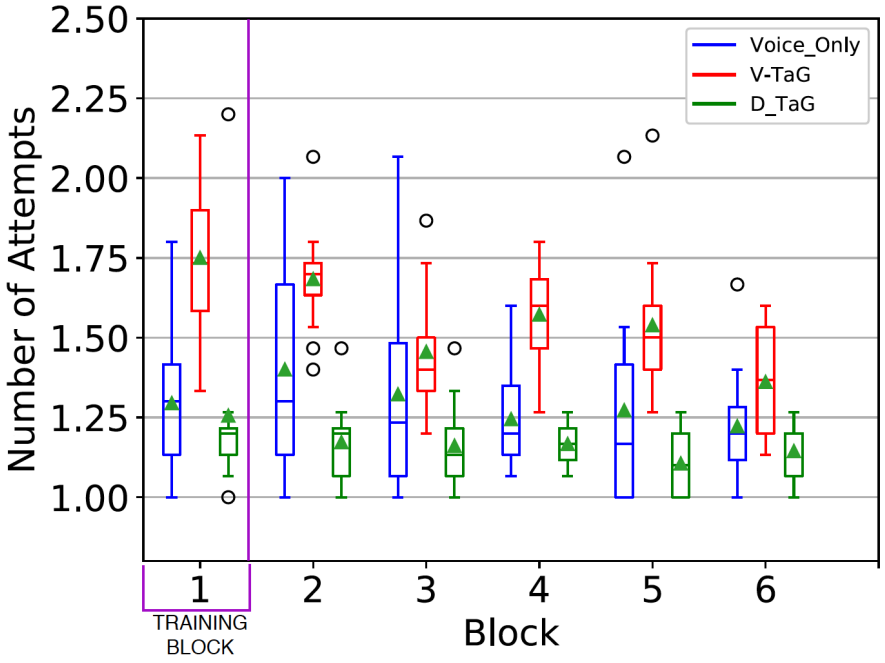


Figure 6.8: Selection effort (count) by edit method and block. (Selection effort is the number of attempts to select an erroneous word)

questionnaire [58] and the usability of the edit method using the System Usability Scale (SUS) [22]. 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), 50.9 (V-TaG), and 31.2 (D-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 ($\chi^2(2) = 4.67, p = .097$).

Interviewing participants 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.[5] 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 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 6.9.

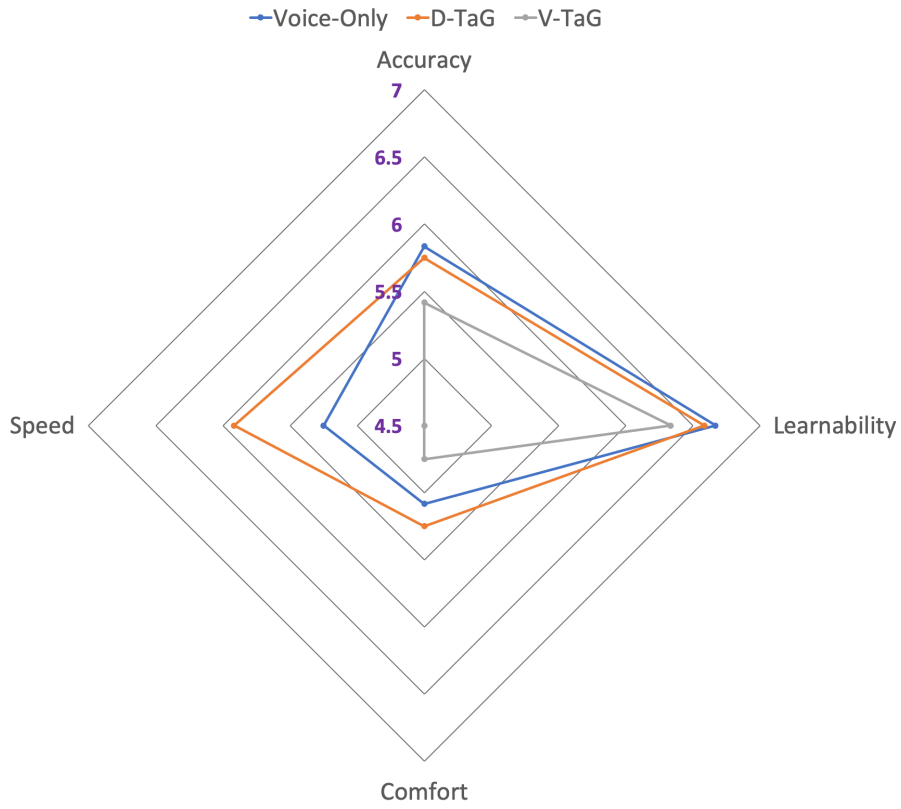


Figure 6.9: Read and Correct Task – average perceived performance on a 1-7 scale with higher scores preferred

5 https://measuringu.com/sus/

2. Image Description Task: A free text formation task was performed, asking participants to describe images presented before them (Figure 6.5). A qualitative evaluation was performed based on their performance.

    - *Subjective Feedback*

        – 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 favor. This was followed by Voice-only and finally V-TaG.

        – Comfort: As seen in Figure 6.10, 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 the 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 6.10), 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: Some participants had difficulty selecting the erroneous word by gaze and confirming it by speaking "select" for the V-TaG edit method. This led V-TaG to the lowest perceived accuracy compared to the other edit methods. While they were comfortable with the selection in D-TaG, participants also appreciated the voice-only approach.

        – 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.

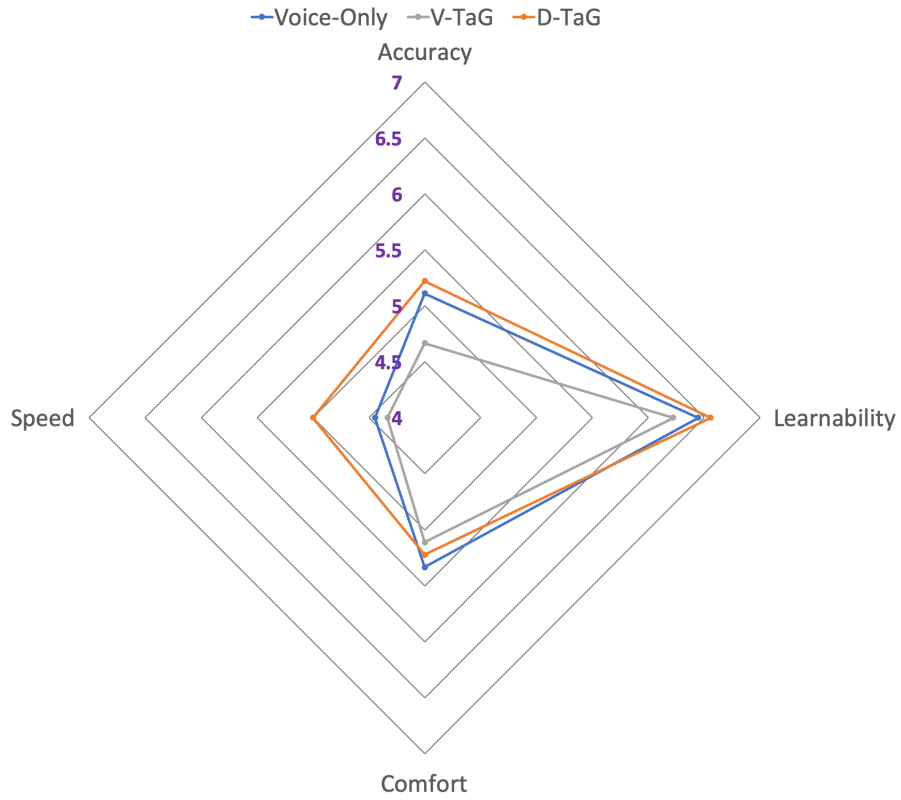


Figure 6.10: Image Description Task – perceived performance on a 1-7 scale with higher scores preferred

6.7 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 the 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. The perceived speed also supports this in the qualitative evaluation. One reason is the speed of gaze input for spatial exploration in the context of user interfaces. This work 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 the 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

performs better in perceived comfort. Designers can leverage these observations 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 questionnaire responses show that the gaze-based D-TaG interface was considered more comfortable than 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, forcing 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 into a fall-back mode that 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.

6.8 LIMITATIONS AND FUTURE WORK

1. Phonetically similar words were often incorrectly rendered (e.g., "little" vs. "Lidl", "year" vs. "yeah"). Incorrect recognition increases error correction time and 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 stand-alone 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 [153, 200].

3. This experiment did not provide visual feedback for the dwelling. Future work would include visual feedback as used in many gaze-based selection methods [77, 112].

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 text editing scenario only considered single-screen text. However, our approach supports editing 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.

6. 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.

6.9 CONCLUSION

Voice-based input offers a fast, hands-free approach to text insertion. With the advancement of several robust language models, speech-to-text ensures that a relatively correct transcription is possible in the English language. However, the challenge still lies in the process of revising or or correcting the incorrect transcription.

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, leading to lower cognitive load and higher usability scores. The work work also highlights the need of a fall back mode when one input modality fails to perform. It ensures that the process of text entry and revision continues.

Our novel approach could be extended to different styles of text editing, thereby expanding the potential of voice and gaze for text-based interactions.

This chapter sheds light on *RQ 2* by showcasing how we can improve hands-free text entry by integrating voice with gaze modality.

7

CONCLUSION AND OUTLOOK

The design and development of text entry methods is a highly complex and challenging task that always entails how efficient and user-friendly the design and interaction are for the user to perform their input actions. In her doctoral dissertation, Anna Felt points out that for the most straightforward keyboard design with 26 letters, one can generate 10^{26} design combinations [44]. This information aptly sheds light on the various keyboard designs found in research and discussed in this thesis. Moreover, this is just the standard keyboard design without considering word predictions.

The growth of online content and our subsequent consumption of it has increased manifolds. Growth is observed in additional domains like entertainment, e-commerce, and information. This massive shift in information consumption also needs great impetus in understanding and adopting content for alternative modalities. Text entry in such context thus becomes a pertinent topic of discussion. Without it, access and interaction to content become extremely limited - not just from an accessibility standpoint but also from a situational impairment scenario. In this thesis, we investigated and enhanced the text entry system's usability with alternate modalities and highlighted the need for understanding instantaneous cognitive load.

This thesis focuses on two alternative modalities for text entry systems: gaze and voice. Our investigation conducted several objective and subjective evaluations to understand the newly designed text entry system and how users felt. The first part of our work focused on increasing the number of text predictions and bringing them closer to the visual fovea for gaze-based predictive keyboard design. Chapter 4 includes a detailed design investigation and evaluation of three keyboard designs to understand the impact of design effectiveness. Chapter 5 discusses an essential contribution of measuring cognitive load while participants experimented. This novel investigation showed how interacting with different designs impacts our cognitive load. The second part investigated the text revision scenario, used voice as the primary modality, and investigated how adding gaze as a secondary modality could improve performance and overall user experience. Chapter 6 explored the possibility of voice being a primary modality

and gaze being a secondary for text editing scenarios. Our design and evaluation showed improved results in a hands-free text editing context.

In both cases, we tried to investigate the two core aspects of measuring user experience: (i) Usability (ease of use) and (ii) Effectiveness (speed, error reduction, etc.). While significant works focused on effectiveness to define user experience, usability and measuring cognitive load to establish usability is a core contribution that this thesis upholds.

Cognitive load has been an important stakeholder in understanding the usability of a system. Surveys like the System Usability Scale (SUS) or NASA Task Load Index (NASA TLX) have thus been very important. Our investigation took it one notch higher to measure and understand instantaneous cognitive load with a non-invasive EEG device. Although we explored the cognitive load for gaze-based text input, such evaluation was not done for multimodal systems. Thus, in the future, the focus would be to measure cognitive information for voice-based interaction and when modality switching happens for multimodal systems.

7.1 OUTLOOK

The investigations conducted for this thesis have far-reaching implications that extend beyond predictive keyboard design employing alternative modalities. The fusion of these two modalities holds the potential for diverse applications in text entry, encompassing tasks as varied as form filling and programming. While the primary objective of this thesis was to delve into the intricacies of predictive text entry and enhance the overall user experience, the findings have elucidated the feasibility of such approaches in an array of scenarios where physical constraints necessitate deviation from the conventional keyboard mouse interface.

Our innovative multimodal approach and keyboard design could be helpful within mixed-reality environments such as Microsoft's Hololens or Facebook's Quest. Even in these advanced settings, logging into applications or networks relies on the laborious clicking of individual keys on a virtual keyboard. Our work in this area opens up exciting possibilities for simplifying and enhancing the user experience within such immersive environments.

Expanding our design in accessibility, a combination of voice and gaze, or any other pointing modality like gesture could benefit older adults

where seamless technology integration is paramount. By aligning technological interfaces with more natural modes of interaction, we mitigate the steep learning curve among older generations. Whether through gaze tracking, voice commands, or gesture-based inputs, these alternative modalities can serve as valuable additions to the existing input methods, fostering inclusivity and usability.

Additionally, our research has delved into EEG signal analysis to comprehend cognitive load. As the market witnesses bio-sensor proliferation and integration into commonplace devices such as smartwatches, the potential for enhancing the user experience becomes increasingly tangible. Smart glasses and virtual or augmented reality headsets can seamlessly incorporate EEG sensors, facilitating real-time cognitive load assessment. Such integration promises to dynamically tailor user interfaces based on cognitive demands, optimizing user experiences across a broad spectrum of applications and contexts.

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

[1] Richard A Abrams, David E Meyer, and Sylvan Kornblum. "Eye-hand coordination: oculomotor control in rapid aimed limb movements." In: *Journal of Experimental Psychology: Human Perception and Performance* 16.2 (1990), pp. 248–267. DOI: 10.1037/0096-1523.16.2.248.

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

[3] Pavlo Antonenko, Fred Paas, Roland Grabner, and Tamara van Gog. "Using Electroencephalography to Measure Cognitive Load." In: *Educational Psychology Review* 22.4 (2010), pp. 425–438. ISSN: 1573-336X. DOI: 10.1007/s10648-010-9130-y. URL: https://doi.org/10.1007/s10648-010-9130-y.

[4] Ahmed Sabbir Arif and Ali Mazalek. "A Survey of Text Entry Techniques for Smartwatches." In: *Human-Computer Interaction. Interaction Platforms and Techniques*. Ed. by Masaaki Kurosu. Cham: Springer International Publishing, 2016, pp. 255–267.

[5] Daniel Ashbrook, Patrick Baudisch, and Sean White. "Nenya: Subtle and Eyes-Free Mobile Input with a Magnetically-Tracked Finger Ring." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '11. Vancouver, BC, Canada: Association for Computing Machinery, 2011, 2043–2046. ISBN: 9781450302289. DOI: 10.1145/1978942.1979238. URL: https://doi.org/10.1145/1978942.1979238.

[6] Michael Ashmore, Andrew T. Duchowski, and Garth Shoemaker. "Efficient Eye Pointing with a Fisheye Lens." In: *Proceedings of Graphics Interface 2005*. GI '05. Victoria, British Columbia: Canadian Human-Computer Communications Society, 2005, 203–210. ISBN: 1568812655.

[7] Vikas Ashok, Yevgen Borodin, Yury Puzis, and I. V. Ramakrishnan. "Capti-Speak: A Speech-Enabled Web Screen Reader." In: *Proceedings of the 12th International Web for All Conference*. W4A '15. Florence, Italy: Association for Computing Machinery, 2015. ISBN: 9781450333429. DOI: 10.1145/2745555.2746660. URL: https://doi.org/10.1145/2745555.2746660.

[8] Behrooz Ashtiani and I. Scott MacKenzie. "BlinkWrite2: An Improved Text Entry Method Using Eye Blinks." In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ETRA '10. Austin, Texas: Association for Computing Machinery, 2010, 339–345. ISBN: 9781605589947. DOI: 10.1145/1743666.1743742. URL: https://doi.org/10.1145/1743666.1743742.

[9] Behrooz Ashtiani and I. Scott MacKenzie. "BlinkWrite2: An Improved Text Entry Method Using Eye Blinks." In: *Proceedings of the 2010 Symposium on Eye-Tracking Research amp; Applications*. ETRA '10. Austin, Texas: Association for Computing Machinery, 2010, 339–345. ISBN: 9781605589947. DOI: 10.1145/1743666.1743742. URL: https://doi.org/10.1145/1743666.1743742.

[10] Jonas Austerjost, Marc Porr, Noah Riedel, Dominik Geier, Thomas Becker, Thomas Scheper, Daniel Marquard, Patrick Lindner, and Sascha Beutel. "Introducing a Virtual Assistant to the Lab: A Voice User Interface for the Intuitive Control of Laboratory Instruments." In: *SLAS TECHNOLOGY: Translating Life Sciences Innovation* 23.5 (2018). PMID: 30021077, pp. 476–482. DOI: 10.1177/2472630318788040. eprint: https://doi.org/10.1177/2472630318788040. URL: https://doi.org/10.1177/2472630318788040.

[11] Shiri Azenkot and Nicole B. Lee. "Exploring the Use of Speech Input by Blind People on Mobile Devices." In: *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '13. Bellevue, Washington: Association for Computing Machinery, 2013. ISBN: 9781450324052. DOI: 10.1145/2513383.2513440. URL: https://doi.org/10.1145/2513383.2513440.

[12] Shiri Azenkot and Shumin Zhai. "Touch Behavior with Different Postures on Soft Smartphone Keyboards." In: *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '12. San Francisco, California, USA: Association for Computing Machinery, 2012, 251–260. ISBN: 9781450311052. DOI: 10.1145/2371574.2371612. URL: https://doi.org/10.1145/2371574.2371612.

[13] Robert W Baloh, Andrew W Sills, Warren E Kumley, and Vicente Honrubia. "Quantitative measurement of saccade amplitude, duration, and velocity." In: *Neurology* 25.11 (1975), pp. 1065–1065.

[14] Erol Başar. *Brain Function and Oscillations: Volume II: Integrative Brain Function. Neurophysiology and Cognitive Processes*. 2012.

[15] Nikolaus Bee and Elisabeth André. "Writing with Your Eye: A Dwell Time Free Writing System Adapted to the Nature of Human Eye Gaze." In: *Perception in Multimodal Dialogue Systems*. Ed. by Elisabeth André, Laila Dybkjær, Wolfgang Minker, Heiko Neumann, Roberto Pieraccini, and Michael Weber. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 111–122. ISBN: 978-3-540-69369-7.

[16] T. R. Beelders and P. J. Blignaut. "Using Vision and Voice to Create a Multimodal Interface for Microsoft Word 2007." In: *Proceedings of the 2010 Symposium on Eye-Tracking Research amp; Applications*. ETRA '10. Austin, Texas: Association for Computing Machinery, 2010, 173–176. ISBN: 9781605589947. DOI: 10.1145/1743666.1743709. URL: https://doi.org/10.1145/1743666.1743709.

[17] T. R. Beelders and P. J. Blignaut. "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. Austin, Texas: ACM, 2010, pp. 173–176. ISBN: 978-1-60558-994-7. DOI: 10.1145/1743666.1743709. URL: http://doi.acm.org/10.1145/1743666.1743709.

[18] Mafkereseb Kassahun Bekele, Roberto Pierdicca, Emanuele Frontoni, Eva Savina Malinverni, and James Gain. "A Survey of Augmented, Virtual, and Mixed Reality for Cultural Heritage." In: *J. Comput. Cult. Herit.* 11.2 (2018). ISSN: 1556-4673. DOI: 10.1145/3145534. URL: https://doi.org/10.1145/3145534.

[19] Harold Bekkering, Jos J. Adam, Herman Kingma, A. Huson, and H. T. A. Whiting. "Reaction time latencies of eye and hand movements in single- and dual-task conditions." In: *Experimental Brain Research* 97.3 (1994), pp. 471–476. ISSN: 1432-1106. DOI: 10.1007/BF00241541. URL: https://doi.org/10.1007/BF00241541.

[20] Chris Berka, Daniel J. Levendowski, Milenko M. Cvetinovic, Miroslav M. Petrovic, Gene Davis, Michelle N. Lumicao, Vladimir T. Zivkovic, Miodrag V. Popovic, and Richard Olmstead. "Real-Time Analysis of EEG Indexes of Alertness, Cognition, and Memory Acquired With a Wireless EEG Headset." In: *International Journal of Human–Computer Interaction* 17.2 (2004), pp. 151–170. DOI: 10.1207/s15327590ijhc1702_3. eprint: https://doi.org/10.1207/s15327590ijhc1702_3. URL: https://doi.org/10.1207/s15327590ijhc1702_3.

[21] Richard A. Bolt. ""Put-That-There": Voice and Gesture at the Graphics Interface." In: 14.3 (1980), 262–270. ISSN: 0097-8930. DOI: 10.1145/965105.807503. URL: https://doi.org/10.1145/965105.807503.

[22] John Brooke et al. "SUS-A quick and dirty usability scale." In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7.

[23] Stuart K Card. *The psychology of human-computer interaction*. Crc Press, 2018.

[24] Stuart K Card, Jock D Mackinlay, and George G Robertson. "A morphological analysis of the design space of input devices." In: *ACM Transactions on Information Systems (TOIS)* 9.2 (1991), pp. 99–122.

[25] Marcus Carter, Fraser Allison, John Downs, and Martin Gibbs. "Player Identity Dissonance and Voice Interaction in Games." In: *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*. CHI PLAY '15. London, United Kingdom: Association for Computing Machinery, 2015, 265–269. ISBN: 9781450334662. DOI: 10.1145/2793107.2793144. URL: https://doi.org/10.1145/2793107.2793144.

[26] Emiliano Castellina, Fulvio Corno, and Paolo Pellegrino. "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, 2008, pp. 79–82. DOI: 10.1145/1344471.1344492.

[27] Jason W Clark, Rathinavelu Chengalvarayan, Timothy J Grost, Dana B Fecher, and Jeremy M Spaulding. *Voice dialing using a rejection reference*. US Patent 8,055,502. 2011.

[28] Leigh Clark et al. "What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents." In: New York, NY, USA: Association for Computing Machinery, 2019, 1–12. ISBN: 9781450359702. URL: https://doi.org/10.1145/3290605.3300705.

[29] Edward Clarkson, James Clawson, Kent Lyons, and Thad Starner. "An Empirical Study of Typing Rates on Mini-QWERTY Keyboards." In: *CHI '05 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '05. Portland, OR, USA: Association for Computing Machinery, 2005, 1288–1291. ISBN: 1595930027. DOI: 10.1145/1056808.1056898. URL: https://doi.org/10.1145/1056808.1056898.

[30] Dan Conway, Ian Dick, Zhidong Li, Yang Wang, and Fang Chen. "The Effect of Stress on Cognitive Load Measurement." In: *Human-Computer Interaction – INTERACT 2013*. Ed. by Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 659–666. ISBN: 978-3-642-40498-6.

[31] Vivian Cook. "Standard Punctuation and the Punctuation of the Street." In: *Essential Topics in Applied Linguistics and Multilingualism: Studies in Honor of David Singleton*. Ed. by Mirosław Pawlak and Larissa Aronin. Cham: Springer International Publishing, 2014, pp. 267–290. ISBN: 978-3-319-01414-2. DOI: 10.1007/978-3-319-01414-2_16. URL: https://doi.org/10.1007/978-3-319-01414-2_16.

[32] Justin Cuaresma and I Scott MacKenzie. "A study of variations of Qwerty soft keyboards for mobile phones." In: *Proceedings of the International Conference on Multimedia and Human-Computer Interaction-MHCI*. 2013, pp. 126–1.

[33] G eds DEUSCHL. "Recommendations for the Practice of Clinical Neurophysiology." In: *Guidelines of the International Federation of Clinical Neurophysiology* (1999). URL: https://ci.nii.ac.jp/naid/10011547941/en/.

[34] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. "Wizard of Oz studies — why and how." In: *Knowledge-Based Systems* 6.4 (1993). Special Issue: Intelligent User Interfaces, pp. 258–266. ISSN: 0950-7051. DOI: https://doi.org/10.1016/0950-7051(93)90017-N. URL: https://www.sciencedirect.com/science/article/pii/095070519390017N.

[35] Alexander De Luca, Martin Denzel, and Heinrich Hussmann. "Look into My Eyes! Can You Guess My Password?" In: *Proceedings of the 5th Symposium on Usable Privacy and Security*. SOUPS '09. Mountain View, California, USA: Association for Computing Machinery, 2009. ISBN: 9781605587363. DOI: 10.1145/1572532.1572542. URL: https://doi.org/10.1145/1572532.1572542.

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

[37] Laurence Devillers et al. "Multifaceted Engagement in Social Interaction with a Machine: The JOKER Project." In: *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. 2018, pp. 697–701. DOI: 10.1109/FG.2018.00110.

[38] Antonio Diaz-Tula and Carlos H. Morimoto. "AugKey: Increasing Foveal Throughput in Eye Typing with Augmented Keys." In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: Association for Computing Machinery, 2016, 3533–3544. ISBN: 9781450333627. DOI: 10.1145/2858036.2858517. URL: https://doi.org/10.1145/2858036.2858517.

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

[40] "Eye Tracking Methodology; Theory and Practice." In: *Qualitative Market Research: An International Journal* 10.2 (2007), pp. 217–220. ISSN: 1352-2752. DOI: 10.1108/13522750710740862. URL: https://doi.org/10.1108/13522750710740862.

[41] C. W. N. F. Che Wan Fadzal, W. Mansor, L. Y. Khuan, and A. Zabidi. "Short-time Fourier Transform analysis of EEG signal from writing." In: *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*. 2012, pp. 525–527. DOI: 10.1109/CSPA.2012.6194785.

[42] Reza Fazel-Rezai, Brendan Z Allison, Christoph Guger, Eric W Sellers, Sonja C Kleih, and Andrea Kübler. "P300 brain computer interface: current challenges and emerging trends." In: *Frontiers in neuroengineering* 5 (2012), p. 14.

[43] Reza Fazel-Rezai, Brendan Allison, Christoph Guger, Eric Sellers, Sonja Kleih, and Andrea Kübler. "P300 brain computer interface: current challenges and emerging trends." In: *Frontiers in Neuroengineering* 5 (2012), p. 14. ISSN: 1662-6443. DOI: 10.3389/fneng.2012.00014. URL: https://www.frontiersin.org/article/10.3389/fneng.2012.00014.

[44] Anna Maria Feit. "Assignment Problems for Optimizing Text Input." English. Doctoral thesis. School of Electrical Engineering, 2018, 182 + app. 56. ISBN: 978-952-60-8016-1 (electronic), 978-952-60-8015-4 (printed). URL: http://urn.fi/URN:ISBN:978-952-60-8016-1.

[45] Torsten Felzer and Bernd Freisleben. "HaWCoS: The "Hands-Free" Wheelchair Control System." In: *Proceedings of the Fifth International ACM Conference on Assistive Technologies*. Assets '02. Edinburgh, Scotland: Association for Computing Machinery, 2002, 127–134. ISBN: 1581134649. DOI: 10.1145/638249.638273. URL: https://doi.org/10.1145/638249.638273.

[46] Torsten Felzer, Ian Scott MacKenzie, Philipp Beckerle, and Stephan Rinderknecht. "Qanti: A Software Tool for Quick Ambiguous Non-standard Text Input." In: *Computers Helping People with Special Needs*. Ed. by Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 128–135.

[47] Torsten Felzer, Ian Scott MacKenzie, Philipp Beckerle, and Stephan Rinderknecht. "Qanti: A Software Tool for Quick Ambiguous Non-standard Text Input." In: *Computers Helping People with Special Needs*. Ed. by Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 128–135.

[48] Paul M Fitts. "The information capacity of the human motor system in controlling the amplitude of movement." In: *Journal of experimental psychology* 47.6 (1954), p. 381.

[49] Christopher Frauenberger and Tony Stockman. "Auditory display design—An investigation of a design pattern approach." In: *International Journal of Human-Computer Studies* 67.11 (2009). Special issue on Sonic Interaction Design, pp. 907–922. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2009.05.008. URL: https://www.sciencedirect.com/science/article/pii/S1071581909000676.

[50] Nestor Garay-Vitoria and Julio Abascal. "Text prediction systems: a survey." In: *Universal Access in the Information Society* 4.3 (2006), pp. 188–203.

[51] Nestor Garay-Vitoria and Julio Abascal. "Text prediction systems: a survey." In: *Universal Access in the Information Society* 4.3 (2006), pp. 188–203. ISSN: 1615-5297. DOI: 10.1007/s10209-005-0005-9. URL: https://doi.org/10.1007/s10209-005-0005-9.

[52] David Goldberg and Cate Richardson. "Touch-Typing with a Stylus." In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*. CHI '93. Amsterdam, The Netherlands: Association for Computing Machinery, 1993, 80–87. ISBN: 0897915755. DOI: 10.1145/169059.169093. URL: https://doi.org/10.1145/169059.169093.

[53] Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. "Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates." In: *Speech Communication* 52.3 (2010), pp. 181–200.

[54] Jason Tyler Griffin, Jerome Pasquero, and Donald Somerset McKenzie. "Touch-screen keyboard providing selection of word predictions in partitions of the touchscreen keyboard." Pat. US Patent 9,116,552. 2015.

[55] Jason Tyler Griffin, Jerome Pasquero, Donald Somerset McKenzie, and Alistair Robert Hamilton. "In-letter word prediction for virtual keyboard." Pat. US Patent 9,122,672. 2015.

[56] Dan Witzner Hansen, Henrik H. T. Skovsgaard, John Paulin Hansen, and Emilie Møllenbach. "Noise Tolerant Selection by Gaze-Controlled Pan and Zoom in 3D." In: *Proceedings of the 2008 Symposium on Eye Tracking Research amp; Applications*. ETRA '08. Savannah, Georgia: Association for Computing Machinery, 2008, 205–212. ISBN: 9781595939821. DOI: 10.1145/1344471.1344521. URL: https://doi.org/10.1145/1344471.1344521.

[57] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research." In: *Human Mental Workload*. Ed. by Peter A. Hancock and Najmedin Meshkati. Vol. 52. Advances in Psychology. North-Holland, 1988, pp. 139–183. DOI: https://doi.org/10.1016/S0166-4115(08)62386-9. URL: https://www.sciencedirect.com/science/article/pii/S0166411508623869.

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

[59] Ramin Hedeshy, Chandan Kumar, Raphael Menges, and Steffen Staab. "Hummer: Text Entry by Gaze and Hum." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2021. ISBN: 9781450380966. URL: https://doi.org/10.1145/3411764.3445501.

[60] Lode Hoste and Beat Signer. "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. Sydney, Australia: Association for Computing Machinery, 2013, 213–220. ISBN: 9781450321297. DOI: 10.1145/2522848.2522861. URL: https://doi.org/10.1145/2522848.2522861.

[61] Anke Huckauf and Mario Urbina. "Gazing with PEYE: New Concepts in Eye Typing." In: APGV '07. Tubingen, Germany: Association for Computing Machinery, 2007, p. 141. ISBN: 9781595936707. DOI: 10.1145/1272582.1272618. URL: https://doi.org/10.1145/1272582.1272618.

[62] Taeho Hwang, Miyoung Kim, Minsu Hwangbo, and Eunmi Oh. "Comparative analysis of cognitive tasks for modeling mental workload with electroencephalogram." In: *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014, pp. 2661–2665. DOI: 10.1109/EMBC.2014.6944170.

[63] Poika Isokoski and Roope Raisamo. "Device Independent Text Input: A Rationale and an Example." In: *Proceedings of the Working Conference on Advanced Visual Interfaces*. AVI '00. Palermo, Italy: Association for Computing Machinery, 2000, 76–83. ISBN: 1581132522. DOI: 10.1145/345513.345262. URL: https://doi.org/10.1145/345513.345262.

[64] Robert JK Jacob. "What you look at is what you get: eye movement-based interaction techniques." In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1990, pp. 11–18.

[65] R Jakob. "The use of eye movements in human-computer interaction techniques: what you look at is what you get." In: *Readings in Intelligent User Interfaces* (1998), pp. 65–83.

[66] Siwacha Janpinijrut, Prakasith Kayasith, Cholwich Nattee, and Manabu Okumura. "Vowel-separated Layout: A Thai Touchscreen Keyboard for People with Hand Movement Disability." In: *Proceedings of the 5th International Conference on Rehabilitation Engineering & Assistive Technology*. i-CREATe '11. Bangkok, Thailand: Singapore Therapeutic, Assistive & Rehabilitative Technologies (START) Centre, 2011, 10:1–10:4. URL: http://dl.acm.org/citation.cfm?id=2500753.2500765.

[67] Anders Sewerin Johansen, John Paulin Hansen, Dan Witzner Hansen, Kenji Itoh, and Satoru Mashino. "Language Technology in a Predictive, Restricted on-Screen Keyboard with Dynamic Layout for Severely Disabled People." In: *Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods*. TextEntry '03. Budapest, Hungary: Association for Computational Linguistics, 2003, 59–66.

[68] Clare-Marie Karat, Christine Halverson, Daniel Horn, and John Karat. "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, 1999, pp. 568–575. DOI: 10.1145/302979.303160.

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

[70] Avi Knoll, Yang Wang, Fang Chen, Jie Xu, Natalie Ruiz, Julien Epps, and Pega Zarjam. "Measuring Cognitive Workload with Low-Cost Electroencephalograph." In: *Human-Computer Interaction – INTERACT 2011*. Ed. by Pedro Campos, Nicholas Graham, Joaquim Jorge, Nuno Nunes, Philippe Palanque, and Marco Winckler. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 568–571. ISBN: 978-3-642-23768-3.

[71] Hiromi Kobayashi and Shiro Kohshima. "Unique morphology of the human eye and its adaptive meaning: comparative studies on external morphology of the primate eye." In: *Journal of Human Evolution* 40.5 (2001), pp. 419–435. ISSN: 0047-2484. DOI: https://doi.org/10.1006/jhev.2001.0468. URL: https://www.sciencedirect.com/science/article/pii/S0047248401904683.

[72] Stefan Kopp, Lars Gesellensetter, Nicole C. Krämer, and Ipke Wachsmuth. "A Conversational Agent as Museum Guide – Design and Evaluation of a Real-World Application." In: *Intelligent Virtual Agents*. Ed. by Themis Panayiotopoulos, Jonathan Gratch, Ruth Aylett, Daniel Ballin, Patrick Olivier, and Thomas Rist. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 329–343.

[73] V. Kostov and S. Fukuda. "Emotion in user interface, voice interaction system." In: *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0*. Vol. 2. 2000, 798–803 vol.2. DOI: 10.1109/ICSMC.2000.885947.

[74] Per Ola Kristensson and Keith Vertanen. "The Potential of Dwell-Free Eye-Typing for Fast Assistive Gaze Communication." In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. ETRA '12. Santa Barbara, California: Association for Computing Machinery, 2012, 241–244. ISBN: 9781450312219. DOI: 10.1145/2168556.2168605. URL: https://doi.org/10.1145/2168556.2168605.

[75] Robert Krueger, Steffen Koch, and Thomas Ertl. "Saccadelenses: interactive exploratory filtering of eye tracking trajectories." In: *2016 IEEE Second Workshop on Eye Tracking and Visualization (ETVIS)*. 2016, pp. 31–34. DOI: 10.1109/ETVIS.2016.7851162.

[76] Chandan Kumar, Raphael Menges, and Steffen Staab. "Eye-Controlled Interfaces for Multimedia Interaction." In: *IEEE MultiMedia* 23.4 (2016), pp. 6–13. DOI: 10.1109/MMUL.2016.52.

[77] Chandan Kumar, Raphael Menges, and Steffen Staab. "Eye-controlled interfaces for multimedia interaction." In: *IEEE MultiMedia*. Vol. 23. 4. New York: IEEE, 2016, pp. 6–13.

[78] Manu Kumar, Andreas Paepcke, and Terry Winograd. "EyePoint: Practical Pointing and Selection Using Gaze and Keyboard." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. San Jose, California, USA: Association for Computing Machinery, 2007, 421–430. ISBN: 9781595935939. DOI: 10.1145/1240624.1240692. URL: https://doi.org/10.1145/1240624.1240692.

[79] Manu Kumar, Andreas Paepcke, Terry Winograd, and Terry Winograd. "EyePoint: Practical pointing and selection using gaze and keyboard." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '07. San Jose, California, USA: ACM, 2007, pp. 421–430. ISBN: 978-1-59593-593-9. DOI: 10.1145/1240624.1240692. URL: http://doi.acm.org/10.1145/1240624.1240692.

[80] Andrew TN Kurauchi. "EyeSwipe: text entry using gaze paths." PhD thesis. Ph. D. dissertation, University of Sao Paulo, 2018.

[81] Andrew Kurauchi, Wenxin Feng, Ajjen Joshi, Carlos Morimoto, and Margrit Betke. "EyeSwipe: Dwell-free Text Entry Using Gaze Paths." In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI '16. San Jose, California, USA: ACM, 2016, pp. 1952–1956. ISBN: 978-1-4503-3362-7. DOI: 10.1145/2858036.2858335. URL: http://doi.acm.org/10.1145/2858036.2858335.

[82] Gordon Kurtenbach and William Buxton. "The Limits of Expert Performance Using Hierarchic Marking Menus." In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*. CHI '93. Amsterdam, The Netherlands: ACM, 1993, pp. 482–487. ISBN: 0-89791-575-5. DOI: 10.1145/169059.169426. URL: http://doi.acm.org/10.1145/169059.169426.

[83] Chris Lankford. "Effective Eye-Gaze Input into Windows." In: ETRA '00. Palm Beach Gardens, Florida, USA: Association for Computing Machinery, 2000, 23–27. ISBN: 1581132808. DOI: 10.1145/355017.355021. URL: https://doi.org/10.1145/355017.355021.

[84] Annabel M. Latham, Keeley A. Crockett, David A. McLean, Bruce Edmonds, and Karen O'Shea. "Oscar: An intelligent conversational agent tutor to estimate learning styles." In: *International Conference on Fuzzy Systems*. 2010, pp. 1–8. DOI: 10.1109/FUZZY.2010.5584064.

[85] SeoYoung Lee and Junho Choi. "Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity." In: *International Journal of Human-Computer Studies* 103 (2017), pp. 95–105. ISSN: 1071-5819. DOI: https://doi.org/10.1016/j.ijhcs.2017.02.005. URL: https://www.sciencedirect.com/science/article/pii/S1071581917300198.

[86] Luis A. Leiva, Alireza Sahami, Alejandro Catala, Niels Henze, and Albrecht Schmidt. "Text Entry on Tiny QWERTY Soft Keyboards." In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. CHI '15. Seoul, Republic of Korea: Association for Computing Machinery, 2015, 669–678. ISBN: 9781450331456. DOI: 10.1145/2702123.2702388. URL: https://doi.org/10.1145/2702123.2702388.

[87] Lee Hoi Leong, Shinsuke Kobayashi, Noboru Koshizuka, and Ken Sakamura. "CASIS: A Context-Aware Speech Interface System." In: IUI '05. San Diego, California, USA: Association for Computing Machinery, 2005, 231–238. ISBN: 1581138946. DOI: 10.1145/1040830.1040880. URL: https://doi.org/10.1145/1040830.1040880.

[88] Esther Levin and Amir M Mané. "Voice user interface design for automated directory assistance." In: *Ninth European Conference on Speech Communication and Technology*. 2005.

[89] J. C. R. Licklider. "Man-Computer Symbiosis." In: *IRE Transactions on Human Factors in Electronics* HFE-1.1 (1960), pp. 4–11. DOI: 10.1109/THFE2.1960.4503259.

[90] Christof Lutteroth, Moiz Penkar, and Gerald Weber. "Gaze vs. Mouse: A Fast and Accurate Gaze-Only Click Alternative." In: *Proceedings of the 28th Annual ACM Symposium on User Interface Software amp; Technology*. UIST '15. Charlotte, NC, USA: Association for Computing Machinery, 2015, 385–394. ISBN: 9781450337793. DOI: 10.1145/2807442.2807461. URL: https://doi.org/10.1145/2807442.2807461.

[91] Otto Hans-Martin Lutz, Antje Christine Venjakob, and Stefan Ruff. "SMOOVS: Towards calibration-free text entry by gaze using smooth pursuit movements." In: *Journal of Eye Movement Research* 8.1 (2015). DOI: 10.16910/jemr.8.1.2. URL: https://bop.unibe.ch/JEMR/article/view/2394.

[92] Kien Hoa Ly, Ann-Marie Ly, and Gerhard Andersson. "A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods." In: *Internet Interventions* 10 (2017), pp. 39–46. ISSN: 2214-7829. DOI: https://doi.org/10.1016/j.invent.2017.10.002. URL: https://www.sciencedirect.com/science/article/pii/S221478291730091X.

[93] I. Scott MacKenzie and R. William Soukoreff. "Phrase Sets for Evaluating Text Entry Techniques." In: *CHI '03 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '03. Ft. Lauderdale, Florida, USA: Association for Computing Machinery, 2003, 754–755. ISBN: 1581136374. DOI: 10.1145/765891.765971. URL: https://doi.org/10.1145/765891.765971.

[94] I Scott MacKenzie and K Tanaka-Ishii. *Evaluation of text entry techniques*. Vol. 2007. Morgan Kaufmann San Francisco, CA, 2007.

[95] I. Scott MacKenzie and Kumiko Tanaka-Ishii. *Text Entry Systems: Mobility, Accessibility, Universality*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2007. ISBN: 0123735912.

[96] I. Scott MacKenzie and Shawn X. Zhang. "The Design and Evaluation of a High-Performance Soft Keyboard." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '99. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 1999, 25–31. ISBN: 0201485591. DOI: 10.1145/302979.302983. URL: https://doi.org/10.1145/302979.302983.

[97] I. Scott MacKenzie and Shawn X. Zhang. "The Design and Evaluation of a High-Performance Soft Keyboard." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '99. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, 1999, 25–31. ISBN: 0201485591. DOI: 10.1145/302979.302983. URL: https://doi.org/10.1145/302979.302983.

[98] I. Scott MacKenzie and Xuang Zhang. "Eye Typing Using Word and Letter Prediction and a Fixation Algorithm." In: *Proceedings of the 2008 Symposium on Eye Tracking Research amp; Applications*. ETRA '08. Savannah, Georgia: Association for Computing Machinery, 2008, 55–58. ISBN: 9781595939821. DOI: 10.1145/1344471.1344484. URL: https://doi.org/10.1145/1344471.1344484.

[99] Päivi Majaranta. *Text entry by eye gaze*. Tampere University Press, 2009.

[100] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. "Fast Gaze Typing with an Adjustable Dwell Time." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '09. Boston, MA, USA: Association for Computing Machinery, 2009, 357–360. ISBN: 9781605582467. DOI: 10.1145/1518701.1518758. URL: https://doi.org/10.1145/1518701.1518758.

[101] Päivi Majaranta, Anne Aula, and Kari-Jouko Räihä. "Effects of Feedback on Eye Typing with a Short Dwell Time." In: *Proceedings of the 2004 Symposium on Eye Tracking Research amp; Applications*. ETRA '04. San Antonio, Texas: Association for Computing Machinery, 2004, 139–146. ISBN: 1581138253. DOI: 10.1145/968363.968390. URL: https://doi.org/10.1145/968363.968390.

[102] Päivi Majaranta, I. Scott MacKenzie, Anne Aula, and Kari-Jouko Räihä. "Effects of feedback and dwell time on eye typing speed and accuracy." In: *Universal Access in the Information Society* 5.2 (2006), pp. 199–208. ISSN: 1615-5297. DOI: 10.1007/s10209-006-0034-z. URL: https://doi.org/10.1007/s10209-006-0034-z.

[103] Päivi Majaranta and Kari-Jouko Räihä. "Twenty Years of Eye Typing: Systems and Design Issues." In: ETRA '02. New Orleans, Louisiana: Association for Computing Machinery, 2002, 15–22. ISBN: 1581134673. DOI: 10.1145/507072.507076. URL: https://doi.org/10.1145/507072.507076.

[104] Päivi Majaranta and Kari-Jouko Räihä. "Twenty Years of Eye Typing: Systems and Design Issues." In: *Proceedings of the 2002 Symposium on Eye Tracking Research amp; Applications*. ETRA '02. New Orleans, Louisiana: Association for Computing Machinery, 2002, 15–22. ISBN: 1581134673. DOI: 10.1145/507072.507076. URL: https://doi.org/10.1145/507072.507076.

[105] Päivi Majaranta and Kari-Jouko Räihä. "Twenty Years of Eye Typing: Systems and Design Issues." In: *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications*. ETRA '02. New Orleans, Louisiana: ACM, 2002, pp. 15–22. ISBN: 1-58113-467-3. DOI: 10.1145/507072.507076. URL: http://doi.acm.org/10.1145/507072.507076.

[106] Bill Manaris and Alan Harkreader. "SUITEKeys: A Speech Understanding Interface for the Motor-Control Challenged." In: *Proceedings of the Third International ACM Conference on Assistive Technologies*. Assets '98. Marina del Rey, California, USA: Association for Computing Machinery, 1998, 108–115. ISBN: 1581130201. DOI: 10.1145/274497.274517. URL: https://doi.org/10.1145/274497.274517.

[107] Bill Manaris and Alan Harkreader. "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, 1998, pp. 108–115. DOI: 10.1145/274497.274517.

[108] Jennifer Mankoff and Gregory D. Abowd. "Cirrin: A Word-Level Unistroke Keyboard for Pen Input." In: *Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology*. UIST '98. San Francisco, California, USA: Association for Computing Machinery, 1998, 213–214. ISBN: 1581130341. DOI: 10.1145/288392.288611. URL: https://doi.org/10.1145/288392.288611.

[109] Chandra Sekhar Mantravadi. "Adaptive multimodal integration of speech and gaze." PhD thesis. Rutgers University, New Brunswick, NJ, 2009.

[110] Raphael Menges. "Improving Usability and Accessibility of the Web with Eye Tracking." doctoralthesis. Universität Koblenz-Landau, Universitätsbibliothek, 2021, pp. xi, 223.

[111] Raphael Menges, Chandan Kumar, Daniel Müller, and Korok Sengupta. "GazeTheWeb: A Gaze-Controlled Web Browser." In: *Proceedings of the 14th Web for All Conference on The Future of Accessible Work*. W4A '17. Perth, Western Australia, Australia: ACM, 2017, 25:1–25:2. ISBN: 978-1-4503-4900-0. DOI: 10.1145/3058555.3058582. URL: http://doi.acm.org/10.1145/3058555.3058582.

[112] Raphael Menges, Chandan Kumar, and Steffen Staab. "Improving User Experience of Eye Tracking-Based Interaction: Introspecting and Adapting Interfaces." In: *ACM Trans. Comput.-Hum. Interact.* 26.6 (Nov. 2019). ISSN: 1073-0516. DOI: 10.1145/3338844. URL: https://doi.org/10.1145/3338844.

[113] Bruno Merlin and Mathieu Raynal. "Evaluation of SpreadKey System with Motor Impaired Users." In: *Computers Helping People with Special Needs*. Ed. by Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 112–119.

[114] Yoshiyuki Mihara, Etsuya Shibayama, and Shin Takahashi. "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, 2005, pp. 76–83. DOI: 10.1145/1090785.1090801.

[115] Adam S. Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. "Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health." In: *JAMA Internal Medicine* 176.5 (May 2016), pp. 619–625. ISSN: 2168-6106. DOI: 10.1001/jamainternmed.2016.0400. eprint: https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2500043/ioi160007.pdf. URL: https://doi.org/10.1001/jamainternmed.2016.0400.

[116] Julio Miró-Borrás and Pablo Bernabeu-Soler. "Text Entry in the E-Commerce Age: Two Proposals for the Severely Handicapped." en. In: *Journal of theoretical and applied electronic commerce research* 4 (Apr. 2009), pp. 101 –112. ISSN: 0718-1876. URL: https://scielo.conicyt.cl/scielo.php?script=sci_arttext&pid=S0718-18762009000100009&nrm=iso.

[117] Louis-Philippe Morency et al. "SimSensei Demonstration: A Perceptive Virtual Human Interviewer for Healthcare Applications." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 29.1 (2015). URL: https://ojs.aaai.org/index.php/AAAI/article/view/9777.

[118] Carlos H. Morimoto and Arnon Amir. "Context Switching for Fast Key Selection in Text Entry Applications." In: *Proceedings of the 2010 Symposium on Eye-Tracking Research amp; Applications*. ETRA '10. Austin, Texas: Association for Computing Machinery, 2010, 271–274. ISBN: 9781605589947. DOI: 10.1145/1743666.1743730. URL: https://doi.org/10.1145/1743666.1743730.

[119] Robert R Morris, Kareem Kouddous, Rohan Kshirsagar, and Stephen M Schueller. "Towards an Artificially Empathic Conversational Agent for Mental Health Applications: System Design and User Perceptions." In: *J Med Internet Res* 20.6 (2018), e10148. ISSN: 1438-8871. DOI: 10.2196/10148. URL: http://www.jmir.org/2018/6/e10148/.

[120] Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. "Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times." In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, 2558–2570. ISBN: 9781450346559. DOI: 10.1145/3025453.3025517. URL: https://doi.org/10.1145/3025453.3025517.

[121] Emilie Møllenbach, John Paulin Hansen, and Martin Lillholm. "Eye Movements in Gaze Interaction." In: 6 (2013). DOI: 10.16910/jemr.6.2.1. URL: https://bop.unibe.ch/JEMR/article/view/2354.

[122] Jakob Nielsen. *Usability engineering*. Morgan Kaufmann, 1994.

[123] Jan Nouza, Tomáš Nouza, and P Cerva. "A multi-functional voice-control aid for disabled persons." In: *Proc. of International Conference on Speech and Computer (SPECOM'05). Patras, Greece*. 2005, pp. 715–718.

[124] Jan Nouza, Jindrich Zdansky, Petr Cerva, and Jan Silovsky. "Challenges in Speech Processing of Slavic Languages (Case Studies in Speech Recognition of Czech and Slovak)." In: *Development of Multimodal Interfaces: Active Listening and Synchrony: Second COST 2102 International Training School, Dublin, Ireland, March 23-27, 2009, Revised Selected Papers*. Ed. by Anna Esposito, Nick Campbell, Carl Vogel, Amir Hussain, and Anton Nijholt. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 225–241. DOI: 10.1007/978-3-642-12397-9_19. URL: https://doi.org/10.1007/978-3-642-12397-9_19.

[125] Stavroula Ntoa, George Margetis, Margherita Antona, and Constantine Stephanidis. "Scanning-based interaction techniques for motor impaired users." In: *Assistive Technologies and Computer Access for Motor Disabilities*. IGI Global, 2014, pp. 57–89.

[126] Marcus Nyström and Kenneth Holmqvist. "An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data." In: *Behavior research methods* 42.1 (2010), pp. 188–204.

[127] Sharon Oviatt. "Multimodal interactive maps: Designing for human performance." In: *Human-Computer Interaction* 12.1 (1997), pp. 93–129.

[128] Sharon Oviatt. "Taming recognition errors with a multimodal interface." In: *Communications of the ACM* 43.9 (2000), pp. 45–45.

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

[130] Sharon Oviatt, Phil Cohen, Lizhong Wu, Lisbeth Duncan, Bernhard Suhm, Josh Bers, Thomas Holzman, Terry Winograd, James Landay, Jim Larson, et al. "Designing the user interface for multimodal speech and pen-based gesture applications: state-of-the-art systems and future research directions." In: *Human-Computer Interaction* 15.4 (2000), pp. 263–322. DOI: 10.1207/S15327051HCI1504_1.

[131] Fred Paas, Alexander Renkl, and John Sweller. "Cognitive Load Theory and Instructional Design: Recent Developments." In: *Educational Psychologist* 38.1 (2003), pp. 1–4. DOI: 10.1207/S15326985EP3801_1. eprint: https://doi.org/10.1207/S15326985EP3801_1. URL: https://doi.org/10.1207/S15326985EP3801_1.

[132] Prateek Panwar, Sayan Sarcar, and Debasis Samanta. "EyeBoard: A fast and accurate eye gaze-based text entry system." In: *2012 4th International Conference on Intelligent Human Computer Interaction (IHCI)*. IEEE. 2012, pp. 1–8.

[133] Diogo Pedrosa, Maria Da Graça Pimentel, Amy Wright, and Khai N. Truong. "Filteryedping: Design Challenges and User Performance of Dwell-Free Eye Typing." In: *ACM Trans. Access. Comput.* 6.1 (2015). ISSN: 1936-7228. DOI: 10.1145/2724728. URL: https://doi.org/10.1145/2724728.

[134] Ken Pfeuffer and Hans Gellersen. "Gaze and Touch Interaction on Tablets." In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. UIST '16. Tokyo, Japan: Association for Computing Machinery, 2016, 301–311. ISBN: 9781450341899. DOI: 10.1145/2984511.2984514. URL: https://doi.org/10.1145/2984511.2984514.

[135] Ian J. Pitt and Alistair D. N. Edwards. "Improving the Usability of Speech-Based Interfaces for Blind Users." In: Assets '96. Vancouver, British Columbia, Canada: Association for Computing Machinery, 1996, 124–130. ISBN: 0897917766. DOI: 10.1145/228347.228367. URL: https://doi.org/10.1145/228347.228367.

[136] Michael Plöchl, José Ossandón, and Peter König. "Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data." In: *Frontiers in Human Neuroscience* 6 (2012), p. 278. ISSN: 1662-5161. DOI: 10.3389/fnhum.2012.00278. URL: https://www.frontiersin.org/article/10.3389/fnhum.2012.00278.

[137] Ondrej Polacek, Zdenek Mikovec, Adam J. Sporka, and Pavel Slavik. "Humsher: A Predictive Keyboard Operated by Humming." In: *The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '11. Dundee, Scotland, UK: Association for Computing Machinery, 2011, 75–82. ISBN: 9781450309202. DOI: 10.1145/2049536.2049552. URL: https://doi.org/10.1145/2049536.2049552.

[138] Ondřej Poláček, Zdeněk M'ikovec, and Pavel Slav'ik. "Predictive scanning keyboard operated by hissing." In: *Proceedings of the 2nd IASTED International Conference Assistive Technologies*. Citeseer. 2012, pp. 862–9.

[139] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. "A review of affective computing: From unimodal analysis to multimodal fusion." In: *Information Fusion* 37 (2017), pp. 98 –125. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2017.02.003. URL: http://www.sciencedirect.com/science/article/pii/S1566253517300738.

[140] Marco Porta and Alessia Ravelli. "WeyeB, an Eye-Controlled Web Browser for Hands-Free Navigation." In: HSI'09. Catania, Italy: IEEE Press, 2009, 207–212. ISBN: 9781424439591.

[141] Matheus Vieira Portela and David Rozado. "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, 2014, pp. 426–429. DOI: 10.1145/2686612.2686679.

[142] Kari-Jouko Räihä and Saila Ovaska. "An Exploratory Study of Eye Typing Fundamentals: Dwell Time, Text Entry Rate, Errors, and Workload." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, 3001–3010. ISBN: 9781450310154. DOI: 10.1145/2207676.2208711. URL: https://doi.org/10.1145/2207676.2208711.

[143] Kari-Jouko Räihä and Saila Ovaska. "An Exploratory Study of Eye Typing Fundamentals: Dwell Time, Text Entry Rate, Errors, and Workload." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, 3001–3010. ISBN: 9781450310154. DOI: 10.1145/2207676.2208711. URL: https://doi.org/10.1145/2207676.2208711.

[144] Kari-Jouko Räihä and Saila Ovaska. "An Exploratory Study of Eye Typing Fundamentals: Dwell Time, Text Entry Rate, Errors, and Workload." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: Association for Computing Machinery, 2012, 3001–3010. ISBN: 9781450310154. DOI: 10.1145/2207676.2208711. URL: https://doi.org/10.1145/2207676.2208711.

[145] D.A. Robinson. "The oculomotor control system: A review." In: *Proceedings of the IEEE* 56.6 (1968), pp. 1032–1049. DOI: 10.1109/PROC.1968.6455.

[146] S Robinson, DR Traum, M Ittycheriah, and J Henderer. "What would you Ask a conversational Agent?" In: *Observations of Human-Agent Dialogues in a Museum Setting* (2008).

[147] David B Roe, Jay G Wilpon, et al., eds. *Voice communication between humans and machines*. Washington, DC: National Academies Press, 1994.

[148] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. "Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones." In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1.4 (2018). DOI: 10.1145/3161187. URL: https://doi.org/10.1145/3161187.

[149] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. "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*. New York: ACM, 2018, 159:1–159:23. DOI: 10.1145/3161187.

[150] Timothy A Salthouse. "Effects of age and skill in typing." In: *Journal of Experimental Psychology: General* 113.3 (1984), p. 345.

[151] Frode Eika Sandnes. "Reflective Text Entry: A Simple Low Effort Predictive Input Method Based on Flexible Abbreviations." In: *Procedia Computer Science* 67 (2015). Proceedings of the 6th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-exclusion, pp. 105 –112. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2015.09.254. URL: http://www.sciencedirect.com/science/article/pii/S1877050915031002.

[152] Simon Schenk, Marc Dreiser, Gerhard Rigoll, and Michael Dorr. "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. Denver, Colorado, USA: Association for Computing Machinery, 2017, 3034–3044. ISBN: 9781450346559. DOI: 10.1145/3025453.3025455. URL: https://doi.org/10.1145/3025453.3025455.

[153] Simon Schenk, Marc Dreiser, Gerhard Rigoll, and Michael Dorr. "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. Denver, Colorado, USA: ACM, 2017, pp. 3034–3044. ISBN: 978-1-4503-4655-9. DOI: 10.1145/3025453.3025455. URL: http://doi.acm.org/10.1145/3025453.3025455.

[154] Daniel Schulman and Timothy Bickmore. "Persuading Users through Counseling Dialogue with a Conversational Agent." In: *Proceedings of the 4th International Conference on Persuasive Technology*. Persuasive '09. Claremont, California, USA: Association for Computing Machinery, 2009. ISBN: 9781605583761. DOI: 10.1145/1541948.1541983. URL: https://doi.org/10.1145/1541948.1541983.

[155] Daniel Schulman, Timothy Bickmore, and Candace Sidner. "An intelligent conversational agent for promoting long-term health behavior change using motivational interviewing." In: *2011 AAAI Spring Symposium Series*. 2011.

[156] I. Scott MacKenzie and Behrooz Ashtiani. "BlinkWrite: efficient text entry using eye blinks." In: *Universal Access in the Information Society* 10.1 (2011), pp. 69–80. ISSN: 1615-5297. DOI: 10.1007/s10209-010-0188-6. URL: https://doi.org/10.1007/s10209-010-0188-6.

[157] Andrew Sears, Jinhuan Feng, Kwesi Oseitutu, and Claire-Marie Karat. "Hands-free, speech-based navigation during dictation: difficulties, consequences, and solutions." In: *Human-Computer Interaction* 18.3 (2003), pp. 229–257. DOI: 10.1207/S15327051HCI1803_2.

[158] Mary C Seiler and Fritz A Seiler. "Numerical recipes in C: the art of scientific computing." In: *Risk Analysis* 9.3 (1989), pp. 415–416.

[159] Korok Sengupta, Sabin Bhattarai, Sayan Sarcar, I. Scott MacKenzie, and Steffen Staab. "Leveraging Error Correction in Voice-Based Text Entry by Talk-and-Gaze." In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: Association for Computing Machinery, 2020, 1–11. ISBN: 9781450367080. DOI: 10.1145/3313831.3376579. URL: https://doi.org/10.1145/3313831.3376579.

[160] Korok Sengupta, Min Ke, Raphael Menges, Chandan Kumar, and Steffen Staab. "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, 2018, p. 88. DOI: 10.1145/3204493.3208338.

[161] Korok Sengupta, Raphael Menges, Chandan Kumar, and Steffen Staab. "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. Limassol, Cyprus: Association for Computing Machinery, 2017, 121–124. ISBN: 9781450348935. DOI: 10.1145/3030024.3038259. URL: https://doi.org/10.1145/3030024.3038259.

[162] Korok Sengupta, Raphael Menges, Chandan Kumar, and Steffen Staab. "Impact of Variable Positioning of Text Prediction in Gaze-Based Text Entry." In: *Proceedings of the 11th ACM Symposium on Eye Tracking Research amp; Applications*. ETRA '19. Denver, Colorado: Association for Computing Machinery, 2019. ISBN: 97814503670 97. DOI: 10.1145/3317956.3318152. URL: https://doi.org/10.1145/3317956.3318152.

[163] Korok Sengupta, Jun Sun, Raphael Menges, Chandan Kumar, and Steffen Staab. "Analyzing the Impact of Cognitive Load in Evaluating Gaze-Based Typing." In: *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. 2017, pp. 787–792. DOI: 10.1109/CBMS.2017.134.

[164] M. Kumar Sharma, Somnath Dey, P. Kumar Saha, and Debasis Samanta. "Parameters effecting the predictive virtual keyboard." In: *2010 IEEE Students Technology Symposium (TechSym)*. 2010, pp. 268–275. DOI: 10.1109/TECHSYM.2010.5469160.

[165] Rajeev Sharma, Vladimir I Pavlović, and Thomas S Huang. "Toward multimodal human–computer interface." In: *Advances In Image Processing And Understanding: A Festschrift for Thomas S Huang*. World Scientific, 2002, pp. 349–365.

[166] Linda E. Sibert and Robert J. K. Jacob. "Evaluation of Eye Gaze Interaction." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '00. The Hague, The Netherlands: ACM, 2000, pp. 281–288. ISBN: 1-58113-216-6. DOI: 10.1145/332040.332445. URL: http://doi.acm.org/10.1145/332040.332445.

[167] Shyamli Sindhwani, Christof Lutteroth, and Gerald Weber. "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, 2019, p. 203. DOI: 10.1145/3290605.3300433.

[168] Man-hung Siu, Herbert Gish, and Fred Richardson. "Improved estimation, evaluation and applications of confidence measures for speech recognition." In: *Fifth European Conference on Speech Communication and Technology*. 1997.

[169] Young Chol Song. "Joystick Text Entry with Word Prediction for People with Motor Impairments." In: *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS '10. Orlando, Florida, USA: Association for Computing Machinery, 2010, 321–322. ISBN: 9781605588810. DOI: 10.1145/1878803.1878892. URL: https://doi.org/10.1145/1878803.1878892.

[170] R. William Soukoreff and I. Scott MacKenzie. "Measuring Errors in Text Entry Tasks: An Application of the Levenshtein String Distance Statistic." In: *CHI '01 Extended Abstracts on Human Factors in Computing Systems*. CHI EA '01. Seattle, Washington: Association for Computing Machinery, 2001, 319–320. ISBN: 1581133405. DOI: 10.1145/634067.634256. URL: https://doi.org/10.1145/634067.634256.

[171] R. William Soukoreff and I. Scott MacKenzie. "Metrics for Text Entry Research: An Evaluation of MSD and KSPC, and a New Unified Error Metric." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '03. Ft. Lauderdale, Florida, USA: Association for Computing Machinery, 2003, 113–120. ISBN: 1581136307. DOI: 10.1145/642611.642632. URL: https://doi.org/10.1145/642611.642632.

[172] Renato de Sousa Gomide, Luiz Fernando Batista Loja, Rodrigo Pinto Lemos, Edna Lúcia Flôres, Francisco Ramos Melo, and Ricardo Antonio Gonçalves Teixeira. "A new concept of assistive virtual keyboards based on a systematic review of text entry optimization techniques." en. In: *Research on Biomedical Engineering* 32 (June 2016), pp. 176 –198. ISSN: 2446-4740. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2446-47402016000200176&nrm=iso.

[173] Adam J Sporka. "Non-speech sounds for user interface control." In: *Czech Technical University in Prague* 116 (2008).

[174] Adam J. Sporka, Sri H. Kurniawan, Murni Mahmud, and Pavel Slavík. "Non-Speech Input and Speech Recognition for Real-Time Control of Computer Games." In: *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*. Assets '06. Portland, Oregon, USA: Association for Computing Machinery, 2006, 213–220. ISBN: 1595932909. DOI: 10.1145/1168987.1169023. URL: https://doi.org/10.1145/1168987.1169023.

[175] Lisa J. Stifelman, Barry Arons, Chris Schmandt, and Eric A. Hulteen. "VoiceNotes: A Speech Interface for a Hand-Held Voice Notetaker." In: *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*. CHI '93. Amsterdam, The Netherlands: Association for Computing Machinery, 1993, 179–186. ISBN: 0897915755. DOI: 10.1145/169059.169150. URL: https://doi.org/10.1145/169059.169150.

[176] Bernhard Suhm, Brad Myers, and Alex Waibel. "Multimodal Error Correction for Speech User Interfaces." In: *ACM Trans. Comput.-Hum. Interact.* 8.1 (2001), 60–98. ISSN: 1073-0516. DOI: 10.1145/371127.371166. URL: https://doi.org/10.1145/371127.371166.

[177] M S Treder, N M Schmidt, and B Blankertz. "Gaze-independent brain–computer interfaces based on covert attention and feature attention." In: *Journal of Neural Engineering* 8.6 (2011), p. 066003. DOI: 10.1088/1741-2560/8/6/066003. URL: https://doi.org/10.1088/1741-2560/8/6/066003.

[178] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L. Sidner, and Timothy Bickmore. "Designing Relational Agents as Long Term Social Companions for Older Adults." In: *Intelligent Virtual Agents*. Ed. by Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 289–302.

[179] Matteo Vescovi. "Soothsayer: a multi-source system for text prediction." In: ().

[180] John Vines, Gary Pritchard, Peter Wright, Patrick Olivier, and Katie Brittain. "An Age-Old Problem: Examining the Discourses of Ageing in HCI and Strategies for Future Research." In: *ACM Trans. Comput.-Hum. Interact.* 22.1 (2015). ISSN: 1073-0516. DOI: 10.1145/2696867. URL: https://doi.org/10.1145/2696867.

[181] David J. Ward, Alan F. Blackwell, and David J. C. MacKay. "Dasher—a Data Entry Interface Using Continuous Gestures and Language Models." In: *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology*. UIST '00. San Diego, California, USA: Association for Computing Machinery, 2000, 129–137. ISBN: 1581132123. DOI: 10.1145/354401.354427. URL: https://doi.org/10.1145/354401.354427.

[182] David J. Ward, Alan F. Blackwell, and David J. C. MacKay. "Dasher: A Gesture-Driven Data Entry Interface for Mobile Computing." In: *Human–Computer Interaction* 17.2-3 (2002), pp. 199–228. DOI: 10.1080/07370024.2002.9667314. eprint: https://www.tandfonline.com/doi/pdf/10.1080/07370024.2002.9667314. URL: https://www.tandfonline.com/doi/abs/10.1080/07370024.2002.9667314.

[183] David J Ward and David JC MacKay. "Fast hands-free writing by gaze direction." In: *Nature* 418.6900 (2002), pp. 838–838.

[184] Leonard J West. "Vision and kinesthesis in the acquisition of typewriting skill." In: *Journal of Applied Psychology* 51.2 (1967), p. 161.

[185] Leonard J West and Yitzchak Sabban. "Hierarchy of stroking habits at the typewriter." In: *Journal of Applied Psychology* 67.3 (1982), p. 370.

[186] Andrew D. Wilson and Maneesh Agrawala. "Text Entry Using a Dual Joystick Game Controller." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, 475–478. ISBN: 1595933727. DOI: 10.1145/1124772.1124844. URL: https://doi.org/10.1145/1124772.1124844.

[187] Jacob O Wobbrock. "Measures of text entry performance." In: *Text entry systems: Mobility, accessibility, universality* (2007), pp. 47–74.

[188] Jacob O. Wobbrock, Brad A. Myers, and Duen Horng Chau. "In-Stroke Word Completion." In: *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*. UIST '06. Montreux, Switzerland: Association for Computing Machinery, 2006, 333–336. ISBN: 1595933131. DOI: 10.1145/1166253.1166305. URL: https://doi.org/10.1145/1166253.1166305.

[189] Jacob O. Wobbrock, James Rubinstein, Michael W. Sawyer, and Andrew T. Duchowski. "Longitudinal Evaluation of Discrete Consecutive Gaze Gestures for Text Entry." In: *Proceedings of the 2008 Symposium on Eye Tracking Research amp; Applications*. ETRA '08. Savannah, Georgia: Association for Computing Machinery, 2008, 11–18. ISBN: 9781595939821. DOI: 10.1145/1344471.1344475. URL: https://doi.org/10.1145/1344471.1344475.

[190] Jacob Wobbrock and Brad Myers. "Trackball Text Entry for People with Motor Impairments." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, 479–488. ISBN: 1595933727. DOI: 10.1145/1124772.1124845. URL: https://doi.org/10.1145/1124772.1124845.

[191] Jacob Wobbrock and Brad Myers. "Trackball Text Entry for People with Motor Impairments." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '06. Montréal, Québec, Canada: Association for Computing Machinery, 2006, 479–488. ISBN: 1595933727. DOI: 10.1145/1124772.1124845. URL: https://doi.org/10.1145/1124772.1124845.

[192] Hisao Yamada. *A historical study of typewriters and typing methods, from the position of planning Japanese parallels*. Journal of Information Processing, 1980.

[193] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. "Designing SpeechActs: Issues in Speech User Interfaces." In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '95. Denver, Colorado, USA: ACM Press/Addison-Wesley Publishing Co., 1995, 369–376. ISBN: 0201847051. DOI: 10.1145/223904.223952. URL: https://doi.org/10.1145/223904.223952.

[194] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. "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, 1995, pp. 369–376.

[195] Xin Yi, Chun Yu, Weijie Xu, Xiaojun Bi, and Yuanchun Shi. "COMPASS: Rotational Keyboard on Non-Touch Smartwatches." In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, 705–715. ISBN: 9781450346559. DOI: 10.1145/3025453.3025454. URL: https://doi.org/10.1145/3025453.3025454.

[196] Yu Zhong, T. V. Raman, Casey Burkhardt, Fadi Biadsy, and Jeffrey P. Bigham. "¡i¿JustSpeak¡/i¿: Enabling Universal Voice Control on Android." In: *Proceedings of the 11th Web for All Conference*. W4A '14. Seoul, Korea: Association for Computing Machinery, 2014. ISBN: 9781450326513. DOI: 10.1145/2596695.2596720. URL: https://doi.org/10.1145/2596695.2596720.

[197] Fritz Zwicky. "Discovery, invention, research through the morphological approach." In: (1969).

[198] Oleg Špakov and Darius Miniotas. "On-Line Adjustment of Dwell Time for Target Selection by Gaze." In: *Proceedings of the Third Nordic Conference on Human-Computer Interaction*. NordiCHI '04. Tampere, Finland: Association for Computing Machinery, 2004, 203–206. ISBN: 1581138571. DOI: 10.1145/1028014.1028045. URL: https://doi.org/10.1145/1028014.1028045.

[199] Oleg Špakov and Darius Miniotas. "Gaze-Based Selection of Standard-Size Menu Items." In: *Proceedings of the 7th International Conference on Multimodal Interfaces*. ICMI '05. Torento, Italy: Association for Computing Machinery, 2005, 124–128. ISBN: 1595930280. DOI: 10.1145/1088463.1088486. URL: https://doi.org/10.1145/1088463.1088486.

[200] Oleg Špakov and Darius Miniotas. "Gaze-based selection of standard-size menu items." In: *Proceedings of the 7th International Conference on Multimodal Interfaces*. ICMI '05. Torento, Italy: ACM, 2005, pp. 124–128. ISBN: 1-59593-028-0. DOI: 10.1145/1088463.1088486. URL: http://doi.acm.org/10.1145/1088463.1088486.

DECLARATION

Erklärung über die Eigenständigkeit der Dissertation

Ich versichere, dass ich die vorliegende Arbeit mit dem Titel „Improving usability of gaze and voice based text entry systems“ selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe; aus fremden Quellen entnommene Passagen und Gedanken sind als solche kenntlich gemacht.

Declaration of authorship

I hereby certify that the dissertation entitled “Improving usability of gaze and voice based text entry systems” is entirely my own work except where otherwise indicated. Passages and ideas from other sources have been clearly indicated.

Stuttgart, March 15, 2022

Korok Sengupta