Lip Reading as an Active Mode of Interaction with Computer Systems

by

Laxmi Pandey

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Merced

Committee in charge:

Assistant Professor Ahmed Sabbir Arif, Chair
Assistant Professor Rachel Ryskin
Professor Shawn Newsam

Summer 2022

Lip Reading as an Active Mode of Interaction with Computer Systems

Inclusive Interaction Lab
https://www.theiilab.com

Publications

I published the following peer-reviewed articles during my doctoral studies. This dissertation includes my work on image-based lip reading, which also resulted in publications 1, 2, 4, 5.

1. **Laxmi Pandey** and Ahmed Sabbir Arif. 2022. Design and Evaluation of a Silent Speech-Based Selection Method for Eye-Gaze Pointing. In Proceedings of the 2022 ACM Interactive Surfaces and Spaces Conference (ISS '22). ACM, New York, NY, USA, to appear.
2. **Laxmi Pandey** and Ahmed Sabbir Arif. 2022. Effects of Speaking Rate on Speech and Silent Speech Recognition. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 231, 1–8. https://doi.org/10.1145/3491101.3519611.
3. **Laxmi Pandey** and Ahmed Sabbir Arif. 2021. Silent Speech and Emotion Recognition from Vocal Tract Shape Dynamics in Real-Time MRI. In ACM SIGGRAPH 2021 Posters (SIGGRAPH '21). Association for Computing Machinery, New York, NY, USA, Article 27, 1–2. https://doi.org/10.1145/3450618.3469176.
4. **Laxmi Pandey** and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 1, 1–19. https://doi.org/10.1145/3411764.3445565.
5. **Laxmi Pandey**, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of Speech and Silent Speech Input Methods in Private and Public. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 251, 1–13. https://doi.org/10.1145/3411764.3445430.
6. **Laxmi Pandey** and Ahmed Sabbir Arif. 2020. Enabling Text Translation Using the Suggestion Bar of a Virtual Keyboard. IEEE International Conference on Systems, Man, and Cybernetics (SMC '20), IEEE, Washington, DC, USA, 4352-4357. https://doi.org/10.1109/42975.2020.9282879.
7. **Laxmi Pandey**, Azar Alizadeh, and Ahmed Sabbir Arif. 2020. Enabling Predictive Number Entry and Editing on Touchscreen-Based Mobile Devices. Proceedings of the 2020 Conference on Human Information Interaction and Retrieval (CHIIR '20). Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/3343413.3377957.
8. Steven J. Castellucci, I. Scott MacKenzie, Mudit Misra, **Laxmi Pandey**, and Ahmed Sabbir Arif. 2019. TiltWriter: design and evaluation of a no-touch tilt-based text entry method for handheld devices. In Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia (MUM '19). Association for Computing Machinery, New York, NY, USA, Article 7, 1–8. https://doi.org/10.1145/3365610.3365629.
9. **Laxmi Pandey** and Ahmed Sabbir Arif. 2019. Context-sensitive app prediction on the suggestion bar of a mobile keyboard. In Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia (MUM '19). Association for Computing Machinery, New York, NY, USA, Article 45, 1–5. https://doi.org/10.1145/3365610.3368414.

Abstract

Lip Reading as an Active Mode of Interaction with Computer Systems

by

Laxmi Pandey

Doctor of Philosophy in Electrical Engineering and Computer Science

University of California, Merced

Assistant Professor Ahmed Sabbir Arif, Chair

Interacting with computer systems with speech is more natural than conventional interaction methods. It is also more accessible since it does not require precise selection of small targets or rely entirely on visual elements like virtual keys and buttons. Speech also enables contactless interaction, which is of particular interest when touching public devices is to be avoided, such as the recent COVID-19 pandemic situation. However, speech is unreliable in noisy places and can compromise users' privacy and security when in public. Image-based silent speech, which primarily converts tongue and lip movements into text, can mitigate many of these challenges. Since it does not rely on acoustic features, users can silently speak without vocalizing the words. It has also been demonstrated as a promising input method on mobile devices and has been explored for a variety of audiences and contexts where the acoustic signal is unavailable (e.g., people with speech disorders) or unreliable (e.g., noisy environment). Though the method shows promise, very little is known about peoples' perceptions regarding using it, their anticipated performance of silent speech input, and their approach to avoiding potential misrecognition errors. Besides, existing silent speech recognition models are slow and error prone, or use stationary, external devices that are not scalable. In this dissertation, we attempt to address these issues. Towards this, we first conduct a user study to explore users' attitudes towards silent speech with a particular focus on social acceptance. Results show that people perceive silent speech as more socially acceptable than speech input but are concerned about input recognition, privacy, and security issues. We then conduct a second study examining users' error tolerance with speech and silent speech input methods. Results reveal that users are willing to tolerate more errors with silent speech input than speech input as it offers a higher degree of privacy and security. We conduct another study to identify a suitable method for providing real-time feedback on silent speech input. Results show that users find an abstract feedback method effective and significantly more private and secure than a commonly used video feedback method. In light of these findings, which establish silent speech as an acceptable and desirable mode of interaction, we take a step forward to address the technological limitations of existing image-based silent speech recognition models to make them more usable and reliable on computer systems. Towards this, first, we develop LipType, an optimized version of LipNet for improved speed and accuracy. We then develop an independent repair model that processes video input for poor lighting conditions, when applicable, and corrects potential errors in output for increased accuracy. We then test this model with LipType and other speech and silent speech recognizers to demonstrate its effectiveness. In an evaluation, the model reduced word error rate by 57% compared to the state-of-the-art without compromising the overall computation time. However, we identify that the model is still susceptible to failure due to the variability of user characteristics. A person's speaking rate, for instance, is a fundamental user characteristic that can influence speech recognition performance due to the variation in acoustic properties of human speech production. We formally investigate the effects of speaking rate on silent speech recognition. Results revealed that native users speak about 8% faster than non-native users, but both groups slow down at comparable rates (34–40%) when interacting with silent speech, mostly to increase its accuracy rates. A follow-up experiment confirms that slowing down does improve the accuracy of silent speech recognition. The method yields the best accuracy rate when speaking at 0.75x of the actual speaking rate. These findings highlight the importance of considering speaking rate in silent speech-based interfaces. Finally, we evaluate the effectiveness of the modality in an actual computer system. Particularly, we study the feasibility of using silent speech as a hands-free selection method in eye-gaze pointing on computer systems. Results revealed that silent speech is significantly better than other hands-free selection methods, namely dwell and speech, in terms of performance, usability, and perceived workload.

To my family, friends, and mentors for their love and support.

A tribute to my late grandfather.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to extend my sincerest gratitude to my doctoral advisor Dr. Ahmed Sabbir Arif for introducing me to the field of Human-Computer Interaction. I feel fortunate to work with such an enthusiastic and supportive advisor and would like to thank him for making himself available almost all the time to rectify all the blockers. I consider myself being blessed for getting an opportunity to approach the problem with open thinking and to get enough time to experiment with my own ideas. I am indebted to him for his kindness, enthusiasm, faith, optimism, and motivation at every step of my dissertation. His advice and guidance have been and will always be immensely valuable to me.

I would also like to thank the members of my dissertation committee, Drs. Rachel Ryskin and Shawn Newsam, for taking time out of their busy schedules to serve on my dissertation committee.

Special thanks go to UC Merced for financial support through various funds and scholarships. I wish to thank the ACM SIG on Information Retrieval and ACM-W for supporting me with a Conference Travel Grant. I would also like to thank the Daryl and Janet Hatano Foundation and Fred and Mitzie Ruiz Foundation for supporting me with a Hatano Cognitive Development Research Fellowship 2020 and Fred and Mitzie Ruiz Fellowship 2021, respectively. I also wish to thank the Department of Computer Science Scholarship Committee for considering me worthy of nomination for the above awards. This dissertation is essentially a consolidation of our recent successive papers on silent speech in CHI 2021-2022 and ISS 2022. Therefore, I am thankful for the efforts of anonymous reviewers and the corresponding organizing committees.

I would also like to thank my friends for having numerous fruitful discussions to give their opinions which helped me in various ways. In addition, I am grateful to my friends in the Inclusive Interaction Lab for their constant support.

Last but not least, I would like to express my profound gratitude to my parents, sisters, and my husband who have been the real power-house of love and encouragement. They have always stood by me through good and tough times. Special thanks to my beloved dog, Cookie, for his unconditional love and excellent companionship.

Chapter 1

Introduction

Speech input, an audio-based language processing method that converts acoustic features into text, is one of the most natural and efficient ways to interact with computer systems. It is also more accessible as it does not require precise selection of small targets or rely entirely on visual elements like virtual keys and buttons [248]. Speech also enables contactless interaction, which is of particular interest when touching public devices is to be avoided, such as the recent COVID-19 pandemic situation. In addition, it can potentially improve user comfort and productivity when traditional input methods, like touch and keyboards, are inefficient, difficult, or inconvenient to use. It allows, for example, people with limited motor skills to interact with mobile devices without using their hands. It is also beneficial to people with Situationally-Induced Impairments and Disabilities (SIID), where the hands are incapacitated due to reasons such as performing a secondary task, wearing gloves, or minor injuries. It also facilitates eyes-free interaction on mobile devices, especially for visually impaired users.

Despite its advantages and proven effectiveness, there are many scenarios where speech is not a viable mode of communication. First, the surroundings may not be favorable for speech-based communication: a person could be near a busy market or in a crowded restaurant where the surrounding noise makes speech difficult to recognize. Second, a person may not wish to speak out loud because of privacy and security concerns or could be in a public setting where others do not want to be disturbed, such as in a library or museum. Finally, and most importantly, many people have difficulties in speaking or are unable to speak entirely due to a range of speech and neurological disorders. Although many augmentative and alternative communication (AAC) devices are available to help them vocalize, these devices produce unnatural sounding vocalization. This prevents users from communicating effectively with other humans and technologies like voice-controlled virtual assistants. This demand developing better communication methods, particularly for mobile devices, that can improve these populations' access to fellow humans and recent technological advancements.

A system that can understand speech by visually interpreting the movements of the speaker's lips, known as lip reading or silent speech recognition (Fig. 1.1), can mitigate many of the these challenges. We envision several benefits of using silent speech over speech.

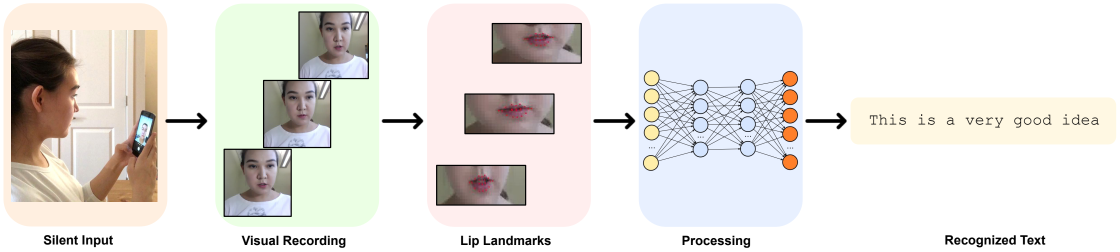


Figure 1.1: An overview of silent speech recognition system: Automatic segmentation of lip sequences and its classification into text with an end-to-end deep neural network.

First, silent speech does not rely on acoustic features, thus can be used in noisy places. Second, silent speech is an unspoken form of communication, hence can largely alleviate the issues surrounding personal privacy and social norms in public environment. Third, silent speech is more inclusive since it can be used by people who cannot vocalize. Researchers explored different video-based [5, 25, 58, 56] and advanced sensor-based [213, 223, 238, 281] recognition methods where they showed relatively high accuracy in speech recognition with silent speech input. A recent work [276] explored silent speech input on mobile devices, where users expressed a higher level of satisfaction with this input method over the tradition speech input. In spite of its advantages, very little is known about peoples' perceptions regarding using it, especially in terms of social acceptance that could influence users' willingness to use this input method. Prior research showed that social acceptability has a significant implication for technological acceptance as it is directly connected to peoples' preferences on using new technologies [282, 160]. In addition, several key questions remain unknown that could influence users' attitude towards using the method. For instance, researchers showed that silent speech input could be prone to high error rates [174, 232, 211, 74]. Consequently, silent speech recognition accuracy could be a key factor in adopting the method. However, little is known of users' error tolerance level for silent speech input. Additionally, silent speech input recognition on mobile devices depends primarily on capturing users' tongue and lip movements via the front camera. Thus, providing appropriate real-time feedback on input recognition is critical for the acceptance of the method. Therefore, to inform the design of silent speech-based interfaces, it is necessary to investigate the error tolerance and suitable feedback mechanism for silent speech input.

Several algorithms and modeling techniques have been proposed for silent speech recognition. Researchers explored different sensor-based [213, 223, 238, 281] recognition methods, many of which use expensive, invasive, and non-portable hardware, including electromagnetic articulography (EMA) [87, 99, 120], real-time magnetic resonance imaging (rtMRI) [219], electroencephalogram (EEG) [231], electromyography (EMG) [150, 291, 146, 145, 147, 189, 254], ultrasound imaging [156, 90, 73, 72, 129, 130, 99, 120], vibrational sensors of glottal activity [213, 223, 238, 281], speech motor cortex implants [32], and non-audible murmur (NAM) microphone [122, 208, 121]. These methods use invasive, impractical, non-portable setups, impeding their scalability in real-world scenarios. More recently, attempts have been made to enable silent speech communication using video-based recognition, referred to as lip reading [5, 25, 58, 56, 57, 275, 57, 36, 14, 59, 229, 276]. Video-based silent speech input methods could be more user friendly and appropriate in private and public settings since it can be used without any wearable devices. Existing video-based speech recognition models, however, are slow (takes about 5 seconds to recognize one word) and error prone (4–47% error rate). In addition, research found that silent speech performance is highly dependent on extraneous factors, including uncontrolled lighting, blur, low resolution, compression artifacts, occlusions, viewing angles, etc. However, most of the factors can be mitigated by replacing the hardware (blur, low-resolution, compression artifacts, etc.) or by the user (occlusions, viewing angles, etc.). Lighting, in contrast, is one of the factors that cannot always be controlled. Not accounting for this in a vision-based speech recognition compromises its fairness and reduces its applicability in real-world scenarios. We believe, a model that does not suffer from these shortcomings could potentially be used as a medium for input and interaction with various computer systems, incorporated in day-to-day usage.

Research also shows that no matter how robust the speech recognition system is, it could still fail due to variabilities in user characteristics, such as high disfluency, non-canonical pronunciation, accents, speaking rate, and acoustic and prosodic variability [101]. To avoid potential speech misrecognition, users often monitor their behaviors to adjust and optimize future task performance according to experienced errors or conflicts [21, 292, 250]. They engage themselves in processes of repairing the errors by either reformulation, simplification, or hyperenunciation [168, 181, 207, 142, 176, 225]. However, peoples' approach to silent speech input to avoid potential misrecognition is unknown. Speaking rate is a fundamental user characteristics that can influence speech recognition performance due to the variation in acoustic properties of human speech production, such as vowel and consonant duration, the transition between phoneme and stops, and distortions in the temporal and spectral domains [101, 91, 305]. Some studies report that faster speaking rates result in higher error rates [91, 261, 265, 201], whereas some identified slower speaking rates to be more error prone [101, 266]. This disagreement encourages re-investigation of the effects of speaking rates on speech recognition performance. Besides, no such investigations have been conducted for silent speech recognition. The findings from such an investigation could provide guidance to users on improving their speech and silent speech input performance.

As of yet, we have focused on understanding the acceptability and usability of silent speech input and how to make it more accurate and efficient. However, to the best of our knowledge, no one has yet investigated the possibility of using silent speech with computer systems. We took a step forward by introducing silent speech as an alternative hands-free selection method for eye-gaze pointing. It could be advantageous to use silent speech since silent speech does not require external hardware, rather both eye tracking and speech recognition can occur through the same webcam.

1.1 Contributions

This dissertation makes the following contributions:

- First, we conduct an online survey to explore users' attitudes towards the speech and silent speech input methods with a particular focus on social acceptance. The survey examined social acceptance of these methods considering different factors, including users' and viewers' perspectives towards using these in different locations and in front of different audiences. The results suggest that social acceptability for the two input methods from users' and viewers' perspectives were different across locations as users considered the less noticeable input method (i.e., silent speech) as their preferred method to interact with mobile devices.

- Second, we conduct user studies to explore users' attitude towards recognition errors associated with speech and silent speech input methods. Results reveal that users are willing to tolerate more errors with silent speech input than speech input as it offers a higher degree of privacy and security. Inspired by the findings, we conduct another study to investigate suitable feedback method for silent speech input. Results show that users find both a commonly used video and an abstract (i.e., a blinking dot) feedback effective but the latter significantly more private, more secure, and less intrusive than the video feedback. Then, based on the findings, we propose a set of recommendations for using silent speech input on mobile devices.

- Third, we develop LipType, an optimized silent speech recognition model for improved speed and accuracy. LipType demonstrated a significant improvement in the performance of state-of-the-art silent speech recognition model. Results revealed 46.9% reduction in word error rate, 39.1% increase in words per minute, and 8.6 seconds reduction in computation time.

- Fourth, we develop an independent repair model that processes video input for poor lighting conditions, when applicable, and corrects potential errors in output for increased accuracy. In an evaluation, the repair model demonstrated its effectiveness with various speech and silent speech recognition models. On average, speech and silent speech models showed 32% and 57% reduction in word error rates, respectively, without severely compromising the overall computation time.

- Fifth, we explore whether native and non-native speakers interact differently with speech and silent speech-based methods, whether speaking rate affects recognition rates of these methods, the optimal speaking rates for increased accuracy, and whether the effects of speaking rate are different for native and non-native speakers. Results revealed that native users speak about 8% faster than non-native users, but both groups slow down at comparable rates (34–40%) when interacting with these methods, mostly to increase their accuracy rates. A follow-up study confirms that slowing down does improve the accuracy of these methods. Both methods yield the best accuracy rates when speaking at 0.75x of the actual speaking rate. A post-hoc error analysis revealed that speech and silent speech methods and native and non-native speakers are susceptible to different types of errors.

- Finally, we investigate silent speech as an alternative selection method for eye-gaze pointing. Towards this, we propose a stripped-down silent speech recognition model that can recognize a small number of silent commands almost as fast as state-of-the-art speech recognition models. Second, we design a silent speech-based selection method and compare it with other hands-free selection methods, namely dwell and speech, in a Fitts' law study. Results revealed that speech and silent speech are comparable in throughput and selection time, but the latter is significantly more accurate than the other methods. We follow-up on this by conducting another study investigating the most effective screen areas for eye-gaze pointing in terms of throughput, pointing time, and error rate. Results revealed that target selection around the center of a display is significantly faster and more accurate, while around the top corners and the bottom are slower and error prone. Finally, we design a silent speech-based menu selection method for eye-gaze pointing and evaluate it in an empirical study. A study revealed that it significantly reduces task completion time and error rate.

1.2 Brief Outline

This dissertation begins with a review of topics that received much attention in the speech-language processing and human-computer interaction research communities, including speech input, silent speech input, social acceptance of technology, silent speech recognition, effects of speaking rate, low-light image enhancement, recognition error correction, hands-free selection methods, and gaze-based menu selection. It then explores the social acceptance of speech and silent speech input in different social contexts in Chapter 3. Chapter 4 investigates the user tolerance of recognition errors in the speech and silent speech input methods, followed by identifying suitable feedback mechanism for silent speech input. Chapter 5 develops and validates LipType, an optimized silent speech recognition model for improved speed and accuracy. Chapter 6 develops an independent repair model accounting for poor lighting conditions and potential recognition errors. Chapter 7 explores whether native and non-native speakers interact differently with speech and silent speech-based methods, whether speaking rate affects recognition rates of these methods, the optimal speaking rates for increased accuracy, and whether the effects of speaking rate are different for native and non-native speakers. Chapter 8 studied the feasibility of using silent speech as a hands-free selection method in eye-gaze pointing. Finally, Chapter 9 concludes this dissertation and speculates on future research opportunities.

Chapter 2

Related Work

This work intersects with the following areas of interest: speech input, silent speech input, social acceptance of technology, silent speech recognition, effects of speaking rate, low-light image enhancement, recognition error correction, hands-free selection methods, and gaze-based menu selection.

2.1 Speech Input

Speech input enabled devices, such as personal voice assistants, allow users to communicate with computer systems using speech commands. Personal voice assistants like Siri, Google Assistant, Alexa, and Cortana can interpret human speech and handle a wide variety of tasks [127, 177]. Such speech interfaces can potentially improve productivity and user comfort when traditional input methods, like touch and keyboards, are inefficient, difficult, or inconvenient to use [248, 116]. Yet, users of speech input are usually unsatisfied with the quality of interaction due to low recognition accuracy [74]. To avoid potential errors, users tend to modify their speaking styles and patterns [168, 181, 207, 142, 176, 225] by shortening their sentences [153, 225], performing repetition [40, 55], increasing the volume [81, 44], and hyper-articulating [217]. However, studies showed that automatic speech recognition (ASR) can fail even when these strategies are applied due to high levels of disfluency, non-canonical pronunciation, accent, speaking rate, and acoustic and prosodic variability [101]. [179] reported that recognition is worse for words that are phonetically similar to other words than for highly distinctive words. [261] found out that longer words have slightly lower error rates than shorter words. [126] showed that infrequent words are more likely to be misrecognized. A different research found a correlation between large fluctuations in the short-term speaking rate and high recognition errors [15]. Another work reported that male speakers have significantly higher recognition error rates than female speakers due to higher rates of disfluency [3]. Relevantly, misrecognized words were found to have higher pitch and energy than correctly recognized words [123]. Another study revealed that words with more possible pronunciations have higher error rates and longer words have slightly lower error rates [101].

Recently, research has mostly focused on improving the performance of speech input by developing robust speech recognition models [194, 255, 304], language models [31, 13] and voice controlled systems [310]. [63] provides a comprehensive review of the literature on speech-based input and interaction methods. With the recent advances in speech recognition technology [106, 2, 230, 234, 260], today's voice-based commercial products [307, 298, 115, 46, 158, 298, 297] can perform streaming, high-accuracy, low-latency speech recognition [28, 172] to revolutionize human-computer interaction [63]. Recently, [118] presented an end-to-end speech recognizer for on-device speech recognition using a recurrent neural network, which has been deployed in the default Google keyboard on the flagship Pixel phones. Despite its popularity, studies show privacy and security concerns for the use of personal voice assistants and voice search commands in public places [82, 85, 83, 235]. A survey[1] revealed that 39% smartphone users use the built-in voice assistants at home but only 6-14% use these in public [228]. To uphold the privacy and security of users, researchers explored whisper input, which is a variant of speech input with a significantly lower energy than normal speech. These works detected whispered speech using a stethoscopic microphone that contacts the skin behind the ear [208], a throat microphone [148], and a non-contact microphone by placing it very close to the front of the narrowly opened mouth [93]. Recently, Amazon included a whisper mode to their personal voice assistant Alexa[2]. When users whisper to Alexa, it whispers back to them. Some have also incorporated state-of-the-art machine learning techniques to improve the performance of whisper speech recognition [97, 108, 96]. However, whispers with a much lower acoustic power and relatively flat spectrum than regular speech are inherently noise-like, thus are highly susceptible to acoustic interference [195]. Moreover, long-term use of whisper voice might have negative effects on our vocal cords [249].

2.2 Silent Speech Input

Silent speech input enables users to communicate with a computer system using speech commands without the need for producing any audible sound. Unlike speech input, silent speech input can be effective when speaking audibly could disturb others or disclose confidential information, to understand elderly and children speech, and to provide people with speech and motor impairments access to computer systems. There have been several previous attempts at achieving silent speech communication. Many have explored silent speech enabled input and interaction methods that use different sensors (e.g., electromagnetic articulography (EMA) [87, 99, 120], electroencephalogram (EEG) [231], electromyography (EMG) [291, 146, 145, 147, 189, 254], ultrasound imaging [156, 90, 73, 72, 129, 130, 99, 120], vibrational sensors of glottal activity [213, 223, 238, 281], speech motor cortex implants [32], and non-audible murmur (NAM) microphone [122, 209, 121]) to recover the speech content produced without vibration of the vocal folds, by detecting tongue, facial, and throat movements. Some have developed intracortical microelectrode Brain-Computer Interfaces

[1] Fortune, https://fortune.com/2016/06/06/siri-use-public-apple/

[2] Digital Trends, https://www.digitaltrends.com/home/how-to-enable-whisper-mode-on-alexa

(BCI) to predict user's intended speech information directly from the brain activities involved in the speech production mechanism [47, 71, 231, 277, 278]. Some have also used multimodal imaging systems for speech recognition, focusing mainly on tongue visualization [130]. A recent work developed a wearable interface that places five EMG sensors above the face to capture the neuromuscular signals for silent speech recognition [150]. Most of these works, however, use invasive, impractical, non-portable setup, impeding their scalability in real-world scenarios.

More recently, attempts have been made to enable silent speech communication using video-based recognition, referred to as lip reading [5, 25, 58, 56, 57, 275, 57, 36, 14, 59, 229]. For example, a work provided smartphone users access to their phone functionalities through silent speech commands [276]. It used the front camera of a smartphone to capture the motion of the mouth, then recognized the silently spoken commands using deep-learning-based image sequence recognition technology. These works suggest that video-based silent speech input method could be more user friendly and appropriate in private and public settings since it can be used without any wearable devices.

Despite these improvements, research found that silent speech could be error-prone due to its dependence on extraneous factors like lighting, skin complexion, posture, head rotation, and facial expression [144, 272, 211, 41]. In recent investigations, users reported a higher level of satisfaction using this method than speech input in some scenarios [276]. Having a model that mitigates the above challenges can widen its usability in more scenarios, allowing input and interactions on private devices without the hands, and on public devices without direct contact as in the COVID-19 situation. It can also help people with speech disorder, muteness, and blindness to input and interact with computer systems, increasing their access to technologies.

2.3 Social Acceptance of Technology

Previous research has explored social acceptability for body-based and device-based gestures [242, 243, 246], around device input [8], head-mounted display (HMD) input [10], and companion drones for blind people [26] in lab or public settings. In a recent work, [30] explored the social acceptability of speech input, which revealed that location influences users' willingness to use the method in public spaces. However, no prior study has explored user attitudes and acceptance of using silent speech input. In a different research, [9] investigated whether social acceptability studies can be conducted on crowdsourced platforms. They showed that crowdsourced platforms could be an alternative to conducting laboratory-style studies for examining social acceptability. Inspired by this work, we conducted our social acceptability study via crowdsourcing.

Prior research also showed that social acceptability has a significant implication for technological acceptance as it is directly connected to peoples' preferences on using new technologies [282, 160]. To examine the social acceptance of new technologies, researchers conducted studies from users' perspective and/or viewers' perspective [8, 242, 243, 10]. To investigate users' perspective, researchers either provided participants with a first-hand experience using a new technology or showed them video clips on how the technology could potentially be used [8]. Later, participants were asked to consider themselves as users of the technology and express their opinion on using it in different contexts. While there are many social acceptability studies conducted from the users' perspective, less attention has been paid to examine social acceptance from viewers' standpoints. A few studies investigated social acceptance from the viewers' perspective where researchers elicited opinions from people watching others using a new technology in different contexts. [203] showed that considering viewers viewpoint is important, especially when using the technology in public places, as users' interactions with the technology might draw bystanders' (or the viewers') unwanted attention. Consequently, viewers' perspective are explored for wearable e-textile interface [237, 236], Augmented Reality (AR) in public space [75], and public interfaces (e.g., public performance act) [241]. Additionally, some studies considered both the users' and the viewers' perspectives while evaluating the social acceptance of new technologies, such as gestural interaction on mobile devices [203], head-worn devices [161, 9, 10, 180], data glass [162], and around device input methods [8]. These studies were commonly conducted by examining observers' impression on watching other people interacting with a technology — either in a real-world setting or in a video. In this work, we examined the social acceptability of speech and silent speech input from both users' and viewers' perspectives.

2.4 Silent Speech Recognition

There is a rich literature on silent speech recognition. Here, we only discuss the works that are closely related to ours (see [312] for a comprehensive review). Recently, there have been attempts to apply deep learning to silent speech recognition [58, 56, 275, 57, 36, 14, 59]. However, most of these approaches perform only at phoneme- or word-level. [163] trained an image classifier using convolutional neural network (CNN) to differentiate between visemes[3] on a sign language dataset of signers mouthing words. [215] also used CNN to predict phonemes in spoken Japanese. [280] used deep bottleneck features (DBF) to encode shallow input features, such as latent dirichlet allocation (LDA) and GA-based informative feature (GIF) [283] for word recognition. [229] also used DBF to encode every video frame and trained a long short-term memory (LSTM) classifier for word-level classification. [290], on the other hand, used an LSTM with histogram of oriented gradient (HoG) input features to recognize words. [58] developed CNN architectures for classifying multi-frame time series of lip movements. LipNet [25] is an end-to-end model for phrase-level lip reading by predicting character sequences (further discussed in a later section). [5] also enabled phrase-level lip reading by utilizing an encoder-decoder structure with multi-head attentions. [60] developed the Watch, Listen, Attend and Spell (WLAS) network that uses dual attention mechanism for visual attention to transcribe videos of mouth motion to characters.

[3] Visemes are visual equivalent of phonemes. A viseme represents the position of the face and mouth when making a sound.

2.5 Effects of Speaking Rate

Different speaking rates can significantly affect speech recognition performance due to a distorted spectrum caused by variations in speaking rate [101, 91, 305]. Natural speaking rate depends on user characteristics like gender, age, accents, and psychological state. [303] showed that older people speak slowly compared to young adults, and women talk slower than men. [240] reported that people usually speak fast when in a hurry or angry, and slow when they are tired, sad, or sick. Studies also showed that non-native speakers talk much slower [109] and exhibit more variation in speaking rate than native speakers [27]. However, suprasegmental characteristics between native and non-native speakers in spontaneous speech suggest that non-native speakers are less variable than native speakers [204], which can affect recognition rate [222], particularly for non-native speakers [296, 70, 24]. However, the research community is divided on how speaking rate affects recognition accuracy. Some associated faster speech with higher error rates [91, 261, 265, 201], while others found slow speech to be more error-prone [101, 266].

2.6 Low-Light Image Enhancement

The problems of underexposed low-light images are very common, solutions to mitigate it have been a popular research topic. Researchers have developed a variety of techniques that can improve image quality. The classical image enhancement methods involve two categories: i) retinex-based methods, which are based on retinex theory [169]. Recent examples of these approaches are Lime [110], naturalness preserved enhancement [293], Retinex [143], and simultaneous reflectance and illumination estimation [92]. ii) histogram equalization methods, which manipulate the gray levels of individual pixels based on the image histogram. Recent examples include contextual and variational contrast enhancement [50], weighted thresholded histogram equalization [17], and layered difference representation [170]. In recent years, several methods based on deep learning image processing techniques have been proposed. One successful example is the developed pipeline for processing low-light images, based on end-to-end training of a fully-convolutional network [52]. However, they reported that their model showed imperfect results for humans faces. Another work [294] utilizes encoder-decoder network to achieve the low-light enhancement for real under exposed images. Other works [299, 4, 173] have also showed the effectiveness of deep learning methods on low light image enhancement.

2.7 Recognition Error Correction

Automatic detection and correction of recognition errors have become an important research area. The aim is to automatically detect and partially or fully correct errors, regardless of the recognition system used. Zhou et al. [311] addressed the issue of error detection in recognition systems using data-mining classifiers such as naive Bayes (NB), neural networks

(NN), and support vector machines (SVM). These classifiers were trained to identify errors using confidence scores and linguistic information present in the recognized output. Another work [12] proposed extraction of additional features from the confusion networks to estimate correctness probability using logistic regression. Pellegrini et al. [227] investigated the use of a Markov chains (MC) classifier with two states: error state and correct state, to model errors. Chen et al. [54] proposed a system for error detection in conversational spoken language translation. This system uses additional features provided as the feedback of statistical machine translation (SMT), including SMT confidence estimates, posteriors from named entity detection (NED), and an automated word boundary detector to verify the word boundaries of recognition output, in order to improve error detection and correction. Sarma et al. [252] built a recognition error detector and corrector using co-occurrence analysis. In the same context, Bassil and Semaan [33] proposed a post-editing ASR error correction method based on Microsoft N-Gram dataset for detecting and correcting spelling errors generated by recognition systems. The detection process detects on-word spelling errors in reference with the Microsoft N-Gram dataset, and the correction process generates correction suggestions for the detected word errors by selecting the best candidate for the correction using contextual information. Other works [226, 259, 94, 175] have explored a non-decoder based post-processing error detection and correction.

2.8 Hands-Free Selection Methods

There is a rich body of work on selection methods for gaze pointing. Most of these works, however, explore manual approaches that require the use of the hands, particularly mid-air gesture (e.g., [51, 233, 247]) and physical keys, buttons, and controllers (e.g., [167, 287, 206, 166]). In this section, we only cover hands-free selection methods that are accessible to people with limited motor skills.

Dwell is the most commonly used hands-free selection method in gaze pointing. It enables users to look at a target for a predetermined period of time to trigger selection [112]. This method is popular due to its simplicity and because it does not require the use of additional sensors like microphones, depth cameras, or motion sensors. However, it is difficult to maintain a sensible balance between speed and accuracy when selecting a dwell time. A short dwell time makes a system faster but increases the chance of unwanted selections, while a long dwell time makes the system slower and can cause users physical and cognitive stress [37, 112]. To address this, several works have enabled users to adjust the dwell time [190] or automatically adjusted dwell time based on user experience [271, 205]. While these approaches improved the performance of dwell, it remains a time-consuming and error prone selection method in gaze pointing.

Many alternatives have been proposed to substitute dwell. [80] explored gaze gestures with eye tracking, where users performed specific eye movements for target selection. Studies suggested users can perform complex gaze gestures intentionally [80, 131]. A follow-up study showed that gaze gestures can enable people with motor impairments to play online games

[135]. Some have used specific types of gaze gestures (e.g., reverse crossing [88] and single gaze gestures [202]) and blinking [23] for target selection. However, performing intentional gaze gestures and blinking are unnatural [138], thus can cause users irritation and fatigue. Several works, in contrast, studied target selection through voluntary facial muscle activation [279, 193], brain signals [125], and foot pedals [198]. These methods use external and invasive hardware, thus not yet scalable in practical situations. Some have also attempted head gestures for target selection [264, 263, 198], which performed well in short-term use, but can cause fatigue in extended use. Many have combined gaze with speech, which is potentially a more natural and efficient mode of interaction [268, 221]. These works either use a single command to confirm selection [38] or multiple commands to facilitate both pointing and selection [257, 200]. Speech is promising but unreliable in noisy places and users are often hesitant to use speech in public places [82, 86, 84, 235]. Besides, speech does not work well with people with severe speech disorder [7, 43].

2.9 Gaze-Based Menu Selection

Not much work has focused on gaze-based menu selection methods. Menu selection is different than individual target selection (e.g., virtual keys, buttons, or links) since the former involves the selection of a sequence of horizontal and vertical targets. Error in one selection task results in an incorrect output, forcing the user to correct the mistake, then re-perform all tasks in the sequence. Menu selection, thus, has a much higher error correction overhead. Almost all gaze-based menu selection methods use a “zooming” approach that dynamically increases the size of a potential target to facilitate precise selection [270, 267, 35, 199]. These methods, however, do not provide an effective mechanism for controlling the zooming behavior, which can cause frustration when the method does not behave as expected. Expanding the menu items can also occlude the content in the background, causing inconvenience. [206] positions the cursor at the center of a target by suppressing cursor movements caused by involuntary eye movements. [149] enable users to select a target my making a “click” sound when the cursor is over it. [206] enable users to speak the items in a menu to select them. Some also explored different menu designs (e.g., radial, semi-circular, etc.) for gaze pointing [149, 284].

Chapter 3

Acceptability of Speech and Silent Speech in Private and Public

This Chapter presents a user study exploring users' attitudes towards speech and silent speech input methods with a particular focus on social acceptance. We conduct a crowd-sourced study examining social acceptance of these methods considering different factors, including users' and viewers' perspectives towards using these in different locations and in front of different audiences. Results show that, in general, people prefer using silent speech input over traditional speech input.

3.1 Social Acceptability

In this study, we will examine the social acceptability of speech and silent speech input from both users' and viewers' perspectives.

3.1.1 Input Modalities

Researchers have explored a number of voice and non-voice input modalities to interact with mobile devices. For instance, they investigated using speech and silent speech input methods that range from noticeable to inconspicuous [310, 87, 146, 156, 223, 47, 276]. Speech or voice input, which is commercially available on smartphones, requires users to make voice commands to send instructions to mobile devices. This input modality is explicit and commonly draws co-located observers' attention due to the nature of its input visibility — thus can make users feel awkward or uncomfortable with the presence of nearby users. On the other hand, silent speech input, which recognizes speech without requiring users to make acoustic signals, interprets users' commands on smartphones by tracking tongue and lip movements. This input method is more subtle than the speech input, and used when acoustics is not an option (e.g., speech-impaired people) or it is undesired (e.g., during a confidential conversation or communication in public places). On one hand, using explicit input modalities can convey clear instructions to the devices; however, this form of input might be less socially acceptable due to the visibility to co-located people. On the other hand, subtle inputs are less explicit; however, co-located observers might not readily interpret these commands, making the interaction more acceptable. Therefore, we first conduct a study to explore the social acceptability of these two input modalities.

3.1.2 Crowdsourced Study

As discussed in the related work, researchers explored social acceptability for a wide range of input modalities, such as smartphone gestures [246, 242, 243], around-device interaction [8], and hand-to-face input methods [258, 171]. They used two common approaches: (i) allowing participants to use the technology in a particular context (e.g., public places) and (ii) showing participants videos of how the technique can be used. To collected feedback, participants are commonly asked to imagine using it in other contexts (e.g., workplace) and provide their feedback on a 5-point Likert scale. Due to the spread of COVID-19, we were unable to recruit participants to run a study in a public place. Thus, we used the second approach for our study.

Crowdsourcing platforms have now become increasingly popular to conduct HCI user studies [9, 10]. They provide researchers with an easy access to large and diverse groups of participants. Additionally, these platforms have been considered as cost-efficient solutions to run user studies remotely. Though there has been concern about the data quality from crowdsourced studies, researchers have taken certain measures to remove outliers, which have been almost as effective as laboratory or field studies [39, 104, 9, 10]. Consequently, we decided to use crowdsourcing platforms to run our first study.

3.1.3 Online Survey

We created an online survey with Qualtrics to collect responses from participants. Figure 3.2 shows a sample of questions from the survey. We divided the survey questions into four sections: (i) *Demographics*: 14 questions to collect demographic information (e.g., age, gender) and prior experience (e.g., experience with smartphones and voice input) from participants; (ii) *Users' perspectives*: 6 questions asking users to share their experience of using speech and silent speech input methods by considering themselves as users of the modalities; (iii) *Observers' perspectives*: 6 questions were used to explore observers' perspective, i.e., seeing other people using the input modalities and (iv) *Overall preference*: 6 questions asking participants to provide their overall preference of using the input modalities on mobile devices. These questions were designed using both open-ended questions, single/multiple-choice questions, and 5-point Likert scale questions. The open-ended questions were used to collect descriptive responses (e.g., justifying their response to a question), while the other types of question were used to collect their preference/perception of using the input modalities and demographic information. When designing the questionnaire, we used similar questions and location-audience contexts used in previous work on social acceptance [8, 10, 9, 242,

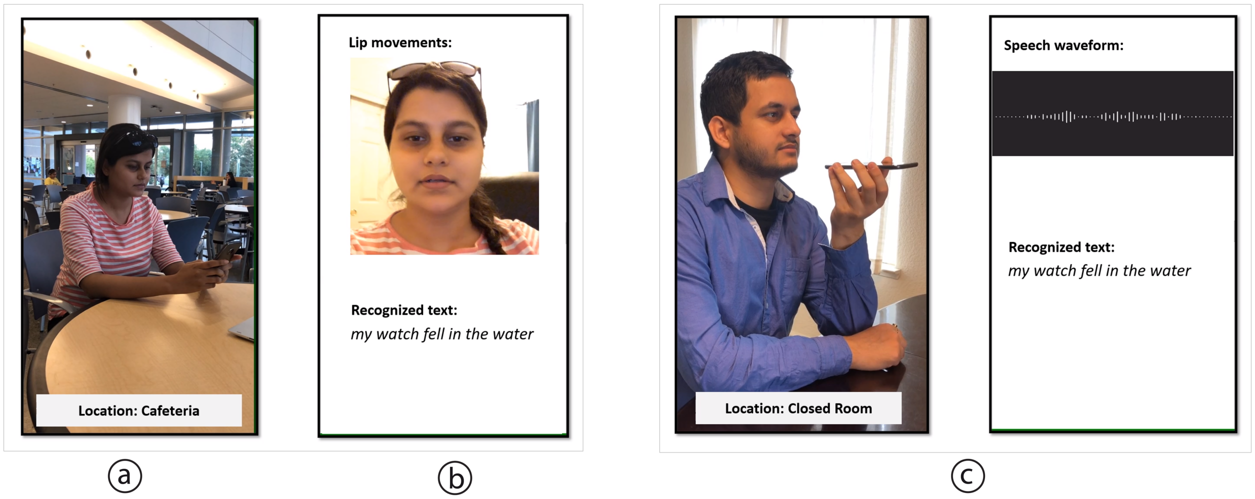
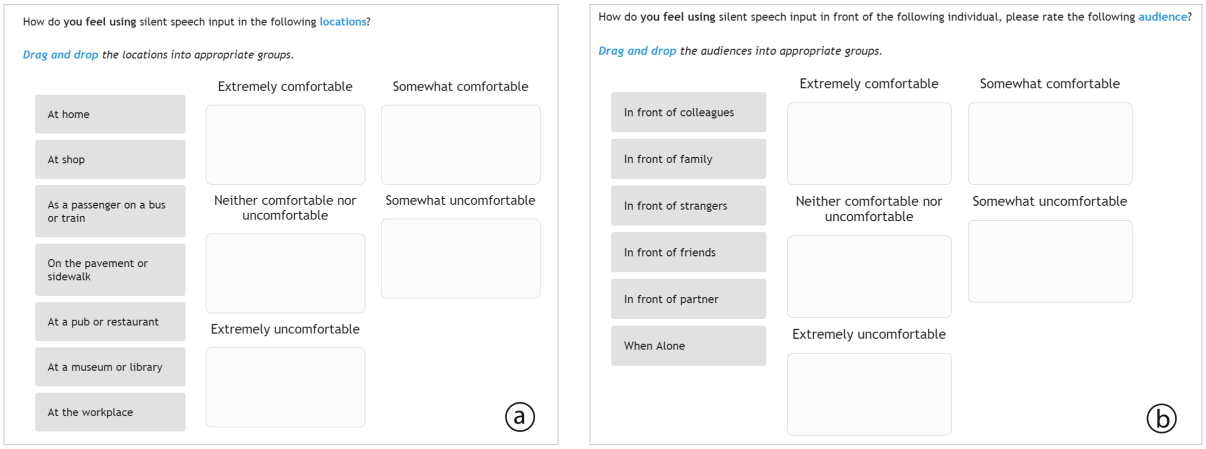


Figure 3.1: Two example videos used in the survey: (a) a user is interacting with a mobile device with silent speech input in a public place, (b) a video clip showing users lip movements and the recognized text, and (c) another video showing a user using speech input on a mobile device in a private room.


How do **you feel using** silent speech input in the following locations?

Drag and drop the locations into appropriate groups.

How do **you feel using** silent speech input in front of the following individual, please rate the following audience?

Drag and drop the audiences into appropriate groups.

Figure 3.2: Example of survey questions to collect user feedback on using silent speech input (a) in seven locations; and (b) in front of six audiences.

243]. We also followed many steps listed by [42], including item generation, context validity, pre-testing with a pilot study, item reductions and others.

Researchers explored a number of ways to measure social acceptabilities of the methods under investigation. One of the commonly used methods is to elicit participants' responses to social acceptability questions through the 'audience-and-location' axes [8, 9, 10, 242], where participants are asked to provide their social comfortness of using a method in front of different audiences and locations. Participants commonly respond by indicating how comfortable they were using the method on a 5-point Likert scale – Extremely comfortable, Somewhat comfortable, Neither comfortable nor uncomfortable, Somewhat uncomfortable, and Extremely uncomfortable. Therefore, we used six audiences (i.e., alone, partner, family, friends, colleagues, and strangers) and seven locations (i.e., home, shop, bus or train, pavement or sidewalk, pub or restaurant, museum or library, and workplace) to explore participants' impression of using the two input methods (i.e., speech and silent speech). As participants might not be familiar with a input method, we used a set of video clips showing users using the two methods to interact with a mobile device in two different contexts – in a busy café surrounded by strangers and at home when alone.

3.1.4 Participants and Study Procedure

To recruit participants, we posted the survey as a task in Amazon Mechanical Turk (AMT), a popular Crowdsourcing platform. All AMT users (i.e., workers) could see the task, however, only the workers who owned a smartphone and had a minimum of 70% approval rate on their previously completed tasks could participate. Workers were compensated with USD $1.50 for their time. We collected data from 109 crowdsourced participants. 62 of them were from the U.S., 6 were from India, 2 were from Brazil, and 1 was from Germany. 8 of them were in the age range of 18–24 years, 28 were in 25–34 years, 18 were in 35–44 years, 10 were in 45–54 years, 5 were in 55–64 years, and 2 were 65 years or older.

The survey was self-paced and the workers were asked to first watch the video clips for an input method, then respond to the questions related to that method. We also clearly instructed them not to relate comfort with physical comfort (e.g., tiredness), rather focus on social and mental aspects of it when providing their responses. Similar strategies were applied in previous studies exploring the social acceptance of new input modalities [8].

As mentioned earlier, data collected from crowdsourcing platforms sometimes raises concerns due to the lack of direct supervision of the workers. Thus, we used the following criteria to remove outliers from our data. (i) Duplicate IP address: we removed any data with the same IP address. This outlier removal technique was also used in prior studies [9, 10]. (ii) Time threshold: as participants were required to watch a set of videos before responding to the questions, they had to spend a minimum time to watch the videos and read and understand the questions before answering them. Consequently, any responses that were submitted within 3 minutes of start were excluded from our analysis. (iii) Incorrect answers: there were a few open-ended questions asking participants to provide justifications for their responses. Any data with incorrect, incomplete, or random answers were rejected. This process excluded in total 38 participants. Hence, we analyzed the data from 71 participants.

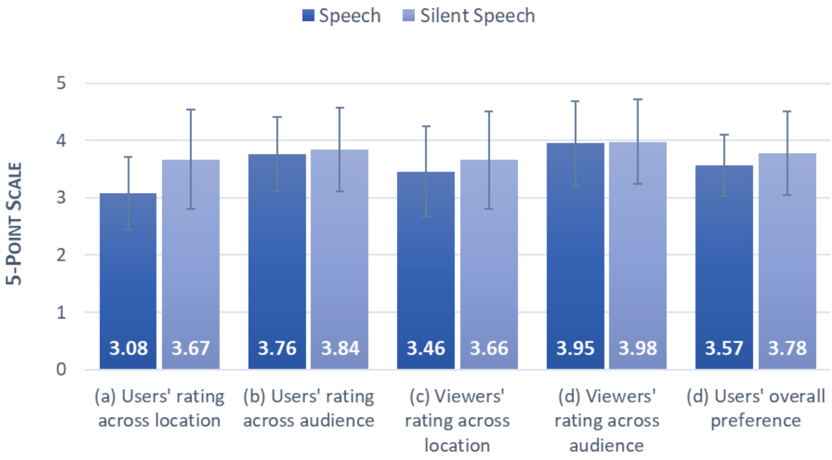


Figure 3.3: Medians of social acceptability for two input methods from users' perspective across (a) location, (b) audiences and from viewers' perspective across (c) location and (d) audiences, and (e) users' overall preference for two input methods. The error bars represent ±1 standard deviation (SD).

3.2 Results

We used non-parametric analyses on the data and, thus, median values are reported. We also report the effect size (r) for the Wilcoxon signed-rank test. Since r for the Friedman test is calculated for pairwise comparison and there is not an agreed method for calculating the confidence interval [245], Kendall's W is most commonly used to assess agreement among the raters. Hence, we report W for the Friedman test. Both r and W use the Cohen's interpretation where 0.1 constitutes a small, 0.3 constitutes a medium, and > 0.5 constitutes a large effect. We aggregated users ratings for each input across all the locations and audiences.

Figure 3.3 (a) and (b) show the median of social acceptability for each input across locations and audiences, respectively, from users' perspective. A Wilcoxon signed-rank test revealed significant differences between the speech and silent speech input methods across locations ($z = -4.59, p < .05, r = 0.54$). However, we found no significant difference between aggregated values for two input methods across audiences ($z = -1.36, p = .17, r = 0.16$). Figure 3.3 (c) and (d) show the median of social acceptability ratings for each input across locations and audiences, respectively, from viewers' perspective. A Wilcoxon signed-rank test showed that silent speech input was significantly different from speech input ($z = -2.5, p < 0.05, r = 0.30$) across locations. However, we did not find any significant difference between two input methods across audiences ($z = -1.14, p = .26, r = 0.14$). We also asked participants to provide their preference for using the two input methods to interact with mobile devices across locations and audiences. Figure 3.3 (e) shows the results. A Wilcoxon signed-rank test revealed significant differences between speech and silent speech input methods

($z = -3.27, p < .05, r = 0.39$). We recommend caution in interpreting the "not significant" results since they yielded a small effect size ($r < 0.3$).

3.3 Discussion

The results suggest that social acceptability for the two input modalities from users' and viewers' perspectives were different across locations as users considered the less noticeable input method (e.g., silent speech) as their preferred method to interact with mobile devices. Similar findings were revealed in a prior work [9], where they suggested that less noticeable input methods (e.g., ring and touchpad) are more socially acceptable than noticeable ones (e.g., hand gestures) to interact with an HMD. The results also show that participants preferred to use silent speech input over speech input. In subjective feedback, participants expressed their interest in using silent speech input as it is more subtle and provide a high degree of privacy and security than the other method. One participant (male, 35–44 years) commented, *"I would still feel that I have a high level of privacy when using silent input"*. Another participant (female, 35–44 years) wrote, *"I prefer whisper or silent because it doesn't bother others and can be used in quiet places like libraries"*.

Though the results showed users' interest in using silent speech input, several key questions remain unknown that could influence their attitude towards using the method. For instance, researchers showed that silent speech input could be prone to high error rates [174, 232, 211, 74]. Consequently, silent speech recognition accuracy could be a key factor in adopting the method. However, little is know of users' error tolerance level for silent speech input. Additionally, silent speech input recognition on mobile devices depends primarily on capturing users tongue and lip movements via the front camera. Thus, providing appropriate real-time feedback on input recognition is critical for the acceptance of the method. Therefore, in the next Chapter, we explore error tolerance and suitable feedback mechanism for silent speech input.

Chapter 4

Error Tolerance and Real-time Feedback for Silent Speech

This chapter first examines users' error tolerance with speech and silent speech input methods, where results revealed their willingness to tolerate more errors with silent speech for the sake of privacy and security. Second, it explores a suitable feedback for silent speech input, we observed that users found both a commonly used video and an abstract (a blinking dot) feedback effective but the latter significantly more private, more secure, and less intrusive than the video feedback. We learned that designing solutions for silent speech input requires careful consideration of various factors and privacy concerns as well as people's tolerance towards using it.

4.1 Error Tolerance

Since the survey results revealed that users put much emphasis on privacy and security, we conducted a Wizard-of-Oz study to investigate whether they are willing to compromise the accuracy of an input method for increased privacy and security.

4.1.1 Apparatus

We developed a custom client/server web application with HTML5 and JavaScript for the Wizard-of-Oz study. The client and server communicated with each other using WebRTC[1]. The client interface looked and felt like the interface depicted in Fig. 3.1. It was launched on a Google Chrome mobile web browser (v71.0.3578.98) on a Motorola Moto G^5 Plus smartphone (150.2x74x7.7 mm, 155 g) at 1080x1920 pixels. The server was hosted on a HP Pavilion 15 laptop computer running on Linux 16.04 at 1920×1080 pixels. The server interface was launched on a Google Chrome web browser (v74.0.3729.157), which included dedicated buttons for each condition for the researcher (wizard) to display the spoken and

[1] Real-time communication for the web, https://webrtc.org

silently spoken phrases on the client side. Both devices were connected to a fast and reliable Wi-Fi network. There were no network dropouts during the study.

4.1.2 Participants

Twelve volunteers from the local university community participated in the user study. Their age ranged from 22 to 25 years (M = 24.25, SD = 1.48). Four of them identified as women and eight as men. They were all experienced smartphone (at least 5 years of experience, M = 7.25, SD = 1.48) and voice assistant (at least one year of experience, M = 2.5 years, SD = 0.65) users. Most of them used multiple voice assistants, including Alexa, Cortana, Google Assistant, and Siri. Two participants used these voice assistants almost every day, eight of them used these occasionally, and the remaining two rarely used these.

4.1.3 Design

The study used a within-subjects design. The independent variables were *method* and *injected error rate* and the dependent variables were the qualitative metrics. In summary, the design was:

12 participants ×

2 methods (speech and silent speech, counterbalanced) ×

5 injected error rates (0%, 5%, 10%, 15%, and 20%, randomized) ×

12 phrases from the [188] set = 1,440 phrases, in total.

4.1.4 Error Injection

Injected errors are commonly used in text entry research to study the effect of errors on performance and preference [164, 20, 22, 11]. In the study, we injected 0%, 5%, 10%, 15%, and 20% misrecognition errors. A misrecognition error occurs when the recognizer incorrectly recognizes a word [22], for example, "take a coffee break" ("coffee" was replaced with "toffee"). The total number of misrecognition errors in a condition was calculated using the following equation: $(w \times e)/100$, where w is the total number of words in *all* presented phrases in the condition and e is the target error rate. We injected errors at word level since both speech and silent speech methods work at either word or phrase level. To inject errors, we randomly replaced a word consisting more than three letters with a similar sounding word, excluding the first word. To assure that all participants encountered the same errors, we randomly pre-selected a subset of phrases from the [188] set, then used those with the methods in a counterbalanced order. The error injection rates were selected based on the findings of a prior investigation the reported that user performance tend to drop significantly when error rate of an input method reaches 20% [22].

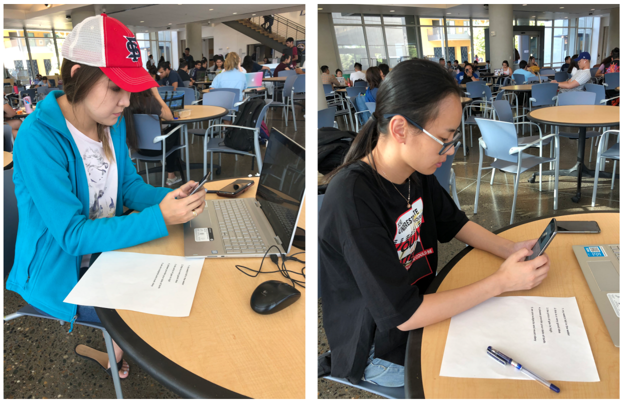


Figure 4.1: Two participants taking part in the second study at a cafeteria.

4.1.5 Procedure

We conducted a Wizard-of-Oz study to control the error rate in each condition. Before the study, participants were told that the purpose of the study was to compare the performance of multiple speech and silent speech recognition methods that may vary in accuracy rate. The study took place at a campus cafeteria. We picked a public place for the study since its purpose was to investigate whether users were willing to tolerate more errors for the sake of increased privacy and security. Note that the survey results suggested that users are likely to be more conscious about their privacy and security when in public. Upon arrival, we demonstrated the speech and silent speech methods on the smartphone and explained the study procedure to each participant. We then collected their consents. The study started after that, where participants were instructed to enter short English phrases from the [188] set using either speech or silent speech at varying injected error rates. The methods were counterbalanced and the error rates were randomly injected to mitigate any potential learning effects. The interface displayed one phrase at a time. Participants were instructed to tap on the screen when they were done speaking or silently speaking the phrase. They all sat at a table in the cafeteria (Fig. 4.1). A researcher (the wizard) sat at a nearby table with the server interface launched on a laptop computer. Upon completion of each phrase, she pressed a key to display the recognized phrase and the next phrase on the smartphone. Participants were asked to speak or silently speak a phrase again when the phrase contained a misrecognized word. Upon completion of each condition (method × injected error rate), participants completed a short questionnaire that asked them to rate their willingness to use the examined methods on a 5-point Likert scale. Upon completion of the complete study, they completed the NASA-TLX questionnaire [210] to rate the methods' perceived workload. We then held a debrief session to explain the study's actual purpose. A complete study session took about 60 minutes.

4.2 Results

We used non-parametric analyses on the data, thus report median values. We also report the effect size r and Kendall's W for the Wilcoxon signed-rank and Friedman tests, respectively (see Section 3.2).

4.2.1 Willingness to Use

A Friedman test identified a significant effect of condition on willingness to use ($\chi^2(9) = 94.04, p < .0001, r = 0.87$). There was a significant effect of injected error rate on willingness to use for both the speech ($\chi^2(4) = 38.06, p < .0001$) and silent speech ($\chi^2(4) = 48.00, p < .0001$) methods. A Dunn's multiple comparisons test identified a significant difference in willingness to use between the methods with both 10% ($z = 2.75, p < .05$) and 15% ($z = 2.83, p < .05$) error rates. Fig. 4.2 illustrates median willingness to use for both methods with the five injected error rates.

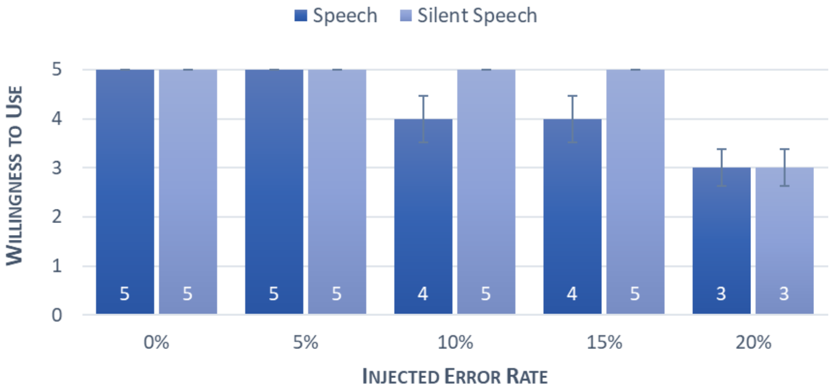


Figure 4.2: Median willingness to use ratings for speech and silent speech with the five injected error rates on a 5-point Likert scale, where where 1 to 5 represented Very unlikely to Very likely. The error bars represent ±1 standard deviation (SD).

4.2.2 Perceived Workload

A Wilcoxon Signed-Rank test identified a significant effect of method on temporal demand ($z = -1.1, p < .05, r = 0.61$) and overall performance ($z = -2.24, p < .05, r = 0.65$). However, no significant effect was identified on mental demand ($z = -1.93, p = .05, r = 0.55$), physical demand ($z = -0.93, p = .35, r = 0.27$), effort ($z = -1.45, p = .15, r = 0.42$), or the level of frustration ($z = -0.99, p = .32$). Fig. 4.3 illustrates median Raw TLX (RTLX) scores for both methods. We analyzed the subscales individually, which is a common modification made to NASA-TLX [113]. Note that the evidence is inconclusive about whether RTLX is more sensitive, less sensitive, or equally sensitive compared to the original version, thus [113] left it to the researchers' discretion.

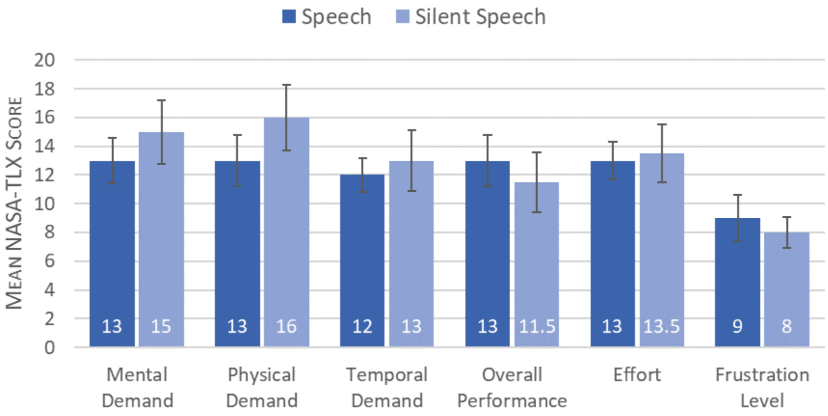


Figure 4.3: Median RTLX scores of the workload related to speech and silent speech methods. The error bars represent ±1 standard deviation (SD).

4.3 Discussion

Results revealed that 0% and 5% error rates yielded the highest and 20% error rate yielded the lowest willingness to use ratings for both methods. This is not surprising since prior investigations reported that user performance with an input method is the best between 0% and 5% error rates, slightly drops between 5% and 10% error rates, and the worst at 20% error rate [20, 22]. Interestingly, for 10% and 15% error rates, the willingness to use ratings for speech dropped at a higher rate that silent speech (Fig. 4.2). A post hoc analysis failed to identify a significant difference between 0–5% and 10–15% error rates for silent speech, while these two groups were significantly different for speech. This suggests that users were willing to tolerate more errors in silent speech. When asked about this during the debrief session, all participants (100%) responded that it was mostly due to concerns about their privacy and security. They feared that speech will violate their privacy and security in public places, especially when they are surrounded by unknown people. One participant (female, 22 years) commented, *"Sometimes, I feel very hesitant to type with my voice publicly because I always feel that someone else is listening to me"*. In contrast, participants felt that silent speech is more private and more secure, thus were willing to compromise accuracy to some extent. One participant (male, 23 years) commented, *"[Silent speech] is very useful for sharing important information in public"*.

There was a significant difference in temporal demand and overall performance for the two methods. Most participants felt that silent speech required more time to use than speech (Fig. 4.3). The debrief session revealed that it was because participants silently spoke the phrases at a much slower rate than speech assuming that it will increase the method's accuracy (although in reality it had no effect since we used a Wizard-of-Oz setup). This also significantly affected their overall rating of the method. There was no significant difference in mental demand, physical demand, effort, and frustration. However, we recommend caution in interpreting these results since in the study participants used the methods while seated at a table. Although we did not instruct them on how to hold the device, they all held the device with both hands for clear view of the interface (Fig. 3.1) and rested their elbow on the table for comfort (Fig. 4.1). Hence, the results may differ when the methods are evaluated in a standing position or while walking.

4.4 Real-time Feedback Mechanism for Silent Speech

Providing appropriate feedback on the system status is the key usability principle while designing any system. Efficient visual feedback helps users to interpret the system status correctly, enabling them to access information rapidly and accurately [182]. However, designing effective visual feedback for mobile devices is challenging due to their limited display space. Besides, some participants of the previous study (see 4.1) commented that the video feedback method occupies much of the smartphone real estate, leaving a little or no space for additional input and interaction tasks (Fig. 3.1). We, therefore, conducted a user study to find out whether it is feasible to replace the commonly used video feedback with a more compact, abstract feedback method.

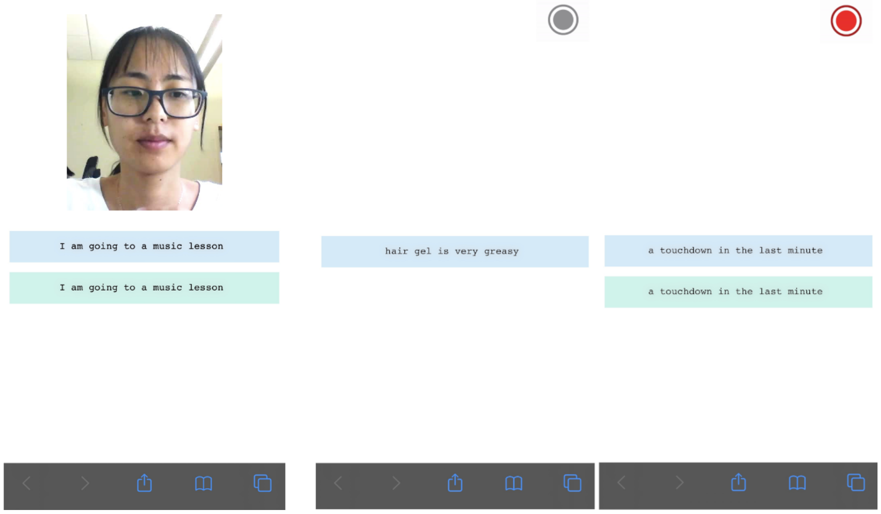
4.4.1 Apparatus

We used the same client/server architecture as the last study (see 4.1), but with an updated user interface (Fig. 4.4). Further, we hosted the app on GitHub[2] to enable people outside the campus network access the client. Six participants used Apple iOS-based smartphones, while the remaining six used Android-based smartphones. Ten of them used a Google Chrome mobile web browser (> v84), while the remaining two used a Safari browser (> v85) to access the client app. The wizard used a Microsoft Surface Book 3 (34.3 cm display, i7 CPU at 1.90GHz, 16GB RAM) to launch the server interface on a Google Chrome web browser (v85.0.4183.102). We did not record any network dropouts during the study.

4.4.2 Feedback Methods

We implement the following two types of visual feedback:

[2] GitHub Pages, https://pages.github.com



I am going to a music lesson

I am going to a music lesson

hair gel is very greasy

a touchdown in the last minute

a touchdown in the last minute

Figure 4.4: The two visual feedback methods used in the study: (1) video feedback that always displays the video captured by the device's front-facing camera on the screen (left) and (2) abstract feedback that displays a grey or a blinking red dot at the top right corner of the device based on whether the camera can see the lips or not, respectively (right).

- **Abstract feedback.** The abstract feedback method is designed to provide minimal feedback on silent speech input. For this, we used a grey dot at the top right corner of the device that turns red and starts blinking when the system tracks the lips (similar to the video recording button on most mobile device). The dot turns grey and stops blinking when the device is unable to see the lips. We use this feedback as it offers a higher level of privacy (does not show users' face or lips) and use minimum screen space on the device.

- **Video feedback.** The video feedback method provides detailed information about users' lip by showing the video captured by the device's front-facing camera. We place the video on the screen as constant feedback to users about the systems status. Though this form of feedback provides precise information on whether the camera can see users' lips, it consumes a considerable portion of the screen real-estate.

4.4.3 Participants

Twelve participants (6 female, 6 male) aged 23 to 34 years (M = 28.75, SD = 2.89) participated in this study. All the participants reported being right-handed, using smartphones for the last 8.58 years (SD = 2.29), and using at least one voice assistant system for 2.26 years (SD = 2.24). None of the participants had prior experience using silent speech input. Note that none of the participants participated in the previous studies.

4.4.4 Error Injection

We injected errors in this study for two reasons. First, to increase the validity of the study since none of the current recognition systems are 100% accurate. Besides, a fully accurate system would have altered some participants about the Wizard-of-Oz setup. Second, to investigate whether users perceive the frequency in which errors occur differently with different feedback methods. For error injection, we used the same approach as the previous study (see 4.1). However, here we maintained a constant 5% error rate over all sessions and injected tracking error rather than misrecognition error. The 5% error rate was chosen as it was found to be an acceptable error rate in various text entry system [20, 22, 11]. A tracking error occurs when the system fails to track the lips because they are out of sight or range, or due to technical issues, resulting in missing words in the final text, for example, "take it to the recycling depot" ("recycling" is removed). We injected tracking error since the purpose of visual feedback on a recognition system is usually to inform users that it is receiving the tracking signals. Hence, tracking error is more appropriate to evaluate the efficiency of visual feedback than misrecognition error.

4.4.5 Design

The study used a within-subjects design. The independent variables was *feedback* and the dependent variables were the qualitative metrics. In summary, the design was:

12 participants ×
2 feedback methods (video and abstract, counterbalanced) ×
30 phrases from MacKenzie & Soukoreff set [188] with 5% injected error
= 720 phrases, in total.

4.4.6 Procedure

The study was conducted remotely due to the spread of COVID-19. We scheduled a video call with each participant ahead of time. They were told that the purpose of the study was to evaluate two different types of visual feedback on a working silent speech recognizer. They were instructed to join the call from a quiet room to avoid any interference during the study. A researcher (the wizard) demonstrated the system and the feedback methods, explained tracking error (that the inability to track the lips results in missing words in the recognized phrase), collected their consents and demographics, and provided all instructions via the video call. The researcher provided the participants with a link to the client app, which they accessed on their smartphone using their preferred web browser. They were instructed to activate the airplane mode but keep the Wi-Fi enabled to avoid any interruptions due to incoming calls. The system displayed one phrase at a time. Participants were asked to silently speak the phrase then tap on the screen to see the recognition and the next phrase. The researcher displayed the recognized phrase and updated the presented phrase using the server interface. We did not instruct the participants on how to hold the device but informed them that the blinking red dot will turn grey when the system cannot track the lips during the graphical feedback condition. The researcher observed all interactions with the smartphone to manually turn the blinking red dot to grey when the front-facing camera is unlikely to capture the lips due to the holding posture or angle. Error correction was not required in this study. Upon completion of the study, participants completed a short questionnaire that asked them to rate various aspect of the two feedback methods on a 5-point Likert scale. We then held a debrief session to inform the participants about the actual nature of the study. The complete study session was recorded using a screen recorder.

4.5 Results

We used non-parametric analyses on the data, thus report median values. We also report the effect size r for the Wilcoxon signed-rank test.

A Wilcoxon signed-rank test identified a significant effect of feedback on whether the method provides enough details about lip detection ($z = -2.06, p < .05, r = 0.6$), occludes, interrupts, and interferes with the task at hand ($z = -2.84, p < .01, , r = 0.82$), and compromise privacy and security ($z = -2.41, p < .05, r = 0.7$). However, there was no significant effect on effectiveness ($z = -0.30, p = .76, r = 0.09$), perceived speed ($z = -1.34, p = .18, r = 0.39$), perceived accuracy ($z = -0.71, p = .48, r = 0.2$), or the overall preference ($z = -1.56, p = .12, r = 0.45$). Fig. 4.5 illustrates median ratings of all aspects of the two feedback methods.

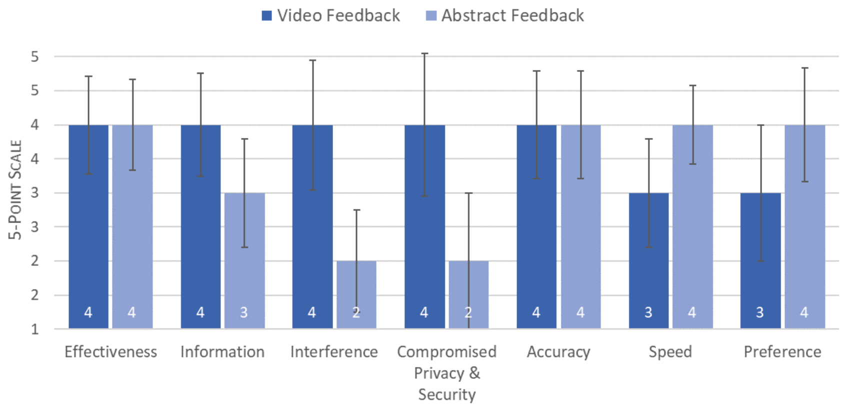


Figure 4.5: Median ratings of various aspects of the two feedback methods on a 5-point Likert scale, where where 1 to 5 represented Strongly disagree to Strongly agree. The error bars represent ±1 standard deviation (SD).

4.6 Discussion

Participants found both feedback methods equally effective. They found the video feedback significantly more informative than abstract feedback. This is not surprising since video feedback displayed a real-time video captured by the device's front-facing camera. Interestingly, participants found the abstract feedback to be the least intrusive (does not occlude, interrupt, or interfere with the task at hand) and most private and secure (does not compromise the user's privacy and security). Once participant (female, 31 years) commented, *"I have privacy concerns with video feedback, I don't want to see my phone camera on when using apps all the time"*. Another participant (male, 27 years) wrote, *"In my opinion, the video feedback mode will always gonna be a concern for my privacy and security"*. In terms of willingness to use, participants were slightly leaning towards the abstract feedback, but this difference was not statistically significant (medium effect size). This is not necessarily a bad thing since it can be interpreted as, users are impartial about the methods, thus using an abstract feedback method is an acceptable design choice. Participants found both methods to be equally reliable (did not compromise accuracy), but interestingly they felt the system with video feedback was slower (statistically not significant) although both used the same Wizard-of-Oz setup. We speculate this is because participants were looking at the video while speaking, which increased the mental demand due to information processing, giving them the impression that it was slower. One limitation of these findings is the lack of generalizability in terms of personality, culture, and ethnic background. Although, the study questionnaire used questions from the SUS questionnaire [45] and custom questions prepared following the [78] guideline, they were not formally validated for the effects of personality, culture, and ethnic background.

Our general intuition may provide initial guidance regarding speech and silent speech input that the latter is likely to be more acceptable than the former due to the nature of the method (it is subtle and less visible). However, without empirical data, it is difficult to come to a conclusion as users' perception towards using the method might be influenced by various factors, such as where they are using the method, in front of whom they are using it, and their acceptance towards the errors committed by the methods. The study results confirm that silent speech input is more socially acceptable as it is subtle, more secure, and less attention-seeking than speech input. Moreover, our results affirm that users are willing to accept more recognition errors with silent speech input than speech input. This is primarily due to the fact that the method is more private, secure, and does not trigger feelings of discomfort. Consequently, users expressed their intention to use the method even with a higher rate of errors than speech input. However, they also showed their preference in limiting the error rate within a reasonable threshold (e.g., 5–10%) for both input methods. We also observed that there is a possible linkage between perceived privacy and security and feedback design for silent speech input. Though video feedback provides users with detailed information (e.g., whether lip movements are captured by the camera), participants expressed their concerns about using this feedback method as it may operate in an always-on manner, continually tracking and analyzing lip movements from the camera. These results further confirm users' strong intention to ensure a high level of privacy and security while inputting on mobile devices.

It is important to note that, while the results are promising, the studies were conducted with Wizard-of-Oz mimicking a mobile silent speech input method. Hence, we were unable to study other technical factors (e.g., silent speech processing delay) that could have affected users' willingness to use the method. Therefore, in the next Chapter, we examine the technical aspects of silent speech input.

Chapter 5

LipType: A Silent Speech Recognizer

This chapter presents the development of LipType, an optimized version of LipNet [25] for improved speed and accuracy. Since developing a new system and acquiring new datasets require an enormous amount of time, effort, and other resources, in this work we exploited a state-of-the-art silent speech recognizer, LipNet [25]. Based on preliminary investigations, we found out that LipNet and other existing recognizers have substantially slower response time and are erroneous due to their architecture. To address these, we develop LipType, an optimized version of LipNet for improved speed and accuracy. LipType demonstrated significant improvements in the performance of LipNet with a 47% reduction in word error rate and 8.6 seconds reduction in computation time.

5.1 An Optimized LipNet Model

We used LipNet as the backbone model based on a study comparing LipNet [25], LCANet [301], Transformer [5], and WAS [60] models. The former two are trained on GRID dataset [66], the latter two on LRS dataset [5]. In an evaluation with 50 random videos from the respective datasets, LipNet and LCANet yielded similar WER (4%), while Transformer and WAS were more error-prone (> 49% WER). Of the two best performed models, we picked LipNet as it is more widely used than LCANet.

LipNet [25] is an existing end-to-end sentence-level model that maps a variable-length sequence of video frames to text, making use of a deep 3-dimensional convolutional neural network (3D-CNN) [141], a recurrent network, and the connectionist temporal classification loss. The model was trained on GRID dataset comprising of highly constrained vocabulary. Although LipNet has proven to be promising, it has several limitations. First, LipNet is focused on capturing spatial and temporal information using deep 3D-CNN that neglects the hidden information between channel correlations in spatial and temporal directions [77], limiting the performance of the architecture. Further, the use of a deep 3D-CNN unnecessarily increases computational complexity and memory intensiveness. We address these issues in LipType, an optimized version of LipNet for improved speed and accuracy.

In LipType, we combined a shallow 3D-CNN (1-layer) and a deep 2D-CNN (34-layer ResNet [117]) integrated with squeeze and excitation (SE) [128] blocks (SE-ResNet) to capture both spatial and temporal information. We used this hybrid-CNN model to address the limitations of 3D-CNN that it neglects the information between channel correlations and increases computational complexity, as well as 2D-CNN's inability to capture temporal information. SE-ResNet adaptively recalibrates channel-wise feature responses by explicitly modelling inter-dependencies between the channels to improve the quality of feature representations. Moreover, it is computationally lightweight and imposes only a slight increase in model complexity and computational burden [128]. Thus, we hypothesize that the proposed hybrid frontend module will reduce the overall computational complexity of LipNet and improve its performance.

5.1.1 The Network

The LipType network consists of two sub-modules (or sub-networks): a *spatiotemporal feature extraction* frontend that takes a sequence of video frames and outputs one feature vector per frame and a *sequence modeling* module that inputs the sequence of per-frame feature vectors and outputs a sentence character by character, as shown in Fig. 5.1. We describe these modules in the following sections.

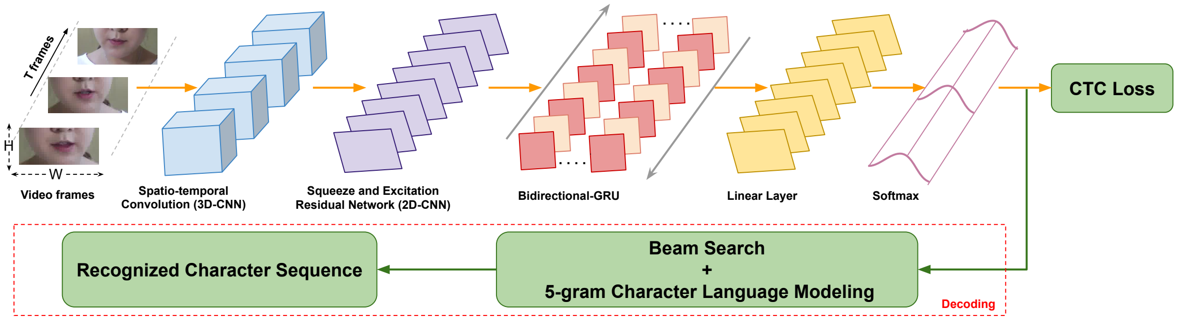

T frames
H
W
Video frames
Spatio-temporal Convolution (3D-CNN)
Squeeze and Excitation Residual Network (2D-CNN)
Bidirectional-GRU
Linear Layer
Softmax
CTC Loss
Recognized Character Sequence
Beam Search
+
5-gram Character Language Modeling
Decoding

Figure 5.1: Architecture of LipType: a sequence of T frames is fed to a 1-layer 3D CNN, followed by 34-layer 2D SE-ResNet for spatiotemporal feature extraction. The extracted features are processed by two Bi-GRUs, followed by a linear layer and a softmax. The network is trained entirely end-to-end with CTC loss.

5.1.1.1 Spatiotemporal Feature Extraction

It starts with the extraction of a mouth-centred cropped image of size H:100 × W:50 pixels per video frame. For this, videos are first pre-processed using DLib face detector [157] and the iBug face landmark predictor [251] with 68 facial landmarks combined with Kalman Filtering.

Then, a mouth-centred cropped image is extracted by applying affine transformations. The sequence of T mouth-cropped frames are then passed to 3D-CNN, with a kernel dimension of T:3× W:5 × H:5, followed by Batch Normalization (BN) [133] and Rectified Linear Units (ReLU) [6]. The extracted feature maps are then passed through 34-layer 2D SE-ResNet that gradually decreases the spatial dimensions with depth, until the feature becomes a single dimensional tensor per time step.

5.1.1.2 Sequence Modeling

The extracted features are processed by 2-Bidirectional Gated Recurrent Units (Bi-GRUs) [61]. Each time-step of the GRU output is processed by a linear layer, followed by a softmax layer over the vocabulary, then an end-to-end model is trained with connectionist temporal classification (CTC) loss [107]. The softmax output is decoded with a left-to-right beam search [64] using Stanford-CTC's decoder [183] and 5-gram character language model [105] to recognize the spoken utterances. The model is capable of mapping variable-length video sequences to text sequences.

5.1.2 Experiment

We conducted an experiment to compare the performance of LipNet and LipType.

5.1.2.1 Dataset

For a fair comparison between the two models, we trained the LipType model on the same GRID dataset [66] on which the LipNet model was trained. It comprises of short and formulaic video clips of a person's face when uttering a highly constrained vocabulary in a specific order ($N = 34$). Similar to a previous experiment investigating the performance of LipNet with overlapped speakers [25], this experiment used 21,635 videos for training and 7,140 videos for evaluation.

5.1.2.2 Implementation

To avoid any potential confounding factor, we trained both models from scratch with the same training parameters. The number of frames was fixed to 75. Longer image sequences were truncated and shorter sequences were padded with zeros. We applied a channel-wise dropout [273] of 0.5. The model was trained end-to-end by the Adam optimizer [159] for 60 epochs with a batch size of 50. The learning rate was set to 10^{-4}. The network was implemented based on the Keras deep-learning platform with TensorFlow [1] as the backend. We trained and tested both models on NVIDIA GeForce 1080Ti GPU board.

5.1.2.3 Performance Metrics

We used the following metrics to benchmark the proposed framework.

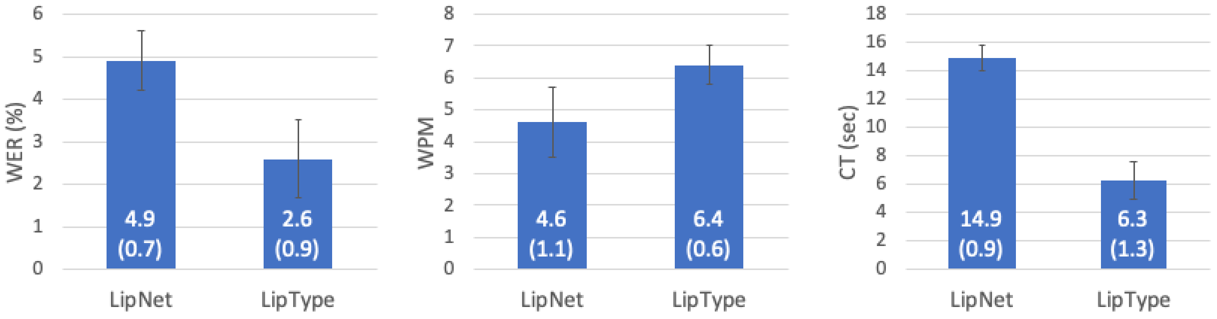
- **Word error rate (WER)** is the minimum number of operations required to transform the predicted text to the ground truth, divided by the number of words in the ground truth. It is calculated using the following equation, where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the ground truth.

$$WER = \frac{S + D + I}{N} \tag{5.1}$$

- **Words per minute (WPM)** is a commonly used text entry metric that signifies the rate in which words (= 5 chars) are entered [19]. It is calculated using the following equation, where T is the number of recognized words, t is the sum of speaking time and computation time in seconds, the constant 60 is the number of seconds per minute, and the factor of one fifth accounts for the average length of a word in the English language.

$$WPM = \frac{|T| - 1}{t} \times 60 \times \frac{1}{5} \tag{5.2}$$

- **Computation time (CT)** is the total time required by the model to predict a phrase. It does not include the time users take to speak a phrase.



(a) (b) (c)

Figure 5.2: Performance comparison of LipNet and LipType in terms of a) word error rate, b) words per minute, and c) computation time. Reported values are the average of all values. Values inside the brackets are standard deviations (SD). Error bars represent ±1 standard deviation.

5.2 Results

In the experiment, LipType outperformed LipNet in terms of input speed, accuracy, and computation time. LipType achieved 2.6% WER, 6.4 WPM, and 6.3 seconds CT (Fig. 5.2).

In comparison with LipNet, it exhibited a 47% reduction in WER, 39% increase in WPM and 8.6 seconds reduction in CT. These findings confirm our intuition that extracting spatiotemporal features using the hybrid of a shallow 3D-CNN and a deep 2D-CNN integrated with SE blocks, instead of only 3D-CNN, will reduce the overall computational complexity and improve performance.

5.3 Discussion

We developed LipType, an optimized version of LipNet for improved speed and accuracy. LipType demonstrated a signifcant improvement in the performance of LipNet. Results revealed 47% reduction in WER, 39% increase in WPM, and 8.6 seconds reduction in CT. Despite these improvements, LipType and other silent speech recognition models remain unreliable in real-world settings. These models do not account for various extraneous factors such as uncontrolled lighting, blur, low resolution, compression artifacts, occlusions, viewing angles, etc. However, most of the factors can be mitigated by replacing the hardware (blur, low-resolution, compression artifacts, etc.) or by the user (occlusions, viewing angles, etc.). Lighting, in contrast, is one of the factors that cannot always be controlled. Not accounting for this in a vision-based speech recognition compromises its fairness and reduces its applicability in real-world scenarios. Therefore, in the next Chapter, we develop an independent repair model that processes video input for poor lighting conditions, when applicable, and corrects potential errors in output for increased accuracy.

Chapter 6

An Independent Repair Model

This Chapter presents the development of an independent repair model, a multi-stage pipeline compensating for poor lighting conditions and potential recognition errors for increased accuracy of speech and silent speech recognizers. It then presents an empirical demonstration of the repair models' effectiveness on multiple speech and silent speech recognizers. Empirical results showed that it improves accuracy rates for all recognizers without substantially compromising the computation time.

6.1 Repair Model: Light Enhancement and Error Reduction

We present a new repair model, a multi-stage pipeline that accounts for poor lighting conditions in input videos and potential errors in the recognition. It includes a *pre-processing* step to enhance videos with poor lighting conditions and a *post-processing* step to automatically detect and correct potential errors generated by the recognizer. A key consideration for this model was its independence, to make sure it is not reliant on a specific recognizer so that it can be used with a variety of speech and silent speech recognition models.

6.1.1 Light Enhancement

There are various factors that can affect the performance of silent speech recognition, for example, uncontrolled lighting, blur, low-resolution, compression artifacts, occlusions, viewing angles, pace of speech, etc. However, most of the factors can be mitigated by replacing the hardware (blur, low-resolution, compression artifacts, etc.) or by the user (occlusions, viewing angles, accent, pace of speech, etc.). Lighting, in contrast, is one the factor that cannot always be controlled.

Making recognition more reliable under uncontrolled lighting conditions is one of the major challenges for practical silent speech recognition models. Existing models do not account for lighting variations, making them unreliable in poorly lit places. We tackle this by adding a pre-processing step to the LipType [See Chapter 5] recognition model. For this, we improved GLADNet [294], a low-light image enhancement network, and adapted it for enhancing input videos. We used GLADNet because it demonstrated a much better performance with actual under-exposed images compared to the other models, both in terms of quality [143, 110, 92, 79] and computation complexity [299, 4, 173, 52].

6.1.1.1 The Network

The light enhancement network learns an end-to-end mapping from low-light images to normal-light images. It processes videos in a frame-by-frame manner, as illustrated in Fig. 6.1. The architecture of the network comprises of two adjacent steps: the first is for *global illumination estimation* and the second is for *detail reconstruction*.

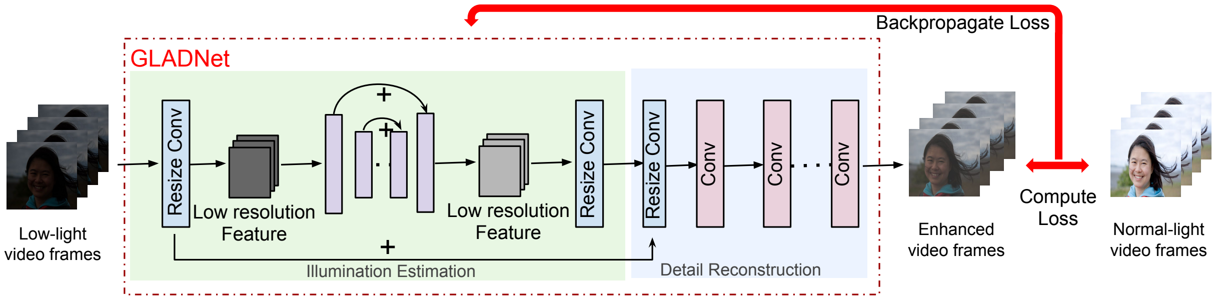


Figure 6.1: Architecture of the pre-processing (light enhancement) network: a sequence of low-light images is fed through the network where the enhanced images are compared with the normal-light images to compute the loss, which is then backpropagated to fine-tune and optimize the model weights and biases.

In the global illumination estimation step, input is down-sampled to a fixed size feature map using nearest-neighbor interpolation. Then, it is passed through an encoder-decoder network[1] to estimate the global illumination of the input. The estimated feature maps are then re-scaled to the original size using a resize convolution block. Then, the re-scaled feature maps are passed to the detail reconstruction step comprising of three convolutional layers. This step adjusts the illumination of the input image by assembling predicted global illumination and input image information, and fills in the details lost during the down- and up-sampling processes. Inspired by a previous work [309], we investigated the consequences of replacing the L1 loss function used in the training of GLADNet with alternative loss functions. Given a collection of N training sample pairs X_i , Y_i, where X_i is low-light input image and Y_i is normal-light ground truth image, the following loss functions can be defined.

[1] In order to reduce computation, we changed the GLADNet network dimension from five down- and five up-sampling blocks to three down- and three up-sampling blocks. A preliminary investigation did not identify a significant effect on variations in layer dimensions on the network's performance.

1. **L1 Loss** (or mean-absolute-error loss) minimizes the sum of the absolute differences between the predicted or generated image and the ground truth.

$$L1(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i) \tag{6.1}$$

2. **L2 Loss** (or mean-squared-error loss) minimizes the sum of the squared differences between the predicted or generated image and the ground truth.

$$L2(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2 \tag{6.2}$$

3. **Multi-scale structural similarity loss** (MSSSIM) [309] minimizes the loss related to the sum of structural-similarity scores across all image pixels, in terms of luminance, contrast, and structure.

$$MSSSIM(X,Y) = -\sum_{i=1}^{N} MSSSIM(X_i - Y_i) \tag{6.3}$$

4. **MSSSIM-L1 loss** captures MSSSIM's ability to preserve the contrast in high-frequency regions and L1's ability to preserves colors and luminance. In the equation below, G is the Gaussian filter, α is the weighting factor to roughly balance the contribution of the two losses. We empirically set $\alpha = 0.81$[2].

$$\text{MSSSIM-L1}(X,Y) = \alpha \cdot MSSSIM + (1-\alpha) \cdot G_\sigma \cdot L1 \tag{6.4}$$

5. **MSSSIM-L2 loss** captures MSSSIM's ability to preserve the contrast in high-frequency regions and L2's ability to remove noise and ringing artifacts. Like MSSSIM-L1, $\alpha = 0.81$ and G is the Gaussian filter.

$$\text{MSSSIM-L2}(X,Y) = \alpha \cdot MSSSIM + (1-\alpha) \cdot G_\sigma \cdot L2 \tag{6.5}$$

6.1.2 Experiment: Light Enhancement Network

We evaluated the performance of the light enhancement network trained with the above five loss functions.

6.1.2.1 Dataset

We trained and validated the network on the GLADNet dataset [294] that comprises of 5,000 image pairs of low and normal light images. We used 4,000 pairs for training and the remaining 1,000 pairs for testing.

[2] In an investigation, results were not affected by small variations in α.

6.1.2.2 Performance Metrics

We used the following two standard image quality metrics [124].

- **Peak signal to noise ratio (PSNR)** computes the peak signal-to-noise ratio between two images in decibels. This ratio is used as a quality measurement between the original and an enhanced image. The higher the PSNR the better the quality of the enhanced image.

- **Structural similarity metric (SSIM)** measures the perceptual difference between two similar images. Unlike PSNR, SSIM is based on visible structures in the image. The lower the SSIM the better the quality of the enhanced image.

6.1.2.3 Implementation

We trained the network for 70 epochs with a batch size of 32. It was optimized using Adam [159]. The learning rate was set to 10^{-3}. The network was implemented on the Keras deep-learning platform with TensorFlow [1] as the backend. We trained and tested the network on NVIDIA GeForce 1080Ti GPU board.

6.1.3 Results

Table 6.1 presents the performance comparison of GLADNet trained on the aforementioned five loss functions in terms of averaged PSNR and SSIM. It can be seen that MSSSIM-L1 achieved the highest PSNR and outperformed other loss functions substantially in the SSIM measure. Therefore, we used GLADNet trained with MSSSIM-L1 loss function to enhance poor lighting input videos for more reliable silent speech recognition.

Metric	**Low-Light**	**Enhanced**				
		Loss Function				
		L1	*L2*	*MSSSIM*	***MSSSIM-L1***	*MSSSIM-L2*
PSNR	19.74	26.22	25.66	26.11	**27.34**	26.13
SSIM	0.46	0.7822	0.7574	0.7890	**0.8091**	0.7911

Table 6.1: Averaged peak signal to noise ratio (PSNR) and structural similarity metric (SSIM) for the five investigated loss functions. For MSSSIM, the reported values are obtained as averages of the three color channels (RGB). The best results are highlighted in bold.

6.1.4 Post-Processing: Error Reduction

This section presents a new algorithm for predicting and automatically correcting potential recognition errors by a speech or silent speech recognizer. It comprises of two sub-modules: an *error minimization* module that corrects potential errors in the recognized character sequence using deep denoising autoencoder (DDA) [288] and a *sequence decoder* module that converts corrected character sequence to meaningful word sequences using spell-checker and a custom language model. The architecture of the network is illustrated in Fig. 6.2.

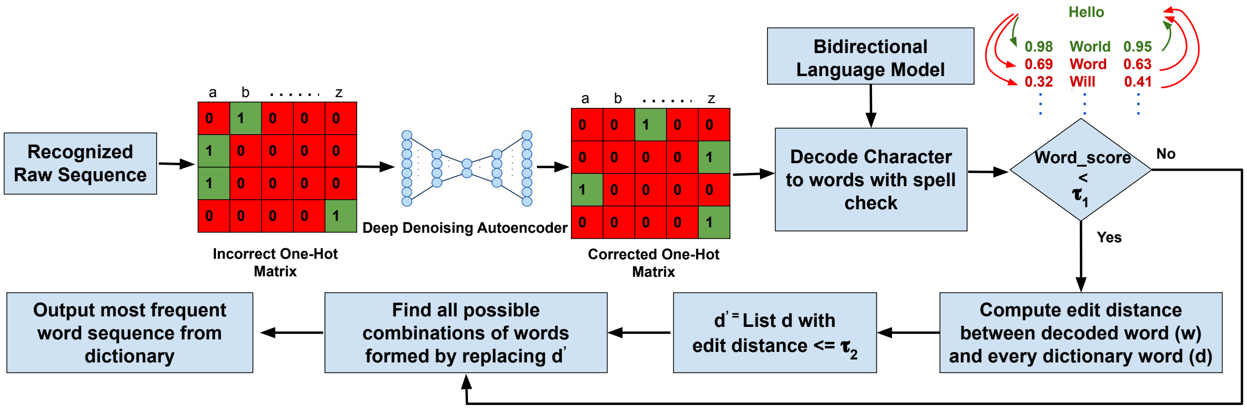

Recognized Raw Sequence
a b ...... z
Incorrect One-Hot Matrix
Deep Denoising Autoencoder
Corrected One-Hot Matrix
Bidirectional Language Model
Hello
0.98 World 0.95
0.69 Word 0.63
0.32 Will 0.41
Decode Character to words with spell check
Word_score < τ1
No
Yes
Compute edit distance between decoded word (w) and every dictionary word (d)
d' = List d with edit distance <= τ2
Find all possible combinations of words formed by replacing d'
Output most frequent word sequence from dictionary

Figure 6.2: Architecture of post-processing (error reduction) network: the predicted raw sequence is fed to DDA, followed by spell checker and a custom language model.

6.1.4.1 Error Reduction

DDA has been successful in the context of reconstructing a noisy signal [89, 178]. In this work, we used DDA to correct the character sequence predicted by the recognizer. The predicted sequence is represented in the form of a matrix, where each row is a one-hot[3] encoded vector, pointing to a particular character out of all. An input to autoencoder is converted to a fixed length sequence: 28 in this case (26 letters of the English alphabet, 1 space character, and 1 newline character), either by subdividing the sequence or by appending zero vectors, depending on the length of the sequence. This fixed length matricized sequence is fed-forwarded through a DDA to obtain an improved character sequence. The DDA is trained with the matricized incorrect character sequence as input and the matricized correct sequence as the labels. This helped in reconstructing the sequence, thus reducing the errors. In order to quantify the errors between incorrect sequence and the ground truth, we used cross-entropy loss [308], which is given by the following equation, where x represents

[3] Encodes categorical data using a one-of-K scheme.

the matricized incorrect character sequence and z represents the matricized ground truth sequence.

$$Loss(x,z) = -\sum_{k=1}^{d}[x_k log z_k + (1-x_k)log(1-z_k)] \tag{6.6}$$

6.1.4.2 Sequence Decoder

The corrected character sequence embedded with the space and newline characters is first combined to form a sequence of words. The resultant word sequence is then passed to the spell checker[4] to be checked for spelling correctness for auto-correction, when necessary. In addtion, a language model (LM) was used to get the most probable sequence of words. We used a traditional count-based LM[5]. Typically, n-gram analysis in count-based LM is a forward n-gram. However, we explored and evaluated the advantage of a bidirectional n-gram modeling that accounts for both forward and backward directions. Formally, we consider a string of n words, $W = w_1, w_2, ..., w_n$. In a forward n-gram, the probability of each word is estimated depending on the preceding words:

$$\begin{aligned} P_{forward}(W) =& P(w_1| < start >) * P(w_2|w_1)* \\ & P(w_3|w_2) * ... * P(< end > |w_n) \end{aligned} \tag{6.7}$$

In contrast, in a backward n-gram the probability of each word is estimated depending on the succeeding words:

$$\begin{aligned} P_{backward}(W) =& P(< start > |w_1) * P(w_1|w_2)* \\ & P(w_2|w_3) * ... * P(w_n| < end >) \end{aligned} \tag{6.8}$$

The combined probability of a sentence, thus, is computed by multiplying the forward and backward n-gram probability of each word:

$$\begin{aligned} P_{combined}(W) =& (P_{forward}(W_1) * P_{backward}(W_1))* \\ & (P_{forward}(W_2) * P_{backward}(W_2))* \\ & ...* \\ & (P_{forward}(W_n) * P_{backward}(W_n)) \end{aligned} \tag{6.9}$$

Applying the values from Equations 6.7 and 6.8, we get:

$$\begin{aligned} P_{combined}(W) =& (P(w_1| < start >) * P(< start > |w_1))* \\ & (P(w_2|w_1) * P(w_1|w_2))* \\ & (P(w_3|w_2) * P(w_2|w_3))* \\ & ...* \\ & (P(< end > |w_n) * (w_n| < end >)) \end{aligned} \tag{6.10}$$

[4] How to write a spelling corrector, http://norvig.com/spell-correct.html

[5] A count-based LM follows the general idea of making n^{th} order Markov assumptions and calculating the n-gram probabilities through the means of counting.

Finally, the network predicts and corrects potential errors committed by the language model in the following three steps. (1) Compare the combined probability of each word, $P_{combined}w_n = P(w_n|w_{n-1}) * P(w_{n-1}|w_n)$ (Equation 6.10), with a pre-defined threshold τ_1. If $P_{combined}w_n$ is less than τ_1, the word is considered erroneous. (2) Compute edit distance (ED) between an erroneous word w_n and each dictionary word d to create a list d' of all dictionary words that have an ED less than a predefined threshold τ_2. (3) Replace each word in d' with $P_{combined}w_n$ in a sentence and output the most frequent word sequence from the dictionary.

We conducted an extensive study to select the best combinations of τ_1 and τ_2 by analyzing the performance of the proposed LM in the defined context.

6.1.5 Experiment: Error Reduction Model

We evaluated each sub-module of the post-processing step. First, we evaluated the architecture for the DDA network. Second, we evaluated the performance of the proposed LM. Finally, we identified the best thresholds values for computing numerical similarities.

6.1.5.1 Dataset

We used LIBRISPEECH LM corpus [218] to train and evaluate the post-processing modules. The dataset contains text from 14,500 public domain books. We first filtered out all punctuation, casing, and non-alphanumeric tokens from the original text and extracted the top 200,000 sentences as vocabulary.

6.1.5.2 Training and Evaluation of Various DDA Architectures

For training DDA, we randomly divided the dataset into 100,000 sentences as correct set and remaining 100,000 as incorrect set. We then synthesized one character-level error to each word of each phrase. To synthesize errors, we simulated the following four types of error to each word in the following sequence: one deletion error (removal of one letter), one transposition error (swapping of two adjacent letters), one replacement error (changing one letter with another), and one insertion error (one additional letter). Table 6.2 presents the statistics of the dataset used for training DDA. It was divided into a split of 80:20% as training:testing set.

To select the best network architecture for DDA, we trained and evaluated four different architectures (Table 6.3). All networks were implemented on the Keras deep-learning platform with TensorFlow [1] as the backend and an NVIDIA GeForce 1080Ti as the GPU board. We used Adam [159] as the optimization method for training. We trained the networks for 50 epochs with learning rate of 10^{-3}, batch size of 128. Results revealed that the DDA architecture with 5-layers having [128 64 32 64 128] nodes performed the best (Table 6.3). Hence, we used the DDA trained with this architecture to minimize potential errors in the recognized output.

Phrase	**Word**	**Char**	**C_Word**	**C_Char**	**I_Word**	**I_Char**
200,000	6,027,754	18,527,816	2,906,117	9,279,253	3,121,637	9,248,563

Table 6.2: Statistics of dataset used for training DDA. C_Word, C_Char, I_Word, and I_Char stands for number of correct words, correct characters, incorrect words, and incorrect characters respectively. The values are the total.

DDA Architecture	**WER**
Number of Layers: [Number of Nodes]	Mean (%)
5: [256 128 64 128 256]	21.8
5: [128 64 32 64 128]	**16.4**
3: [128 64 128]	19.1
3: [64 32 64]	26.3

Table 6.3: Evaluation of various DDA architectures in terms of word error rate (WER).

6.1.5.3 Training and Evaluation of N-Gram Language Model

We evaluated the directional advantage of a count-based n-gram LM with state-of-the-art bi-directional neural LM in terms of sentence error rate (SER)[6], perplexity[7], and computation time. For a fair comparison, we trained both models from scratch using the LIBRISPEECH dataset (Section 6.1.5.1). We divided the dataset in a split of 80:20% as training: testing set. Count-based n-grams models were trained using the Natural Language Toolkit (NLTK)[8] with Kneser-Ney smoothing [119, 53] to better estimate probabilities of unseen n-grams. Bi-directional neural LM (Bi-LSTM) was trained using LSTM based recurrent units that have two recurrent layers with 4,096 LSTM nodes in each layer, an input projection layer of size 128, and an output softmax layer over vocabulary. The model was trained end-to-end using cross-entropy loss [308] with Adam [159] as the optimization method. The model was trained for 60 epochs with batch size of 64 and learning rate of $1e^{-3}$. It was implemented based on the Keras deep-learning platform with TensorFlow [1] as the backend. Both LMs were trained and tested on NVIDIA GeForce 1080Ti GPU.

In the experiment, Bi-LSTM performed better than the count-based LMs in terms of SER and perplexity (Table 6.4). However, it required extra computation time. Among count-based LMs, the combined trigram LM (forward and backward) performed much better.

[6] Sentence error rate (SER) signifies the percentage of recognized sentences that are not an exact match of the ground truth.

[7] Perplexity is the multiplicative inverse of the probability assigned to the sentence by the language model, normalized by the number of words in the sentence. The lower the perplexity the better the language model.

[8] Natural Language Toolkit (NLTK), https://www.nltk.org/api/nltk.lm.html

Sentence Error Rate (SER) %

Bigram			Trigram			Bi-LSTM
Forward	Backward	Combined	Forward	Backward	**Combined**	
27.4	30.9	26.7	24.4	27.6	**16.5**	15.3

Perplexity

Bigram			Trigram			Bi-LSTM
Forward	Backward	Combined	Forward	Backward	**Combined**	
51.3	60.1	48.7	44.1	48.3	**41.4**	39.8

Computation Time (Second)

Bigram			Trigram			Bi-LSTM
Forward	Backward	Combined	Forward	Backward	**Combined**	
1.8	1.7	3.1	1.5	1.9	**3.4**	9.2

Table 6.4: Comparison between forward, backward and combination of both (forward + backward) n-gram LM with Bi-LSTM LM. Reported sentence error rate (SER), perplexity, and computation time are average of all values. The proposed repair model uses the combined trigram model.

Besides, it yielded a 7.27% and 3.86% higher SER and perplexity, respectively, and a 5.8 seconds ($\sim$ 170.5%) lower computation time than Bi-LSTM. Hence, considering the negligible percentage differences in SER and perplexity and a large difference in computation time, we decided to use the combination of forward and backward trigram LM in our repair model.

6.1.5.4 Selection of Best Combinations of τ_1 and τ_2 to Compute Numerical Similarity

To select the best combinations of τ_1 and τ_2, we evaluated the proposed LM for various combinations of τ_1 and τ_2, in terms of true positive rate (TPR) and false positive rate (FPR), defined as:

$$TPR = \frac{TP}{TP + FN} \quad \text{and} \quad FPR = \frac{FP}{FP + TN} \tag{6.11}$$

TP: True positive is the total number of correct words identified as correct.
FP: False positive is the total number of incorrect words identified as correct.
TN: True negative is the total number of incorrect words identified as incorrect.
FN: False negative is the total number of correct words identified as incorrect.

Each curve in Fig. 6.3 signify TPR vs. FPR for different sets of τ_1 and τ_2. It can be clearly seen that the LM with $\tau_1 = 0.7$, $\tau_2 = 2$ performed best among all cases since it has a much higher TPR and a lower FPR.

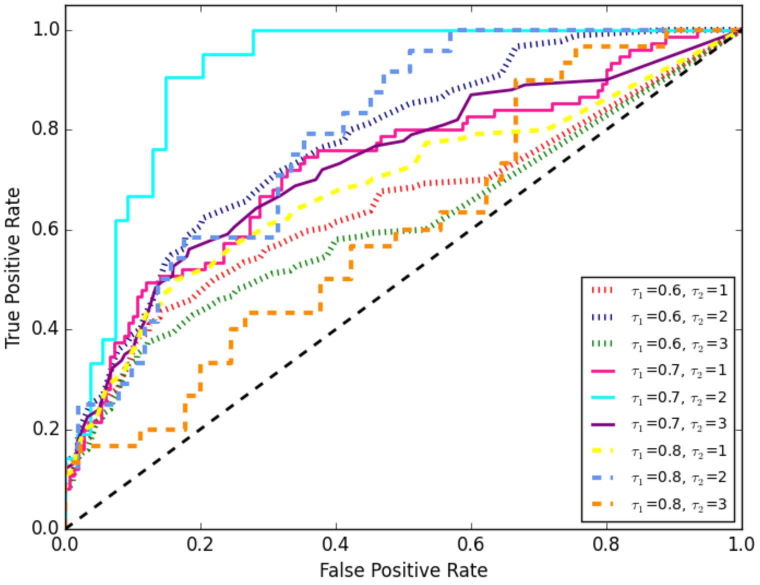


Figure 6.3: Performance comparison in terms of TPR and FPR of proposed LM for various values of τ_1 and τ_2

To summarize, the post-processing step include: 5-layer DDA with bi-directional count-based trigram LM, followed by numerical similarity with $\tau_1 = 0.7$, $\tau_2 = 2$.

6.1.6 Performance Evaluation: Independence of the Model

Since our goal was to develop a repair model that can be used with a range of speech and silent speech recognizers, we evaluated its effectiveness with both LipType [See Chapter 5] and several other popular speech and silent speech recognizers. Particularly, we picked the following six pre-trained models.

6.1.6.1 Silent Speech Recognizers

1. **LipNet** [25] model uses a neural network architecture for lip reading that maps variable-length sequences of video frames to text sequences, making use of deep 3-dimensional convolutions, a recurrent network, and the connectionist temporal classification loss [107], trained entirely end-to-end. It was trained on the GRID dataset [66] which comprises of short and formulaic videos that show a well-lit person's face while uttering a highly constrained vocabulary in a specific order.

2. **LipType** [See Chapter 5] model follows the same architecture as LipNet except it

replaces deep 3-dimensional convolutions with a combination of shallow 3-dimensional convolutions (1-layer) and deep 2-dimensional convolutions (34-layer ResNet) integrated with squeeze and excitation (SE) blocks (SE-ResNet). It was also trained on the GRID dataset.

3. **Transformer** [5] model comprises of two sub-modules: a *spatio-temporal visual frontend* that takes a sequence of video frames to extract one feature vector per frame and a *sequence processing backend* comprised of encoder-decoder structure with multi-head attention layers [286] that generates character probabilities over the vocabulary. It was trained on Lip Reading in the Wild (LRW) [58] and the Lip Reading Sentences 2 (LRS2) [5] datasets.

6.1.6.2 Speech Recognizers

1. **DeepSpeech** [111] is a speech recognition model developed using end-to-end training of a large recurrent neural network (RNN). It converts an input speech spectrogram into a sequence of character probabilities. It was trained on the Wall Street Journal (WSJ) [224], Switchboard [100], and Fisher [62] datasets.
2. **Kaldi** [234] is an open-source toolkit for speech recognition written in C++, which uses Finite State Transducer (OpenFST) library [244] for training recognition models. It comprises of multiple speech recognition recipes. For our work, we used a pre-trained chain English model (Api.ai) recipe, trained on the LIBRISPEECH dataset [218].
3. **Wave2Letter** [65] is an end-to-end model for speech recognition, that combines a convolutional network-based acoustic model and a graph decoding. It is trained to output letters without the need for force aligning them. It was trained on the LIBRISPEECH [218] dataset.

We evaluated these models on seen and unseen data. For seen data, we randomly selected 30 phrases from each model's training dataset. For unseen data, we randomly selected 30 phrases from MacKenzie and Soukoreff dataset [188]. Unseen data was common for all models. All selected phrases are listed in the Appendix A.

6.1.6.3 Experimental Conditions

We evaluated the silent speech models under three lighting conditions. Due to the spread of COVID-19, all conditions were simulated in a private room without any artificial light sources.

- **Dark light**: video recorded during nighttime (9:00–11:00 PM).
- **Dusky light**: video recorded during evening time (6:00–8:00 PM).
- **Daylight**: video recorded during daytime (1:00–3:00 PM).

Likewise, speech models were evaluated under three noisy conditions, simulated in a private room.

- **Indoor noise**: audio recording with an indoor noise, simulated by playing a prerecorded CNN news report in the background.
- **Outdoor noise**: audio recording in a public place, simulated by playing a prerecorded busy marketplace noise.
- **Quiet**: audio recording in a quiet room.

6.1.6.4 Apparatus

We developed a custom Android application with Android Studio 3.1.4 for data collection. The application included a *landing* page and a *data collection* page. The landing page included a drop-down menu to select recording conditions and a Start button to start a session. The data collection page included a video viewer to display the device's front camera, an area to presented phrases, and a Record/Stop toggle button to start and stop recording. The application recorded all videos and automatically logged the duration of a session, device specification (display and camera resolution, etc.), light intensity, and sound level.

6.1.6.5 Participants

Twelve volunteers aged 19—54 years (M = 27.9, SD = 9.15) took part in the study (Fig. 6.4). They were all proficient in the English language. Five of them identified themselves as women and seven identified as men. They all had at least five years of experience with smartphones. All of them were Android-based smartphone users, and users of a voice assistant system for at least one year. Most of them had experience with multiple voice assistants, including Amazon Alexa, Google Assistant, and Apple Siri. They all received US $20 for participating in the study.

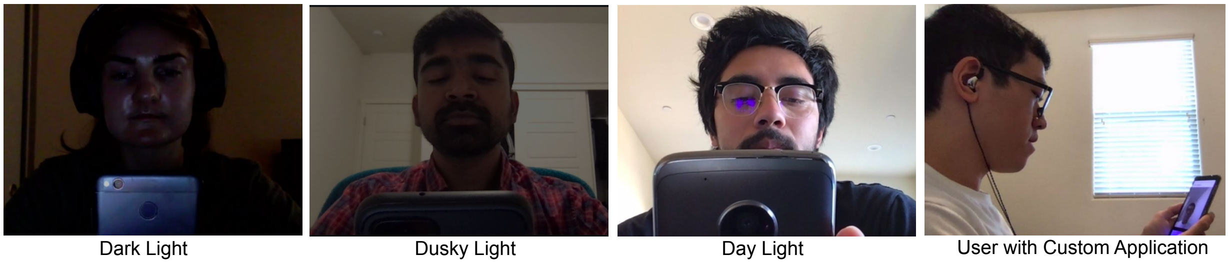


Figure 6.4: Four volunteers participating in the user study.

6.1.6.6 Design

We used the following within-subjects design for the study:

12 participants ×
2 methods (speech, silent speech) ×
3 conditions (indoor, outdoor, quiet / dark, dusky, day), counterbalanced ×
2 data types (seen, unseen) ×
3 models (DeepSpeech, Kaldi, Wave2Letter / LipNet, LipType, Transformer), counterbalanced ×
30 phrases = 12,960 phrases in total.

6.1.6.7 Procedure

The study was conducted remotely due to the spread of COVID-19. We explained the purpose of the study and scheduled individual Zoom[9] video calls with each participant ahead of time. We instructed them to join the call from a quiet room to avoid any interruptions during the study. In the first call, we demonstrated the application and collected their consents and demographics using electronic forms. We then shared the application (APK file) with them and guided them through the installation process on their smartphones. The first session started shortly after that. The application displayed one phrase at a time. Participants pressed the Record button, spoke or silently spoke[10] the phrase, then pressed the Stop button to see the next phrase. In the noisy conditions (Section 6.1.6.3), we shared the respective audio clips with the participants and instructed them to play the clips slightly louder than a normal conversation. Log analysis reveled that, on average, participants played the indoor noise at 48.75 db (min = 42 db, max = 58 db) and outdoor noise at 55.25 db (min = 49 db, max = 66 db). To simulate different lighting conditions, silent speech sessions were scheduled at different times of the day. Log analysis revealed that, on average, room light intensity was 0.93 lux (min = 0 lux, max = 2 lux) in the dark light condition, 7.86 lux (min = 6 lux, max = 11 lux) in the dusky light condition, and 58.0 lux (min = 52 lux, max = 61 lux) in the daylight condition. All sessions followed the same format, expect for demonstration and installation. Upon completion of each session, participants shared the logged data with us by uploading those to a cloud storage under our supervision. In total, there were 24 recording sessions (Table 6.5). A researcher monitored all sessions via Zoom. Upon completion of the study, we evaluated the repair model with the six recognition models using the collected audio and video clips. For speech, first, we passed the recorded audio to a speech recognizer, then we post-processed the output to auto-correct errors. We did not pre-process the data since speech only utilizes audio information, thus, is not affected by poor lighting conditions. For silent speech, first, we processed each recorded video with the pre-processing technique to enhance the lighting of the clips, then we passed the processed

[9] Zoom, https://zoom.us

[10] Uttering phrases without vocalizing any sound

Speech			
Session	**Condition**	**Model**	**Dataset**
1	Indoor	DeepSpeech	Fisher [62] (S)
2	Outdoor	DeepSpeech	Fisher [62] (S)
3	Quiet	DeepSpeech	Fisher [62] (S)
4	Indoor	Kaldi	LIBRISPEECH [218] (S)
5	Outdoor	Kaldi	LIBRISPEECH [218] (S)
6	Quiet	Kaldi	LIBRISPEECH [218] (S)
7	Indoor	Wave2Letter	LIBRISPEECH [218] (S)
8	Outdoor	Wave2Letter	LIBRISPEECH [218] (S)
9	Quiet	Wave2Letter	LIBRISPEECH [218] (S)
10	Indoor	DeepSpeech/Kaldi/Wave2Letter	Mackenzie and Soukoreff [188] (U)
11	Outdoor	DeepSpeech/Kaldi/Wave2Letter	Mackenzie and Soukoreff [188] (U)
12	Quiet	DeepSpeech/Kaldi/Wave2Letter	Mackenzie and Soukoreff [188] (U)
Silent Speech			
13	Dark	LipNet	Grid [66] (S)
14	Dusky	LipNet	Grid [66] (S)
15	Day	LipNet	Grid [66] (S)
16	Dark	LipType	Grid [66] (S)
17	Dusky	LipType	Grid [66] (S)
18	Day	LipType	Grid [66] (S)
19	Dark	Transformer	LRS [5] (S)
20	Dusky	Transformer	LRS [5] (S)
21	Day	Transformer	LRS [5] (S)
22	Dark	LipNet/Transformer/LipType	Mackenzie and Soukoreff [188] (U)
23	Dusky	LipNet/Transformer/LipType	Mackenzie and Soukoreff [188] (U)
24	Day	LipNet/Transformer/LipType	Mackenzie and Soukoreff [188] (U)

Table 6.5: Recording sessions for different noisy and lighting conditions with their corresponding recognition models and datasets. S and U stands for seen and unseen data, respectively.

videos to a silent speech recognizer, finally we post-processed the output to auto-correct errors.

6.1.7 Results

For evaluation, we considered all pre-trained models as baselines and compared with their respective repaired versions in terms of WER, WPM, and CT. To ensure a fair comparison of computation time, we evaluated all models on NVIDIA GeForce 1080Ti GPU board. Results revealed that the proposed repair model significantly reduce error rates of all pre-trained models regardless of data type and experimental conditions.

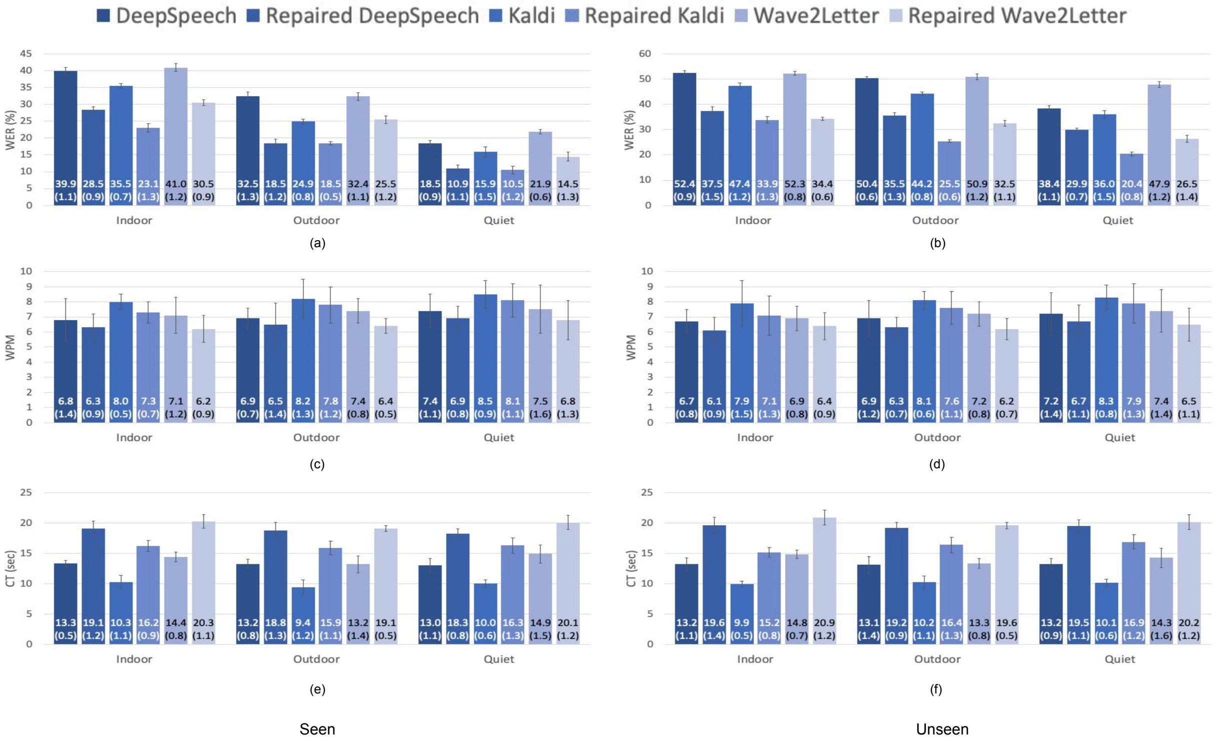


Figure 6.5: Performance evaluation of the three investigated speech recognition models without/with the proposed repair model in terms of a) WER-Seen, b) WER-Unseen, c) WPM-Seen, d) WPM-Unseen, e) CT-Seen, and f) CT-Unseen. Each condition has 360 data points. Reported values are the average of all values. The values inside the brackets are standard deviations (SD). Error bars represent ±1 SD.

Fig. 6.5 shows the effectiveness of repair model on the three examined speech recognition models. It can be clearly observed that the repair model resulted in substantial reductions in error rates for all pre-trained models under all noisy conditions. With DeepSpeech, it showed 37.5% reduction in WER for seen data and 26.7% reduction for unseen data. With

Kaldi, it showed 31.5% reduction in WER for seen data and 38% reduction for unseen data. With Wave2Letter, it showed 26.8% reduction in WER for seen data and 38.3% reduction for unseen data. On average, for all models, we observed 8.4% reduction in WPM and 5.9 seconds increase in CT on both seen and unseen data. Overall, Repaired Kaldi performed the best among all pre-trained models.

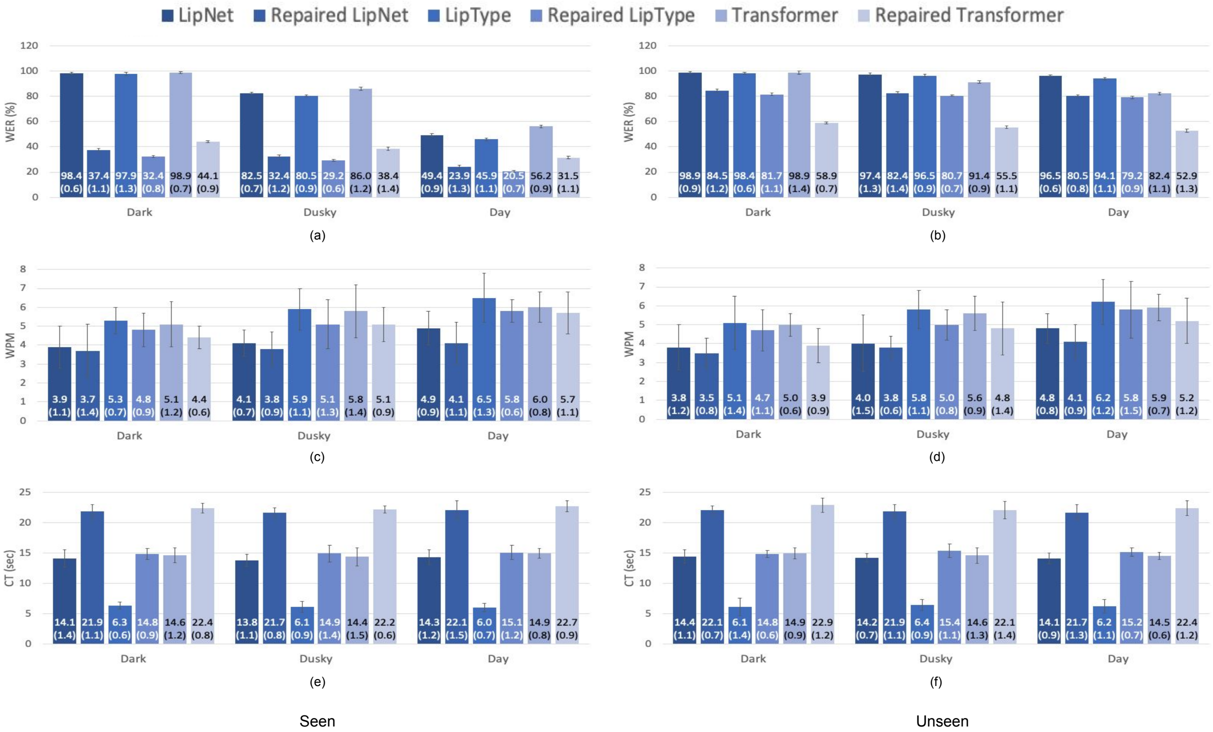


Figure 6.6: Performance evaluation of the three examined silent speech recognition models without/with the proposed repair model in terms of a) WER-Seen, b) WER-Unseen, c) WPM-Seen, d) WPM-Unseen, e) CT-Seen, and f) CT-Unseen. Each condition has 360 data points. Reported values are the average of all values. The values inside the brackets are standard deviations (SD). Error bars represent ±1 SD.

Fig. 6.6 shows the effectiveness of the repair model on silent speech recognition models. The performance of the repair model followed a similar trend as the speech models. It showed substantial reductions in error rates for all lighting conditions. With LipNet, it showed 58.1% reduction in WER for seen data and 15.5% reduction for unseen data. With LipType, it showed 61.9% reduction in WER for seen data and 16.3% reduction for unseen data. With Transformer, it showed 51.5% reduction in WER for seen data and 38.5% reduction for unseen data. On average, for all models, we observed 10.9% reduction in WPM and 8 seconds increase in CT on both seen and unseen data. For unseen data, we observed a negligible reduction in WER for LipNet and LipType compared to the Transformer model.

We speculate that this is because LipNet and LipType are trained on a relatively small GRID dataset [66] that has a smaller number of word-level classes (shorter phrases). This resulted in a much better performance for their repair models with seen data as most of the silently spoken words were in its vocabulary. Likewise, it did not perform as well with unseen data as many of the silently spoken words were not in its vocabulary (thus could not be fully processed by the language model). Transformer, in contrast, is trained on LRS dataset [5] that has a larger number of word-level classes (longer phrases). This resulted in a much lower WER for repaired Transformer with unseen data as it provided the language model with more accurate words than LipType. Note that the language model is part of the repair model not the recognizer. It is trained on a more comprehensive LIBRISPEECH dataset [218]. But its effectiveness is reliant on the vocabulary of the recognizer.

We also performed extensive ablation studies on each submodule of our model to demonstrate their contribution to the overall performance gains.

6.1.8 Ablation Studies

In this section, we present the results of various ablation studies performed to demonstrate the contribution of each submodule of our model to the overall performance gains.

6.1.8.1 With only Pre-processing

The purpose of this study was to analyze the effects of pre-processing on silent speech recognition model's performance in terms of WER, WPM, CT. For evaluation, we considered all pre-trained models as baselines and compared with their conjunction with pre-processing. Results revealed that the proposed pre-processing module substantially reduced the error rates of all pre-trained models (Table 6.6). In the study, pre-processing with LipNet showed 15% reduction in WER with seen and 7% reduction with unseen data. With LipType, it showed 12% reduction in WER with seen and 5.5% reduction with unseen data. With Transformer, it showed 24% reduction in WER with seen and 8% reduction with unseen data. On average, for all models, there were 5% reduction in WPM and 2 sec. increase in CT with both seen and unseen data. Note that the performance of these models with pre-processing and post-processing (repaired) are shown in Fig. 6.6.

6.1.8.2 Effects of Individual Error Correction Module

We also analyzed the effects of individual error correction modules with the LipType model in terms of WER and CT. All the presented results are calculated with seen data. Results demonstrated that each submodule made a significant contribution to the overall performance improvement of the repair model (Table 6.7).

Model	Seen			Unseen		
	WER	WPM	CT	WER	WPM	CT
LipNet	49.4	4.9	14.3	96.5	4.8	14.1
PP + LipNet	42.0	4.8	16.4	89.5	4.6	15.9
LipType	45.9	6.5	6.0	94.1	6.2	6.2
PP + LipType	40.9	5.6	8.5	88.9	6.0	8.1
Transformer	56.2	6.0	14.9	82.4	5.9	14.5
PP + Transfomer	42.5	5.9	16.4	76.0	5.6	15.8

Table 6.6: Performance evaluation of the three examined silent speech recognition models without/with the Pre-processing (PP) module in terms of WER, WPM and, CT for seen and unseen data.

Method	WER	CT
LipType	45.9	6.0
PP + LipType	40.9	8.3
PP + LipType + DDA	29.7	11.1
PP + LipType + DDA + SC	27.5	11.7
PP + LipType + DDA + SC + LM	24.1	14.2
PP + LipType + DDA + SC + LM + ED	20.5	15.1

Table 6.7: Effect of individual error correction module on LipType's WER and CT with seen data (Pre-processing: PP; DDA: Deep denoising autoencoder; SC: Spell Checker; LM: Language Model; ED: Edit Distance). We considered DDA + SC + LM + ED as the post-processing module.

6.1.8.3 Correction Classification

In this study, we analyzed the types of correction made by the post-processing module. For this, we classified all errors by the following criteria:

- Whether the correct word is substituted with other word(s), substitution error.
- Whether the new word(s) is inserted, insertion error.
- Whether the correct word(s) is deleted, deletion error.

After analysis, we observed that Silent Speech has 2% insertion, 27% deletion, 71% substitution, (34% of these were on short words <= 3 chars, 11% of these were on long words > 3 chars, 29% of these were in starting of the phrase <= length(phrase)/2-1, 14% of these were in ending of the phrase > length(phrase)/2). However, speech has 38% insertion, 21% deletion, 41% substitution (12% of these were on short words <= 3 chars, 18% of these were on long words > 3 chars, 25% of these were in starting of the phrase <= length(phrase)/2-1, 11% of these were in ending of the phrase > length(phrase)/2).

Silent speech has 94.7% fewer insertion errors than speech. We speculate that this is because, for speech input, the recognition model captures background noises and recognizes them as words which resulted in more insertion errors. Unlike speech recognition, silent speech recognition just uses visual information for recognition which does not get affected by background noise. Besides, silent speech has 73.1% more substitution errors. We hypothesize that this is because it is more difficult to distinguish between homophones with just visual information due to ambiguity in lip movements i.e., different characters that produce exactly the same lip sequence (e.g. ‘p’ and ‘b’). This may have resulted in more substituted words.

6.2 Discussion

We developed an independent repair model that processes video input for poor lighting conditions and corrects potential errors in output for increased accuracy. We evaluated the repair model’s effectiveness with various speech and silent speech recognizers. To demonstrate its benefit, we selected six pre-trained models, i.e., three for speech and three for silent speech. We then conducted a user study with twelve participants to collect diverse data under real-world conditions. For speech models, we collected data in indoor, outdoor, and quiet noisy conditions. For silent speech, we collected data in dark, dusky, and day lighting conditions. We then evaluated the impact of the repair model on each model’s performance using the collected data. Results showed significant improvement in the performance of all models. Models augmented with the repair model outperformed the original models drastically for all experimental conditions. For speech, we observed 32% reduction in WER, 5.8 seconds increase in CT, and 8.1% reduction in WPM; whereas for silent speech, we observed 57.2% reduction in WER, 7.9 seconds increase in CT, and 10.3% reduction in WPM. Since speech models do not involve preprocessing, their repaired models showed 26.2% less CT than silent speech models.

On comparing the performance of LipNet and LipType [See Chapter 5] from Fig. 5.2 and Fig. 6.6(a):Day, we observed a 45-50% reduction in their WER. We speculate that this is because the dataset used to evaluate both models for seen speakers comprises of uniform visual attributes (same skin tone, accent, pace of speech, etc.) (Fig. 5.2). However, the dataset for final evaluation used new speakers’ data that solicited more variability in terms of speaker characteristics (Fig. 6.6(a):Day). We also observed that the repaired Transformer performed much better than the other silent speech models on unseen data. We speculate that this is because Transformer is trained on LRS dataset that has a larger number of word-level classes (longer phrases). This resulted in a much lower WER for repaired Transformer with unseen data as it provided the language model with more accurate words than LipType. Overall, empirical results demonstrate the effectiveness of the repair model on all recognition models for improving accuracy.

Despite these improvemnts, research shows that no matter how robust the speech recognition system is, it could still fail due to a variablity in user characteristics. Speaking rate for example, is a fundamental user characteristics that can influence speech recognition performance due to the variation in acoustic properties of human speech production, such as vowel and consonant duration, the transition between phoneme and stops, and distortions in the temporal and spectral domains. Therefore, in the next step, we will investigate the effects of speaking rate on silent speech recognition.

Chapter 7

Effects of Speaking Rate on Silent Speech Recognition

Speaking rate is a fundamental user characteristics that can influence speech recognition performance due to the variation in acoustic properties of human speech production, such as vowel and consonant duration, the transition between phoneme and stops, and distortions in the temporal and spectral domains [101, 91, 305]. Some studies report that faster speaking rates result in higher error rates [91, 261, 265, 201], whereas some identified slower speaking rates to be more error prone [101, 266]. This disagreement encourages re-investigation of the effects of speaking rates on speech recognition performance. Besides, no such investigations have been conducted for silent speech recognition. This Chapter explores whether native and non-native speakers interact differently with speech and silent speech-based methods, whether speaking rate affects recognition rates of these methods, the optimal speaking rates for increased accuracy, and whether the effects of speaking rate are different for native and non-native speakers.

7.1 User Study 1: Speaking Rate

This study investigates whether native and non-native speakers speak at different rates when interacting with speech and silent speech-based methods.

7.1.1 Apparatus

We developed a custom app with Android Studio 3.1.4 (Fig 7.1). Participants used it on their own Android smartphones. Its *landing page* included a drop-down menu to select a recording condition (speech, silent speech) and a Start button to start data collection. The *data collection page* displayed the front camera in real-time, random phrases from a set [188] for participants to speak or silently speak, and a Record and Stop toggle button to start and stop recording, respectively. The app stored all videos locally and automatically logged the duration of each spoken phrase.

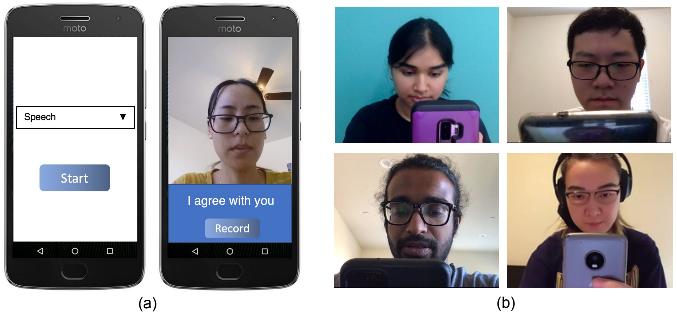


Figure 7.1: (a) Screenshots of the custom app used in the study: the landing page (left) and the data collection page (right). (b) Four volunteers participating in the first study through a teleconferencing system.

7.1.2 Participants

Twelve volunteers took part in the user study (Fig. 7.1). Table 7.1 presents the demographics of the participants divided into native and non-native groups. Originally, we wanted to recruit equal number of native and non-native speakers, but were unable to do so due to the spread of COVID-19.

Table 7.1: Demographics of the participants.

	Native (N = 4)	**Non-native (N = 8)**
Age	22–54 years (M = 32.2, SD = 14.7)	19–33 years (M = 25.8, SD = 4.7)
Gender	1 female, 3 male	4 female, 4 male
Experience with speech	1–8 years (M = 3.5, SD = 3.3)	1–4 years (M = 1.2, SD = 0.8)
Experience with silent speech	None	None

7.1.3 Design and Metrics

The study had one within-subjects independent variable: *medium*, with three levels: *baseline*, *speech*, and *silent speech*; and one between-subjects independent variable: *speaker*, with two levels: *native* and *non-native*. The baseline condition recorded participants' speaking rates in human-human communication, while the speech and silent speech conditions recorded their speaking rates with a speech and silent speech recognizer, respectively, through the mobile app. We used a Wizard-of-Oz setup, that is, the app did not include actual recognizers but pretended to accurately recognize all spoken and silently spoken phrases as long as the participant's face was visible to the app. For the baseline condition, we extracted one minute of speech from the conversations we had with the participants during the app installation and demonstration process. In the speech and silent speech conditions, participants spoke and silently spoke 30 phrases from a set [188], respectively (720 phrases, in total). The dependent variables were:

- **Time per phoneme** (TPP) is the average time participants took to utter a phoneme (in milliseconds), calculated using the following equation: $TPP = \frac{\text{time per phrase}}{\text{total phoneme in phrase}}$. Total phoneme in a recognized phrase was counted with the Pronouncing API[1] that uses the Carnegie Mellon University (CMU) Pronouncing Dictionary[2] to identify phonemes.

- **Actual words per minute** (A-WPM) is the most commonly used metrics for calculating speaking rate [49, 16]. It measures the average number of actual words spoken in a minute. This metric is different from the traditional WPM metric that considers five characters as one word regardless of the actual number of words in a phrase [19]. A-WPM is calculated using the following equation: $WPM = \frac{\text{total words}}{\text{number of minutes}}$.

7.1.4 Procedure

The study was conducted remotely via Zoom due to COVID-19. We scheduled individual video calls with each participant. They were instructed to join the call from a quiet room to avoid any interruptions during the study. In the call, we first explained how speech and video-based silent speech recognition systems work, then demonstrated the custom app and collected their informed consents and demographics using electronic forms. We then shared the app installation file (APK) with them and guided them through the installation process on their smartphones. The data collection session started after that, where the app displayed one phrase at a time. Participants were instructed to press the Record button, speak or silently speak the presented phrase, then pressed the Stop button. They were told that the system will process the spoken or silently spoken phrase when they press the Stop button. If the phrase is correctly recognized, it will display the next phrase, otherwise will ask them to re-speak the same phrase. However, in reality, the app did not include a recognizer, instead pretended to correctly recognize all spoken and silently spoken phrases. The Zoom sessions were recorded to extract one minute of speech for the baseline condition (Section 7.1.3). Participants were not informed of this during the study to avoid a potential Hawthorne effect [184]. Upon completion, participants shared all locally stored video clips and log files with us by uploading those to a cloud storage. They then took part in an interview about their experience with the app. Finally, we debriefed them about the Wizard-of-Oz setup

[1] Pronouncing API: https://pronouncing.readthedocs.io/en/latest/pronouncing.html#pronouncing.phones_for_word

[2] CMU Pronouncing Dictionary: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

and informed them that clips from the demo and installation Zoom session will be used to measure their natural speaking rates.

7.2 Results

A complete study took 45–60 minutes. A Shapiro-Wilk test revealed that the response variable residuals were normally distributed. A Mauchly's test indicated that the variances of populations were equal. Hence, we used a one-way repeated-measures ANOVA to study the effects of *medium*, a one-way between-subjects ANOVA for the effects of *speaker*, and a mixed-design ANOVA for the *medium* × *speaker* interaction effects [18].

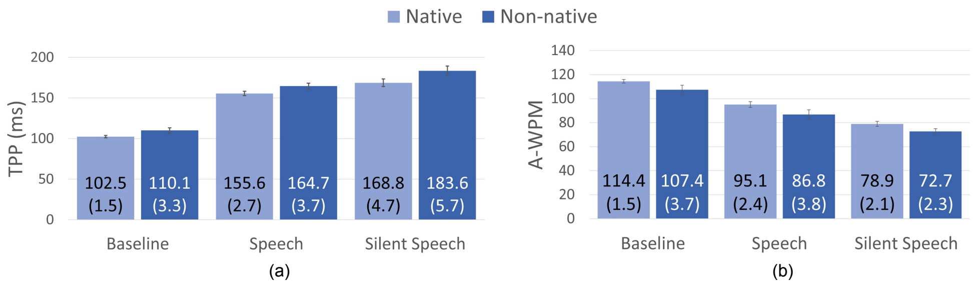


Figure 7.2: (a) Average time per phoneme (TPP) and (b) average actual words per minute (A-WPM) for native and non-native speakers with the three investigated mediums. The values inside the brackets are standard deviations (SD). The error bars represent ±1 SD.

7.2.1 Time per Phoneme (TPP)

An ANOVA identified a significant effect of medium ($F_{2,11} = 697.59, p < .0001$) on TPP. On average, participants took 107.5 ms (SD = 4.6), 161.6 ms. (SD = 5.5), and 178.7 ms (SD = 8.8) to utter a phoneme in the baseline, speech, and silent speech conditions, respectively. An ANOVA also identified a significant effect of speaker ($F_{1,10} = 1212.35, p < .0001$) on TPP. On average, native participants took 161.26 ms (SD = 10.78) to utter a phoneme, while non-native participants took 173.14 ms (SD = 13.36). There was also a medium × speaker interaction effect ($F_{1,20} = 66.02, p < .0001$). Fig. 7.2 (a) presents average TPP for native and non-native speakers with the three mediums.

7.2.2 Actual Words per Minute (A-WPM)

An ANOVA identified a significant effect of medium ($F_{2,11} = 1783.18, p < .0001$) on A-WPM. On average, participants yielded 109.7 (SD = 4.6), 89.5 (SD = 5.2), and 74.8 (SD = 3.6) A-WPM in the baseline, speech, and silent speech conditions, respectively. An ANOVA also identified a significant effect of speaker ($F_{1,10} = 1467.96, p = .0001$) on A-WPM. On average, native participants yielded 87.49 A-WPM (SD = 9.04), while non-native participants yielded 80.26 A-WPM (SD = 8.42). There was also a medium × speaker interaction effect ($F_{1,20} = 17.18, p < .0001$). Fig. 7.2 (b) presents average A-WPM for native and non-native speakers with the three mediums.

7.3 Discussion

Both native and non-native speakers spoke at much slower rates compared to their usual speaking rates while using the speech and the silent speech recognizers. On average, participants took 33.4% and 39.8% extra time to utter a phoneme with speech and silent speech, respectively. Likewise, A-WPM dropped by 22.5% and 46.6%, respectively. Consequently, a post-hoc Tukey-Kramer multiple-comparison test identified two distinct groups: {baseline} and {speech, silent speech}. The post-study interview revealed that participants spoke slowly while using these methods thinking that it would increase their recognition rates. However, there was no actual effects on phrase recognition as the Wizard-of-Oz approach pretended to correctly recognize all spoken or silently spoken phrases. Since all participants were experienced users of various voice assistant systems, it is likely that the unreliability of these systems encouraged them to reduce the rate of their speech. Relevantly, a participant (female, 27 years, non-native) said, *"It [speaking slowly] is mostly due to lack of proficiency and different accent. I always try to speak slowly and try to match accent to make the speech assistant understand me which is sometimes awkward and irritating"*. Surprisingly, they spoke at a much slower rate when using a silent speech recognizer compared to when using a speech recognizer. This could be either because participants never used a silent speech-based method before or the fact that video-based silent speech recognizers detect speech based on lip movements rather than the sound produced by the speakers (Section 7.1.4), giving them the impression that the method requires extra finesse for an acceptable accuracy rate. Post-study interview revealed that participants overemphasized their lip movements during silent speech to "aid" the recognition process.

Results revealed that non-native speaker spoke at a slower rate than native speakers (about 7% slower TPP). This is not surprising since many studies found out that average speaking rate for non-native speakers is slower than for native speakers as *"a general lack of proficiency and experience can result in slower speaking rates"* [27, 76, 109, 69, 68]. However, both native and non-native speakers slowed down at comparable rates when interacting with speech (∼34% slower TPP) and silent speech (∼40% slower TPP) recognizers. This finding is interesting as it suggests that these slower speaking rates were not caused by the lack of proficiency or experience but due to the speakers' skepticism about the reliability of the state-of-the-art speech and silent speech recognizers. Based on these findings, we recommend evaluating new speech and silent speech recognizers with both native and non-native speakers of the target language, and report the results of the two groups separately due to their significantly different speaking rates. The fact that users slow down when interacting with speech and silent speech recognizers can also be exploited for improved performance.

We were unable to study any potential effects of recognition error on speaking rate since the Wizard-of-Oz setup collected data without any errors. However, users are likely to adjust their interaction behavior when interacting with an error-prone system, like observed in other recognition systems [21]. Another limitation of the study is using different scenarios in the baseline and the speech conditions. Speaking rate for the baseline was calculated in continuous computer-mediated communication, while the same for the speech and the silent speech were calculated from manually segmented phrases. It is unknown whether the additional latency introduced by the manual segmentation affected the speaking rate in any way. It is also unclear if the speaking rates are different for computer-mediated and face-to-face communications, although prior works reported other behavioral changes [154].

7.4 User Study 2: Effects of Speaking Rate

This study investigated whether speaking rate affects recognition rates of state-of-the-art speech and silent speech recognizers.

7.4.1 Participants and Design

We invited the participants of the previous study (Section 7.1.2) to take part in this study. The study had two within-subjects independent variable: *medium* and *speaking rate*. The former had two levels: *speech* and *silent speech*, and the latter had seven levels: 0.25x, 0.5x, 0.75x, 1x, 1.25x, 1.5x, and 1.75x of the actual speaking rates of the participants. These rates were selected based on YouTube's playback speed scale, ranging from quarter speed (0.25x) to double speed (2x). Among these, we selected the actual rate (1x), the top three slower rates (0.25x, 0.5x, 0.75x) and the top three faster rates (1.25x, 1.5x, 1.75x), resulting in seven rates in total. The study had one between-subjects independent variable: *speaker*, with two levels: *native* and *non-native*. Participants spoke 30 phrases from the respective model's training dataset [see Chapter 6 A.1 & A.5]], which were post-processed to achieve the seven speaking rates, resulting in 30 phrases × 2 mediums × 7 speaking rates = 420 phrases per participant. The dependent variable was the following performance metric:

- **Word accuracy** (WA) measures the total number of words accurately recognized from the total number of spoken words. It is calculated using the following equation, where S is the number of substitutions, D is the number of deletions, I is the number of insertions, N is the number of words in the ground truth: $WA = 1 - \frac{(S+D+I)}{N}$.

7.4.2 Apparatus and Procedure

We modified the custom app used the previous study to replace the phrases [188] with phrases from the examined speech and silent speech recognition models' training datasets. We also

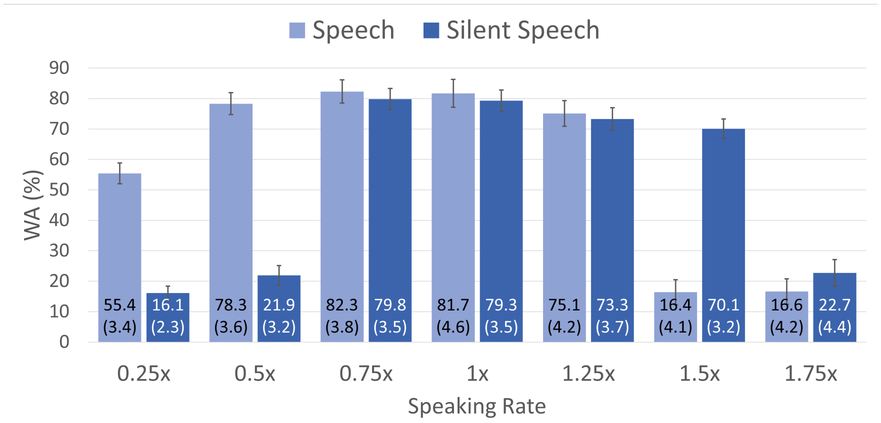


Figure 7.3: Average word accuracy rates (%) of the speech and the silent speech recognition methods with the seven examined speaking rates. The values inside the brackets are standard deviations (SD). The error bars represent ±1 SD.

included a new condition in the app, where participants are instructed to read the presented phrases. Recorded video clips were time-expanded for the slower rates and time-compressed for the faster rates using the FFmpeg [3] platform. All clips were then processed using two state-of-the-art pre-trained recognition models for speech and silent speech: Kaldi (Api.ai) [234] and LipType [see Chapter 6], respectively.

The study used the same procedure as the first study except for the demonstration and the post-study debrief and interview. The custom app displayed one phrase at a time, and participants were instructed to read it at a rate in which they would usually speak with another person. Note that, despite the different speaking rates, each participant spoke exactly the same number of words in each condition.

7.5 Results

A complete study took about 30 minutes. A Shapiro-Wilk test revealed that the response variable residuals were normally distributed. A Mauchly's test indicated that the variances of populations were equal. Hence, we used a two-way repeated-measures ANOVA to study the effects of *medium* and *speaking rate*, a one-way between-subjects ANOVA to to study the effects of *speaker*, and a mixed-design ANOVA to study the interaction effects [18].

[3] A Complete, Cross-Platform Solution to Record, Convert and Stream Audio and Video: https://www.ffmpeg.org

7.5.1 Word Accuracy (WA)

An ANOVA identified a significant effect of medium ($F_{1,11} = 64769.13, p < .0001$) and speaking rate ($F_{6,66} = 697.21, p < .0001$) on WA. The medium × speaking rate interaction effect was also statistically significant ($F_{6,66} = 33009.02, p < .0001$). Fig. 7.3 illustrates the average WA of the speech and the silent speech recognition methods with the seven speaking rates. An ANOVA also identified a significant effect of speaker ($F_{1,10} = 805.74, p < .0001$). The speaker × medium ($F_{1,10} = 64.54, p < .0001$) and the speaker × speaking rate × medium ($F_{6,60} = 543.30, p < .0001$) interaction effects were also statistically significant. Fig. 7.4 illustrates the average WA of the speech and the silent speech recognition methods for native and non-native speakers with the seven examined speaking rates.

7.5.2 Error Analysis

We conducted a post-hoc analysis of the recognized phrases at the usual speaking rate (x1) to find out the distribution of insertion errors (extra words are incorrectly inserted), deletion errors (correct words are incorrectly omitted), and substitution errors (words are substituted with incorrect words) [29]. Table 7.2 presents the results.

7.6 Discussion

Both speech and silent speech methods performed well with regular (1x) speaking rate. On average, speech and silent speech methods yielded 82% (SD = 4.6) and 80% (SD = 3.5) WA, respectively, with regular speaking rate. The effects of speaking rate was different for native and non-native speakers. At regular rate, speech and silent speech methods were 9.9% and 7.5% more accurate, respectively, for native speakers than non-native speakers. However,

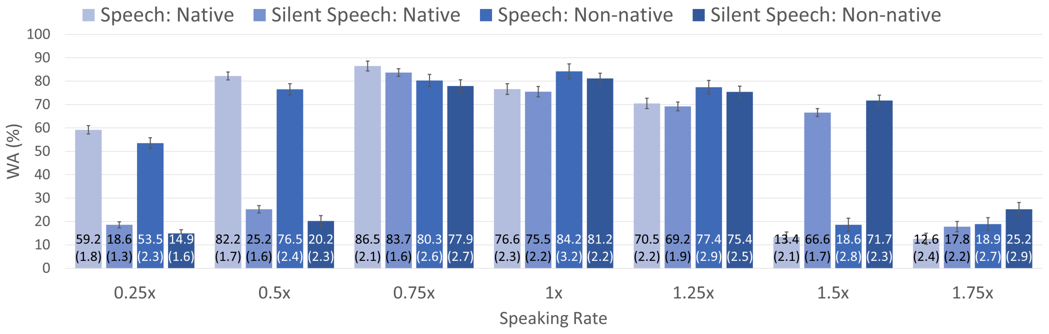


Figure 7.4: Average word accuracy rates (%) of the speech and silent speech recognition methods for native and non-native speakers with the seven examined speaking rates. The values inside the brackets are standard deviations (SD). The error bars represent ±1 SD.

Table 7.2: Distribution of insertion, deletion, and substitution errors in the phrases recognized by the speech and the silent speech recognizers.

	Speech			Silent Speech		
	All	Native	Non-Native	All	Native	Non-Native
Insertion	38%	12%	29%	2%	4%	22%
Deletion	21%	41%	37%	27%	30%	21%
Substitution	41%	47%	34%	71%	66%	57%

for both native and non-native speakers, the performance of the speech recognition method dropped substantially with speaking rates lower than 0.5x and higher than 1.25x (Fig. 7.4). A post-hoc Tukey-Kramer multiple-comparison test revealed that 0.75x speaking rate was significantly more accurate than the other speaking rates. Likewise, the performance of the silent speech recognition method dropped substantially with rates lower than 0.75x and higher than 1.25x for both native and non-native speakers (Fig. 7.4). Like speech, a post-hoc Tukey-Kramer multiple-comparison test identified 0.75x as significantly more accurate than the other examined speaking rates. These findings suggest that speaking slightly slower than usual can indeed increase the reliability of speech and silent speech recognizers, regardless of the speaker's proficiency and experience in English. Results also suggest that 0.5–1.25x is the optimal range for speech and 0.75–1.25x is the optimal range for silent speech for higher accuracy rates. We speculate, this is due to the fact that much faster speaking rates can cause frequent and stronger pronunciation changes while much slower speaking rates tend to add unnecessary pauses between phonemes [191]. The average natural speaking rate was slower in this study than the first study since, here, participants read the phrases, which is slower than speaking [137, 136, 192, 212].

Error analysis revealed that silent speech had 94.7% lower insertion errors than speech. We speculate, this is because ambient noise in the audio affected the recognition of the speech method. Silent speech, in contrast, uses visual information for recognition, thus was not affected by background noise. Interestingly, speech committed 11% higher deletion errors and 38% higher substitution errors for native speakers than non-native speakers. This could be because faster rates resulted in overlaps between the words, making it difficult to segment them. Silent speech also resulted in 81% lower insertion errors, 42% higher deletion errors, and 16% higher substitution errors than non-native speakers, presumably for the same reasons. Silent speech had 73.1% higher substitution errors than speech, which could be due to the difficulty in distinguishing between different homophones with visual information as multiple characters can produce the same lip movement sequence, such as for the letters 'p' and 'b'.

The findings of this work highlight the importance of considering speaking rate in speech and silent speech-based interfaces. While designing interfaces for these methods, the recognition algorithms must be optimized for varying speaking rates and the characteristics of native and non-native speakers. Error analysis presented in this work could be used to identify areas that require extra effort to increase the respective method's accuracy rates. The findings could also provide guidance to users on improving speech and silent speech input performance.

In spite of encouraging results, it is vital to note that the presented findings are either based on Wizard-of-Oz or simulation studies, so they may vary when tested with the actual system in a real-world setting. It is therefore imperative to test the modality with the real system. In the next Chapter, we will investigate silent speech as a hands-free selection method in eye-gaze pointing.

Chapter 8

Silent Speech-Based Selection Method for Eye-Gaze Pointing

Eye-gaze-based interaction is a promising modality for faster and seamless hands-free (also known as contactless or touchless) interaction [262]. It enables people with limited motor skills to interact with computer systems without using the hands [166, 7, 43, 140, 67]. It is also beneficial in Situationally-Induced Impairments and Disabilities (SIID) [300, 253], when the hands are incapacitated due to reasons such as performing a secondary task, minor injuries, or unavailability of a keyboard [239]. Hands-free interaction is also of a particular interest in situations when touching public devices is to be avoided to prevent the spread of an infectious disease [134].

Eye tracking technologies measure a person's eye movements and positions to understand where the person is looking at any given time. In the past, eye tracking required expensive, often non-portable extramural devices, which were slow and error prone [165, 306]. Recent developments have made eye tracking more affordable, portable, and reliable. Modern algorithms can track eyes using webcams almost as fast and accurately as commercial tracking technologies [165, 256, 302]. The most common application of eye tracking is to direct control a mouse cursor using eye movements [306]. While the idea of performing tasks simply by looking at the interface is empowering, eye tracking has yet to become a pervasive technology due to the "Midas Touch" problem [139], which refers to the classic eye tracking problem where the system cannot distinguish between users simply scanning the items versus their intention to select them, resulting in unwanted selections wherever the user looks, making the system unusable. One solution to this problem is to use a different action to activate selection. The most commonly used selection method with eye tracking is dwell, where users look at a target for 100–3,000 ms [112] to select it. It is, however, difficult to pick the most effective dwell time for a population since a short dwell time makes the system faster but increases false positives, while a long dwell time makes the system slower and causes users physical and cognitive stress [37, 112]. Many alternatives have been proposed to substitute dwell, including head and gaze gestures, blinking, voluntary facial muscle activation, brain signals, and foot pedals. Most of these approaches either use external, invasive hardware that are not yet scalable in practical situations or exploit unnatural behaviors that can cause users irritation and fatigue [138]. Speech is promising but not reliable in noisy places (e.g., when listening to music). Users are also hesitant to use speech when in public places (e.g., in a library) [82, 86, 84, 235]. Besides, speech does not work well with people with severe speech disorders since it relies on the sound produced by the users [7, 43].

In this chapter, we investigate silent speech as an alternative selection method for eye-gaze pointing. First, we propose a stripped-down image-based model that can recognize a small number of silent commands almost as fast as state-of-the-art speech recognition models. Second, we design a silent speech-based selection method and compare it with other hands-free selection methods, namely dwell and speech, in a Fitts' law study. We follow-up on this by conducting another study investigating the most effective screen areas for eye-gaze pointing in terms of throughput, pointing time, and error rate. Finally, we design a silent speech-based menu selection method for eye-gaze pointing and evaluate it in an empirical study.

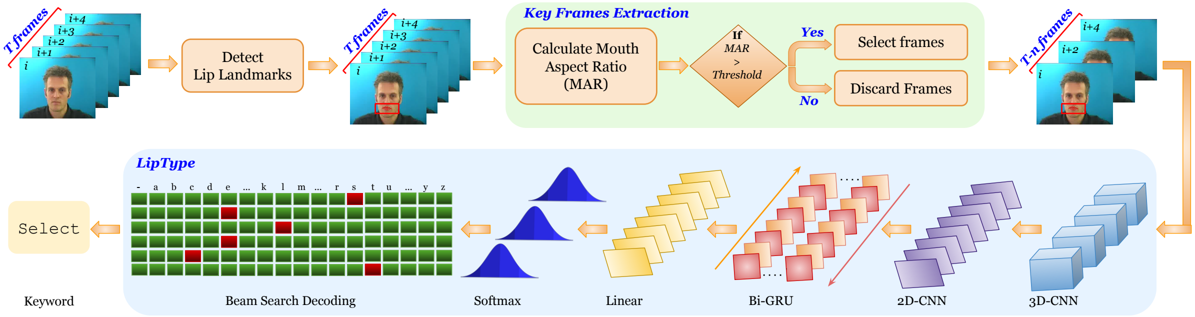


Figure 8.1: The architecture of the proposed silent command recognition model. It pre-processes a sequence of T frames for mouth-centered cropped images to extract key frames. The key frames are fed to a 1-layer 3D CNN, followed by a 34-layer 2D SE-ResNet for spatiotemporal feature extraction. The features are then processed by two Bi-GRUs, a linear layer, and a softmax. Finally, the softmax output is decoded with a left-to-right beam search using the Stanford-CTC decoder.

8.1 A Model for Silent Command Recognition

We customized an existing silent speech recognition model LipType [see Chapter 6] to recognize silent commands. We did not use an off-the-shelf recognizer since they are optimized for recognizing phrases, thus trained on large corpora (≥1,000 phrases [66]). This increases the variability and ambiguity in lip movements (similar movements for different characters), which are disambiguated in post-processing using language models [25],[see Chapter 6]. This affects both speed and accuracy. State-of-the-art silent speech recognition models can take up to 5,000 ms to recognize one word with accuracy rates between 53–96% [see Chapter 6]. Since voice assistants usually use a small number of words as commands, we used a smaller set of words that can be distinguished based on mouth aspect ratios (MAR) and scraped off all word and phrase-level language models. The proposed model consists of three sub-modules: a *key frames extraction* frontend that takes a sequence of video frames and extracts key frames to create a compact representation, a *spatiotemporal feature extraction* module that takes a sequence of key frames and outputs one feature vector per frame, and a *sequence modeling* module that inputs the sequence of per-frame feature vectors to recognize a keyword. The model is capable of mapping variable-length video sequences to text sequences. Fig. 8.1 illustrates the architecture of the model.

Module 1: Key frames extraction. This module crops one w:100 × h:50 pixels mouth-centered image per video frame to extract key frames. The module pre-processes each video clip with the DLib face detector [157] and the iBug face landmark predictor [251] with 68 facial landmarks (L) and Kalman filtering (Fig. 8.2, left). Then, mouth-centered cropped images are extracted by applying affine transformations. These images are used to measure MAR by dividing the distance between the upper and the lower lips (h) with the distance between the left and the right corners of the mouth (w) (Eq. 8.1). All frames with a MAR greater than 20 are considered as key frames and the remaining frames are discarded to reduce computation time. This threshold was picked based on an ablation study that revealed that a MAR greater than 20 is sufficient to distinguish between words in a corpus with 10 words (Fig. 8.2, right).

$$MAR = \frac{\|L_{61} - L_{56}\| + \|L_{60} - L_{57}\| + \|L_{59} - L_{58}\|}{2 * \|L_{44} - L_{50}\|} \tag{8.1}$$

Module 2: Spatiotemporal feature extraction. This module passes the extracted key frames to a 3D-CNN with a kernel dimension of $T{:}3 \times W{:}5 \times H{:}5$, followed by Batch Normalization (BN) [133] and Rectified Linear Units (ReLU) [6]. Then, the extracted feature maps are passed through a 34-layer 2D SE-ResNet to gradually decrease the spatial dimensions with depth until the feature becomes a single dimensional tensor per time step.

Module 3: Sequence modeling. This module processes the extracted features using two Bidirectional Gated Recurrent Units (Bi-GRUs) [61]. Each time-step of the GRU output is processed by a linear layer and a softmax layer over the vocabulary, and an end-to-end model is trained with connectionist temporal classification (CTC) loss [107]. The output is then decoded with a left-to-right beam search [64] using the Stanford-CTC decoder [183] to recognize spoken keywords.

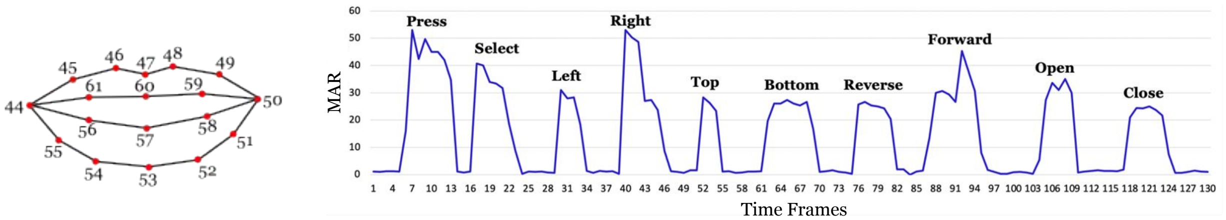


Figure 8.2: From left, lip landmarks detected by DLib and iBug [157], and average mouth aspect ratios (MAR) of the ten keywords.

8.1.1 Training and Implementation

We trained the model for ten keywords: *Press*, *Select*, *Left*, *Right*, *Top*, *Bottom*, *Reverse*, *Forward*, *Open*, *Close*, with the data collected from 20 participants: 9 female, 11 male, average age 26.95 years (SD = 3.03). The data collection process occurred remotely. Participants sat in front of their computers and silently spoke each keyword in a random order for 50 times (20 participants × 10 keywords × 50 repetitions = 10,000 samples). We enabled them to use the embedded cameras to increase the variability of the dataset. They were instructed to take 1–2 minutes breaks between the words and ∼3 seconds breaks between the repetitions to reduce the effects of fatigue. A researcher guided them and observed the whole process via a videotelephony system. Before training, we pre-processed the data by applying a horizontally mirrored transformation, color space augmentations, and random cropping on the cropped mouth images, resulting in 42,981 samples in total (4,290/keyword). We augmented the dataset with simple transformations to reduce overfitting. The number of frames was fixed to 75. Longer image sequences were truncated and shorter sequences were padded with zeros. We applied a channel-wise dropout [274] of 0.3. The model was trained end-to-end by the Adam optimizer [159] for 60 epochs with a batch size of 50. The learning rate was set to 10^{-3}. The network was implemented on the Keras deep-learning platform with TensorFlow [1] as the backend. Wll models were trained and tested on an NVIDIA GeForce 1080Ti GPU board.

8.1.2 Performance Evaluation

We conducted a study to compare the performance of the proposed silent command model with a state-of-the-art speech (Google Speech-to-Text API [103], Kaldi (Api.ai) [234]) and silent speech (LipType [see Chapter 6]) recognition models to determine if it is reliable enough as a selection method in gaze-based interfaces. Twelve volunteers participated in the study (M = 27.67 years, SD = 2.77). Six of them identified themselves as female and six as male. None of them took part in the data collection process. In the study, participants either spoke or silently spoke (counterbalanced) each keyword for 12 times in a random order (12 participants × 2 methods × 2 models × 10 keywords × 12 repetitions = 5,760

Table 8.1: Average recognition time (seconds) and accuracy rates (%) for the investigated models.

Metric	**Method**	*Press*	*Select*	*Left*	*Right*	*Top*	*Bottom*	*Reverse*	*Forward*	*Open*	*Close*
Time	*Google*	1.73	1.64	1.65	1.54	1.69	1.82	1.82	1.68	1.65	1.61
	Kaldi	2.27	2.17	2.30	2.19	2.10	2.02	2.08	2.13	2.33	2.14
	Command	1.99	1.96	2.04	1.90	2.09	2.03	1.88	1.76	2.04	1.96
	LipType	3.09	3.28	2.95	3.24	3.02	3.09	3.18	3.15	3.09	3.38
Accuracy	*Google*	97.92	97.71	98.11	98.36	98.18	97.42	98.15	98.53	97.97	97.82
	Kaldi	88.05	88.65	90.19	87.48	89.04	88.41	85.83	88.04	89.62	88.02
	Command	77.12	79.36	73.44	72.48	72.37	71.91	71.84	72.76	79.52	76.18
	LipType	87.51	87.55	85.89	86.86	88.57	89.06	87.31	88.24	86.04	88.86

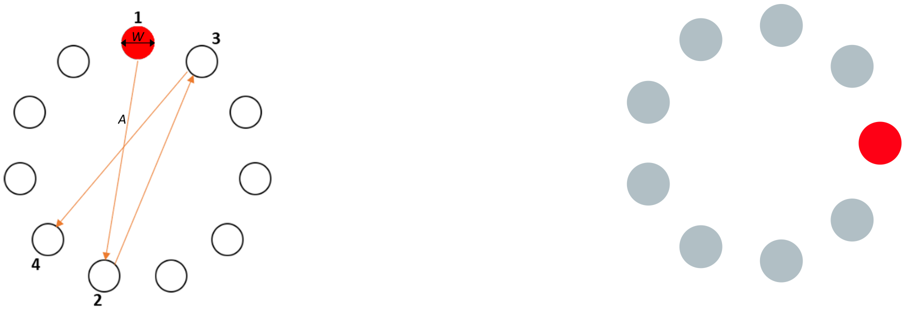
samples). A custom web application, developed with HTML5, CSS, PHP, and JavaScript, presented one keyword at a time, processed and displayed the recognized word on the screen, then presented the next keyword. The application was loaded on a Chrome web browser v92.0.4515.131 running on a MacBook Pro 16″ laptop with 2.6 GHz Intel Core i7 processor, 16 GB RAM, 3072×1920 at 226 ppi. Its built-in FaceTime HD webcam (1.2 megapixel with 1,280×720 pixel resolution) was used to track lip movements. The application automatically calculated and recorded *recognition time* (seconds): the average time to recognize a word and *accuracy rate* (%): the average percentage of words accurately recognized by a model.

8.1.2.1 Results

On average, Google Speech-to-Text and Kaldi took 1.68 seconds (SD = 0.27) and 2.17 seconds (SD = 0.42), respectively, to recognize the keywords, whereas LipType and Silent command took 3.14 seconds (SD = 0.39) and 1.97 seconds (SD = 0.34), respectively. The differences were statistically significant ($F_{3,11} = 159.65, p < .0001$). The average accuracy rates for Google Speech-to-Text and Kaldi were 97.91% (SD = 1.15) and 88.32% (SD = 5.11), respectively, whereas 87.58 (SD = 5.22) and 73.47% (SD = 7.33) for LipType and Silent command, respectively. The differences were statistically significant ($F_{3,11} = 506.53, p < .0001$). Table 8.1 presents recognition time and accuracy rates for all keywords with each method. Within the investigated models, we selected the relatively best-performed models for speech and silent speech recognition: Google Speech-to-Text and Silent command. Silent command was almost as fast as Google Speech-to-Text (1.97 vs. 1.68 seconds) but was about 24% more error prone. However, this rate was recorded in a quiet room, while research showed that the accuracy rate of speech drops by 45–55% in presence of a background noise (42–58 db) [see Chapter 6]. The performance of silent speech, in contrast, is unaffected by this. Besides, an ablation study showed that the accuracy rate of the proposed model further improves with a much smaller corpus or a larger training dataset. The model reached a 100% accuracy rate with 1 keyword and close to 95% accuracy rate with 6 keywords, which are acceptable in the context of speech and silent speech input [see Chapter 3]. In this work, we use 1 keyword: *Select*, during the Fitts' law study, and the 6 most relevant keywords: *Select*, *Left*, *Right*, *Top*, *Bottom*, *Close*, in the menu selection study.

8.2 Eye Tracking

This work uses the GazeCloudAPI for real-time eye-tracking using a webcam [95]. It tracks eyes in three stages: facial features extraction, eyes features detection, and point of gaze estimation. The process starts with capturing RGB color space images with a web camera and converting them to grayscale. These images are then normalized with histogram equalization to enhance facial feature accuracy [102]. Afterward, a Haar-like feature classifier is used to classify the images into face and non-face regions [289]. The classifier further classifies the face into subregions, such as the eyes, the nose, the lips, etc. Once the eye region is detected, the system first identifies the position of the pupil by detecting the iris from the eye region. Then, locates the pupil as the center of the iris using a Hough circle transform [155]. Finally, the point of gaze is estimated using the pupil location [98]. In an empirical evaluation [285], the API yielded 0.9°, 1° accuracy on the x, y coordinates with a Logitech Pro 9000 Webcam at 1600×1200, where participants could freely move their head. Note that eye tracking accuracy is measured in angles, representing the deviation in degrees between the actual and the predicted gaze directions. An average below 1.2° is considered to be a good measurement of accuracy in free head conditions, while an accuracy below 0.8° is desired when the head is fixed using a chinrest [285].



(a) The 2D Fitts' law task in ISO 9241-9

(b) A screenshot of the web application ($A = 780, W = 140$ pixels)

Figure 8.3: (a) The target is highlighted in red. The arrows and the numbers demonstrate the sequence in which the targets are selection. (b) The custom web application also highlights the intended target in red and uses the same selection sequence as ISO 9241-9.

8.3 Fitts' Law Protocol

Fitts' law is a well-established method for evaluating target selection on computing systems [186]. In the 1990s, it was included in the ISO 9241-9 (revised: ISO 9241-411) standard for evaluating non-keyboard input devices by using Fitts' throughput as a dependent variable [269]. The most common multi-directional protocol evaluates target selection movements in different directions. The task is 2D with targets of width W equally spaced around the circumference of a circle (Fig. 8.3a). Participants select the targets in a sequence moving across and around the circle, starting and finishing at the top target. Each movement covers an amplitude A, which is the diameter of the layout circle. A *trial* is defined as one target selection task, whereas completing all tasks with a given amplitude is defined as a *sequence*. Throughput cannot be calculated on a single trial because a sequence of trials is the smallest unit of action in ISO 9241-9. Traditionally, the difficulty of each trial is measured in bits using an index of difficulty (ID), calculated as follows:

$$ID = log_2(\frac{A}{W} + 1)$$

The movement time (MT) is measured in seconds for each trial, then averaged over the sequence of trials. It is then used to calculate the performance throughput (TP) in bits/second (bps) using the following equation:

$$TP = \frac{ID}{MT}$$

The revised ISO 9241-9 (9241-411) used here [132] measures throughput using an effective index of difficult ID_e, which is calculated from the effective amplitude A_e and the effective width W_e to make sure that the real distance traveled form one target to the next is measured. It also takes into effect how far the participants were from the target center.

$$TP = \frac{ID_e}{MT} \qquad ID_e = log_2(\frac{A_e}{W_e} + 1)$$

The effective amplitude is the real distance travelled by the participants and the effective width is calculated as follows, where SD_x is the standard deviation of the selection coordinates projected on the x-axis for all trials in a sequence. This accounts for any targeting errors by the participants, assuming that they were aiming at the center of the targets.

$$W_e = 4.133 * SD_x$$

8.4 Experimental System

We developed a custom web application[1] with HTML5, CSS, PHP, and JavaScript for the Fitts' law study protocol (Section 8.3). It enables users to control a cursor with eye-gaze

[1] Based on an existing application: http://simonwallner.at/ext/fitts.

by translating gaze position into x, y coordinates of the cursor on the display. It uses the GazeCloudAPI for eye-tracking with a webcam (Section 8.2). We used it instead of other APIs [302, 220] due to its robustness [295]. The application uses the following free-hand target selection methods.

- **Dwell.** Users point at a target then fixate (or hold the sight) for 500 ms to select the target. The threshold was picked based on studies identifying 500 ms as the most effective dwell time for novice eye-gaze users [205, 51, 185].
- **Speech Command (Google).** Users point at a target then speaks the voice command *Select* to select the target.
- **Silent Speech Command.** Users point at a target then silently speaks the command *Select* (without vocalizing the word) to select the target.

8.5 User Study 1: Fitts' Law

We conducted a Fitts' law study to investigate the performance of different hands-free selection methods (dwell, speech, silent speech) with eye tracking.

8.5.1 Participants & Apparatus

Twelve volunteers participated in the user study. Their age ranged from 24 to 40 years (M = 29.01, SD = 4.78). Four of them identified themselves as women and eight as men. Four of them wore corrective eyeglasses and one wore corrective contact lenses. One participant had experience working with the MediaPipe Iris API, but none used eye tracking to interact with their computer systems. Each of them received US $15 for volunteering in the study. We used the web application described in Section 8.4 (Fig. 8.3b) and the apparatus described in Section 8.1.2.

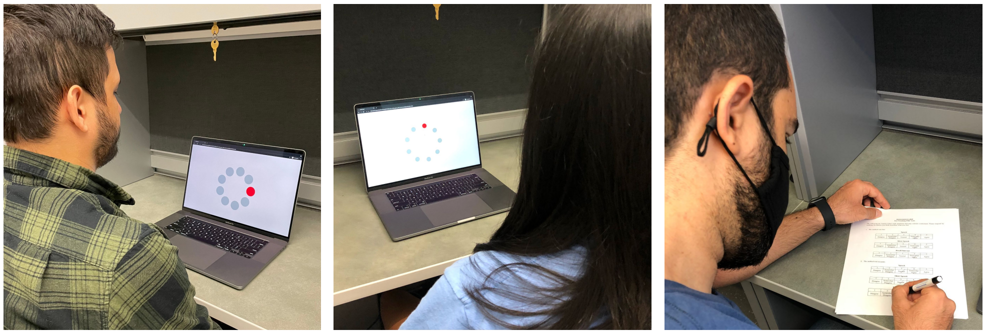


Figure 8.4: Three participants taking part in the first user study.

8.5.2 Design

The study was a 3 × 3 × 4 within-subjects design. The independent variables and the levels were as follows:

- Selection method (*Dwell*, *Speech*, *Silent Speech*) counterbalanced
- Amplitude (260, 520, 780 pixels)
- Width (35, 70, 140, 220 pixels)

The three amplitudes were selected based on the minimum and maximum distance possible on the experimental device's 16″ display. Likewise, the four widths were selected since 35 pixels is one of the smallest widths used in prior eye tracking research [199], 70 pixels is the recommended width in eye tracking applications [287], while targets with widths over 220 pixels are unrealistic. The dependent variables in the study were as follows:

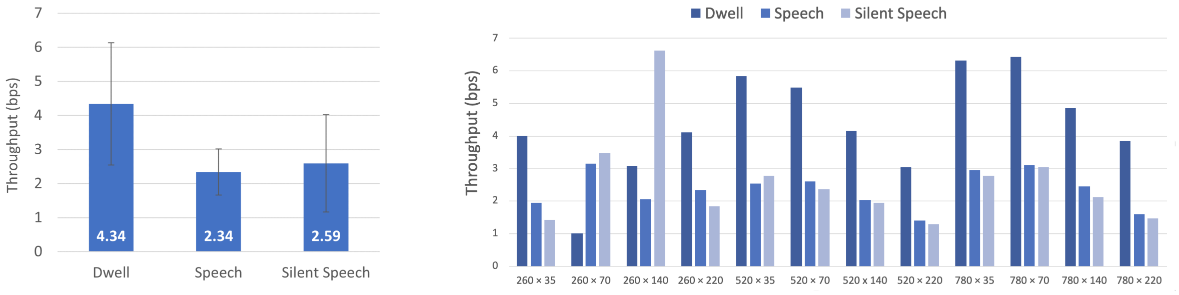
- **Throughput** (bps) as described in Section 8.3.
- **Selection time** (seconds) represents the average time users took to perform a selection task, measured from the moment the cursor entered the target (including re-entries, when the cursor mistakenly left the target, then re-entered) to the moment it was selected. This metric does not include **pointing time** (seconds) that signifies the time to move the cursor over a target as all selection methods used the same eye tracking method for pointing.
- **Error rate** (%) signifies the average percentage of incorrect target selections per trial (%), where users performed a selection action outside the target.

8.5.3 Procedure

The study was conducted in a quiet room. Upon arrival, we explained the research and demonstrated the application to the participants. They then signed an informed consent form and completed a short demographics questionnaire. We then calibrated the eye tracking system for each participant by using a 4-point calibration method. The display was located about 65–75 cm in front of the participants' eyes (Fig. 8.4), as recommended in eye tracking research [285]. After calibration, we enabled participants to practice with the application by using the three selection methods for ∼5 minutes. They could extend the practice period on request. Once familiar with the methods, they started the study by performing point-select tasks by pointing at a target using eye tracking, then selecting it using either dwell, speech, or silent speech. As per ISO 9241-411, the targets were highlighted one-by-one clockwise for all levels, starting from the top target. The amplitude and width values were selected randomly. As a target was selected, the next target was highlighted. We did not instruct participants to fix their head, thus could freely move their heads during the study. We enforced a 2-minute break after each four sequences and a 5-minute break after each condition to avoid the effect of fatigue. Upon completion of the study, participants completed a short questionnaire to rate their willingness to use and perceived physical and mental efforts of the methods on a 5-point Likert scale. All researchers involved in this study were fully vaccinated for COVID-19, wore face covering, and maintained a $3''$ distance from the participants at all times. Participants were pre-screened for COVID-19 symptoms during recruitment and on the day of the study. They wore face coverings at all times, except for when taking part in the study. All study devices and all surfaces were disinfected before and after each session. This protocol was approved by the Institutional Review Board (IRB).

8.5.4 Results

A complete study session took about 60–80 minutes, including demonstration, questionnaires, and breaks. A Shapiro-Wilk test revealed that the response variable residuals were normally distributed. A Mauchly's test indicated that the variances of populations were equal. Hence, we used a repeated-measures ANOVA for all quantitative within-subjects factors (described in Section 8.5.2). We used a Friedman test for the questionnaire data [18]. We did not identify any effects of the between-subjects factors, namely age, gender, and the use of corrective eyeglasses or contact lenses.



(a)

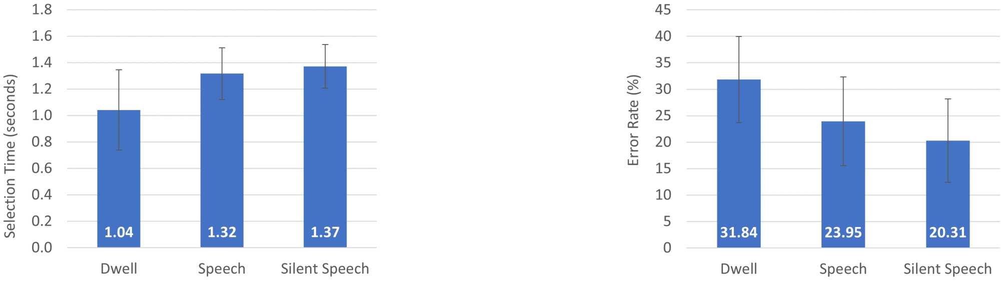
(b)

Figure 8.5: Average throughput (bits/second) by (a) selection method and (b) selection method, amplitude, and width. Error bars represent ±1 standard deviation (SD).

8.5.4.1 Throughput

An ANOVA identified a significant effect of selection method on throughput ($F_{2,22} = 2367.84, p < .0001$). Average throughput for dwell, speech, and silent speech were 4.34 (SD = 1.79), 2.34 (SD = 0.68), and 2.59 bps (SD = 1.43), respectively (Fig. 8.5a). A Tukey-Kramer test found the three selection methods significantly different from one another. There was also a significant effect of amplitude ($F_{2,22} = 189.88, p < .0001$) and width ($F_{3,33} = 487.72, p < .0001$). The method × amplitude × width interaction effect was also statistically significant

($F_{12,132} = 225.83, p < .0001$). Fig. 8.5b illustrates average throughput by selection method, amplitude, and width.



(a) Selection time (seconds)

(b) Error rate (%)

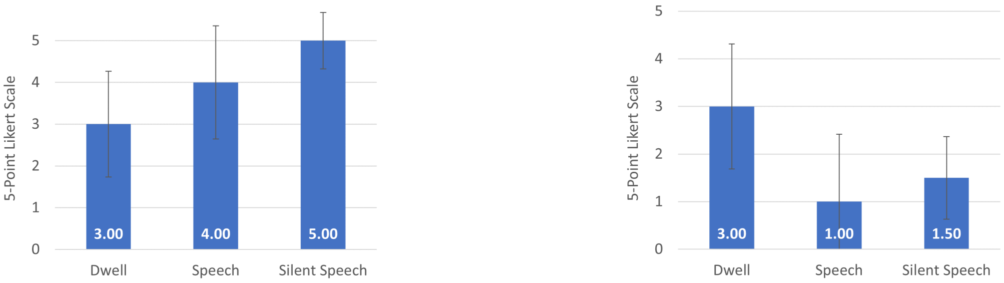
Figure 8.6: Average selection time and error rate by selection method. Error bars represent ±1 standard deviation (SD).

8.5.4.2 Selection Time

An ANOVA identified a significant effect of selection method on selection time ($F_{2,22} = 1001.30, p < .0001$). Average selection time for dwell, speech, and silent speech were 1.04 (SD = 0.30), 1.32 (SD = 0.20), and 1.37 seconds (SD = 0.17), respectively (Fig. 8.6a).

8.5.4.3 Error Rate

An ANOVA identified a significant effect of selection method on selection time ($F_{2,22} = 3932.24, p < .0001$). Average error rate for dwell, speech, and silent speech were 31.84% (SD = 8.15), 23.95% (SD = 8.38), and 20.31% (SD = 7.88), respectively (Fig. 8.6b).



(a) Willingness-to-use

(b) Physical and mental effort

Figure 8.7: Median willingness-to-use and physical and mental effort. Error bars represent ±1 standard deviation (SD).

8.5.4.4 User Feedback

A Friedman test identified a significant effect of selection method on willingness-to-use ($\chi^2 = 8.31, df = 2, p < .05$). However, no significant effect was identified on physical and mental effort ($\chi^2 = 3.33, df = 2, p = .11$). Fig. 8.7 presents median willingness-to-use and perceived physical and mental effort ratings of the three methods.

8.5.5 Discussion

Results confirmed that target amplitude and width influence the selection methods in accordance to the Fitts' law (Fig. 8.5b), except for dwell's unusual throughput for A:260 × W:140, which we identified as an outlier. Dwell was the best performed selection method in terms of throughput. Its 4.34 bps throughput was 85% and 68% higher than speech and silent speech (2.34 and 2.59 bps), respectively. However, it was also the most unreliable, which is reflected in its average selection time (Fig. 8.6a) and error rate (Fig. 8.6b). Participants took on average 1.04 seconds to select targets with dwell. Since the dwell time was set at 500 ms, this suggests that there were many target re-entries, where the cursor left the target before selecting it, thus had to re-enter, forcing participants to spend extra time with the method. Fig. 8.8 illustrates cursor traces from a random participant for the three selection methods, where one can see that dwell required much more target re-entries than speech and silent speech. Dwell also yielded a 33% and 57% higher error rates than speech and silent speech, which suggests that participants frequently dwelled outside the targets. Dwell's unreliability had an impact on user preference. Participants were least willing to use the method and found it to be the most physically and mentally demanding (Fig. 8.7). One participant (male, 28 years) commented, *"Dwell was the most difficult because it was causing eye fatigue"*. This suggests that dwell can be useful in short-term use, but is likely to affect user performance, preference, and comfort in extended use. Silent speech was the second best performed selection method in terms of throughput. A Tukey-Kramer test found its throughput to be significantly better than speech. Silent speech was also the most accurate. A Tukey-Kramer test identified its error rate to be significantly lower than both dwell and speech (36% and 15% lower, respectively). Participants were also willing to use the method the most on their computers. They found it slightly more physically and mentally demanding than speech (Fig. 8.7b), but this effect was not statistically significant. These results identify silent speech as an effective selection method in eye-gaze pointing.

8.6 User Study 2: Screen Location

We conducted a user study to inform the design of the final study. Its purpose was to identify the most effective screen areas for eye-gaze pointing, in terms of throughput, pointing time, and error rate, which can essentially help designing more effective interactive systems for eye tracking.

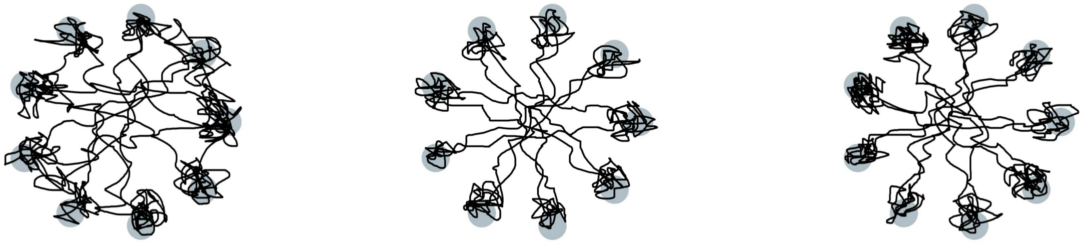
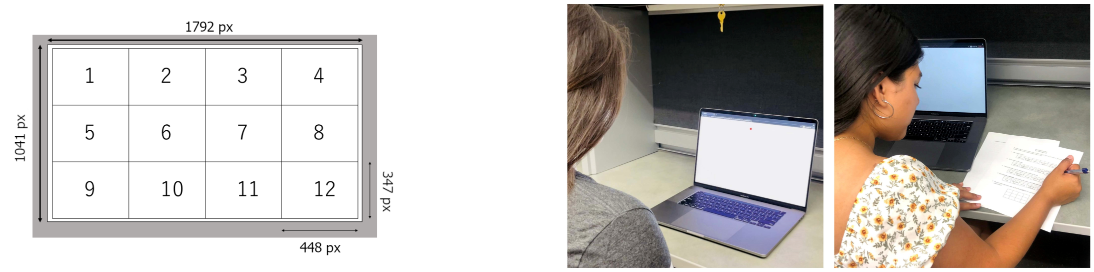

(a) Dwell (b) Speech (c) Silent speech

Figure 8.8: Cursor trace examples for the three selection methods (A:520 × W:70 pixels).

8.6.1 Participants

Twelve volunteers (M = 27.75 years, SD = 4.11) participated in the second study (Fig. 8.9b). None of them participated in the first study. Six of them identified themselves as women and six as men. Four of them wore corrective eyeglasses. None of them had experience with an eye-gaze-based system. They all received US $15 for volunteering.

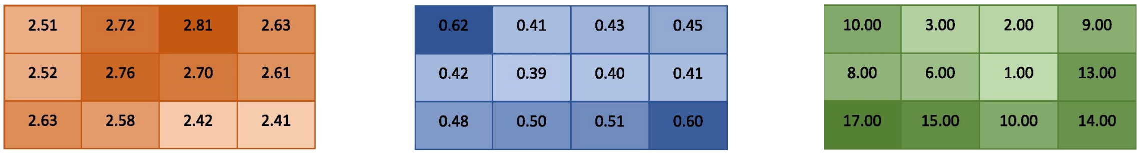


(a) (b)

Figure 8.9: (a) The twelve zones used in the second study and (b) two participants taking part in the study.

8.6.2 Apparatus, Design, & Procedure

The study used the apparatus described in Section 8.1.2. To investigate the most effective screen areas, the 1792×1041 display area (excluding the dock and the menu bar) was divided into 12 equal 448×347 pixels zones (Fig. 8.9a). The application displayed circular targets (35 pixels in diameter) at random locations in the zones for the participants to select using silent speech command. The study used the following within-subjects design: 12 participants × 12 zones × 12 targets per zone = 1,728 targets. The independent variable was "zone" and dependent variables were throughput, pointing time, and error rate (Section 8.5.2). This study used the same procedure as the first study (Section 8.5.3) except for the task. In this study, participants performed the point-select tasks by pointing at a target using eye tracking then selecting the target using the silent speech command *Select*. A sequence of trials consisted of 12 circular targets (35 pixels in diameter) per zone. The targets were presented at random locations in the zones (Fig. 8.9b). Hence, all trials had the same width (W) but different amplitudes (A). Upon completion of all trials, participants completed a short questionnaire where they could rate the difficulty levels of the 12 zones on a 5-point Likert scale.



2.51	2.72	2.81	2.63
2.52	2.76	2.70	2.61
2.63	2.58	2.42	2.41

(a) Throughput (bps)

0.62	0.41	0.43	0.45
0.42	0.39	0.40	0.41
0.48	0.50	0.51	0.60

(b) Pointing time (seconds)

10.00	3.00	2.00	9.00
8.00	6.00	1.00	13.00
17.00	15.00	10.00	14.00

(c) Error rate (%)

Figure 8.10: Average throughput, pointing time, and error rate per zone.

8.6.3 Results & Discussion

A complete study session took about 40–60 minutes, including demonstration, questionnaires, and breaks. A Shapiro-Wilk test revealed that the response variable residuals were normally distributed. A Mauchly's test indicated that the variances of populations were equal. Hence, we used a repeated-measures ANOVA for the quantitative within-subjects factors. We did not identify any effects of the between-subjects factors, namely age, gender, and corrective eyeglasses.

An ANOVA identified a significant effect of zone on throughput ($F_{11,121} = 4.37, p < .0001$). A Tukey-Kramer test identified three distinct groups: {1, 11, 12}, {4, 5, 7, 8, 9, 10}, and {2, 3, 6}, from the worst to the best performed zones. There was also a significant effect of zone on pointing time ($F_{11,121} = 8.93, p < .0001$). A Tukey-Kramer test identified three distinct groups: {1, 10, 11, 12}, {2, 3, 4, 5, 8, 9}, and {6, 7}, from the slowest to the fastest performed zones. An ANOVA also identified a significant effect on error rate ($F_{11,121} = 4.16, p < .0001$). A Tukey-Kramer test identified three distinct groups: {9}, {1, 4, 5, 8, 10, 11}, and {2, 3, 6, 7}, from the least to the most accurate zones. Fig. 8.10 illustrates these.

In summary, the study identified the central zones {2, 3, 6, 7} as the most accurate and the fastest. The top corners and bottom zones {1, 4, 9, 10, 12} were the most error prone and the slowest. The remaining zones {5, 8, 11} performed moderately well. User responses to the post-study questionnaire mirrored the quantitative data. We speculate, this is due to the increase in participants' viewing angle when looking at the top corners and bottom zones. Prior work showed that eye tracking systems achieve the best accuracy at narrow visual angles and even a slight increase in visual angles can increase gaze errors significantly [151]. Participants also expressed their enthusiasm about the system. One participant (male, 29 years) wrote, *"The technology felt good. It will be helpful to disable people to simplify their life"*. Another participant (female, 28 years) commented, *"This could be useful in self-checkout kiosk"*.

8.7 Menu Selection with Eye-Gaze and Silent Speech

We designed a method for menu selection with silent speech and gaze pointing. It facilitates the selection of small targets from a grid by adopting the *target gravity* metaphor from traditional graphical user interfaces [216, 34] and using six silent speech commands for cursor positioning and target selection. Target gravity uses a snap-to effect [216] that automatically moves the cursor to a target's center when it is within 10 pixels of the target, and then remains locked on the target until the gaze path exceeds 10 pixels or the user silently speaks the release command. We used this behavior because cursor drift and jitter during fixation due to involuntary eye movements causes irritation and affects performance [206]. The 10 pixels threshold was used because it felt the most natural in multiple lab trials. The method uses two silent commands to select and close/release targets, and four commands for directional movements of the cursor (Table 8.2). Fig. 8.11 illustrates a menu selection scenario with the proposed system.

Table 8.2: The six silent commands and corresponding actions used in the proposed menu section method.

Command	**Direction**	**Action**
Select		Selects the current item
Right	Horizontal	Moves the cursor to the right item. If there are no items on the right of the current item, the cursor is moved to the first item in the menu
Left	Horizontal	Moves the cursor to the left item. If there are no items on the left of the current item, the cursor is moved to the last item in the menu
Top	Vertical	Moves the cursor one item above the current item. If there are no items above the current item, the cursor is moved to the last item in the menu
Bottom	Vertical	Moves the cursor one item below the current item. If there are no items below the current item, the cursor is moved to the first item in the menu
Close		Unlocks the cursor by releasing target gravity

8.8 User Study 3: Menu Selection

We conducted a study to compare the silent speech-based selection method with and without menu selection commands.

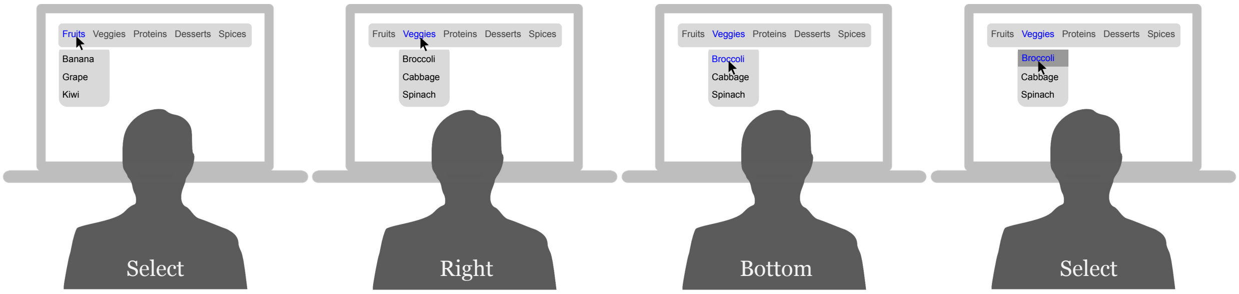

Fruits Veggies Proteins Desserts Spices
Banana
Grape
Kiwi
Select
Fruits Veggies Proteins Desserts Spices
Broccoli
Cabbage
Spinach
Right
Fruits Veggies Proteins Desserts Spices
Broccoli
Cabbage
Spinach
Bottom
Fruits Veggies Proteins Desserts Spices
Broccoli
Cabbage
Spinach
Select

Figure 8.11: A menu selection scenario with the proposed method. To select "Broccoli", the user starts scanning the horizontal menu from the left. The system locks the cursor on the first item when the gaze is within 10 pixels of the item. The user silently speaks the command "Select" to expand the current menu (display the sub-menu). The user silently speaks "Right" to move the cursor horizontally to the next item. The user locates the target, silently speaks "Bottom" to move the cursor to the target below the current item, then silently speaks "Select" to select the target.

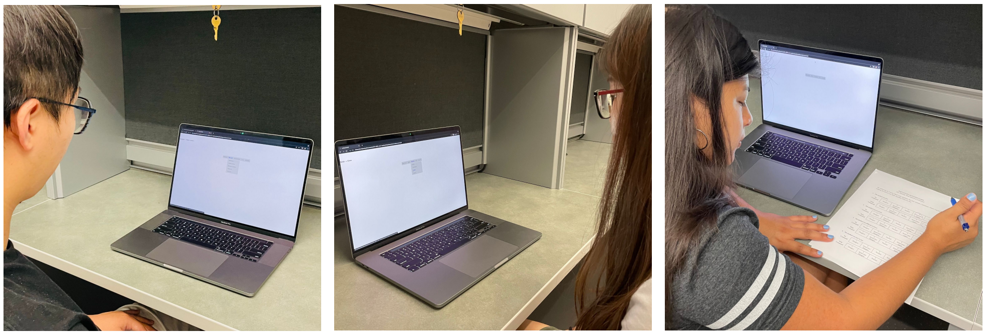


Figure 8.12: Three participants taking part in the final user study.

8.8.1 Participants & Apparatus

Twelve volunteers took part in the study. Neither of them participated in the first study. Their age ranged from 22 to 36 years (M = 28.25, SD = 4.63). Six of them identified themselves as women and six as men. Two of them wore corrective eyeglasses. None of them had experience with an eye-gaze-based system. Each of them received US $15 for volunteering in the study. The study used the apparatus described in Section 8.1.2.

8.8.1.1 Task Selection

We customized the web application to display four menus (one at a time) categorizing different types of animals, food, popular books, and famous people. Simple categories were used to assure that the selection tasks do not require specialized knowledge. All categories had five vertical menu items. The vertical sub-menus under the horizontal menus had either three, four, or seven items. We did not use more than seven items per sub-menu to avoid memory overload [197]. Fifteen random targets were selected per category: five with target distances between 2–5, five between 6–7, and five between 8–12. Target distance signifies the total number of horizontal and vertical items before the target. Horizontal items are counted from left to right and vertical items are counted from top to bottom since research revealed that users tend to scan items from left-to-right and top-to-bottom [48]. The menus were designed following the macOS guidelines [196] to provide a familiar look-and-feel. Each menu item was 150×38 pixels. Current items were highlighted in a blue font (Fig. 8.13) and selected items were highlighted in a dark gray background (Fig. 8.11).

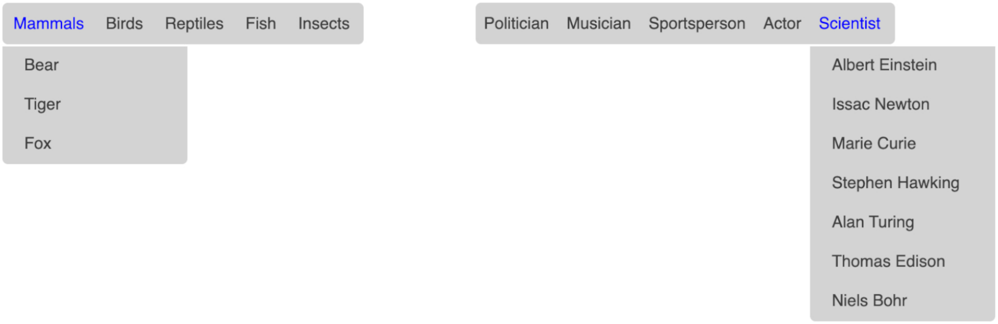


Figure 8.13: Examples of two menus categorizing different types of animals and famous people.

8.8.2 Design & Procedure

The study used the following within-subjects design: 12 participants × 2 methods (command, menu command, counterbalanced) × 2 unique menus per method × 15 tasks per menu = 720 menu selection tasks. The independent variable was “method” and dependent variables were as follows:

- **Task completion time** (seconds) represents the average time users took to perform a menu selection task.
- **Look-back rate** (%) represents the average percentage of times users entered a correct sub-menu, then left to explore the other sub-menus. This occurred when users were

unable to locate a target despite entering the correct sub-menu, thus explored other sub-menus to find the target.

- **Error rate** (%) signifies the average percentage of incorrect menu selections per method (%), where users either selected an incorrect item or performed a selection task outside the menu.

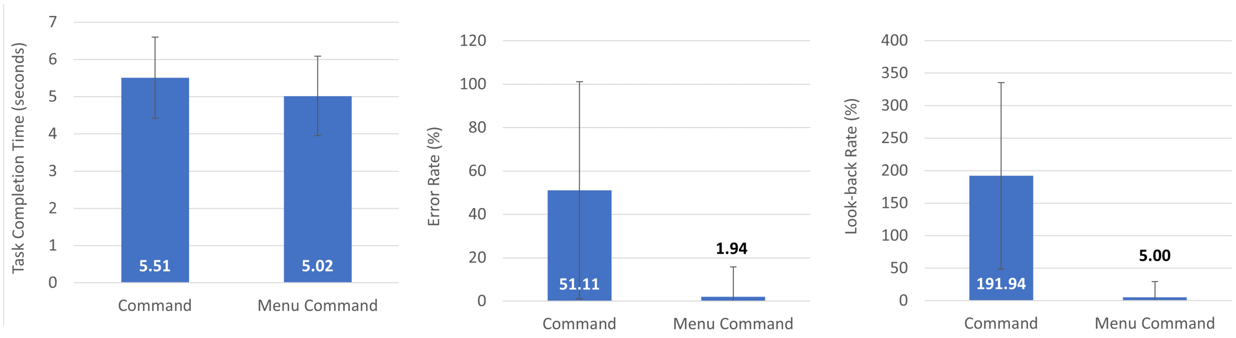
The study used the same procedure as the previous studies (Section 8.5.3). During practice, participants selected two items using both methods (with and without menu commands) from a menu that was not used in the study. Once they were familiar with the methods, they started the main study, where they performed 15 target selection tasks per menu category with both methods. In the menu command condition, participants used the commands presented in Table 8.2 for navigation and selection. In the command condition, they used eye-gaze exclusively for positioning the cursor and the "Select" command to select a target. Tasks with different distances were presented on the screen in a random order. Considering some participants could be more familiar with the categories than the others, the application also displayed the complete target path. For example, for the scenario depicted in Fig. 8.11, the application displayed the task as "Select Veggies > Broccoli", indicating that the participants first have to go to the "Veggies" sub-menu then select "Broccoli". Two menu categories were assigned to each method in a counterbalanced order. We did not use the same menu categories with both methods to avoid any potential effects of knowledge (using the knowledge acquired in one condition to achieve the goals in another). Participants were instructed to select the targets as fast and accurate as possible. Error correction was not required. Timing started from the moment they lifted their gaze from the presented task to the moment a sub-menu item was selected. We enforced a 2-minute break after each menu category and a 5-minute break after each condition to avoid the effect of fatigue. Upon completion of the study, participants completed a custom and the NASA-TLX questionnaire [114] to rate the methods' perceived performance, usability, and workload.

8.8.3 Results

A complete study session took about 40–60 minutes, including demonstration, questionnaires, and breaks. A Shapiro-Wilk test revealed that the response variable residuals were normally distributed. A Mauchly's test indicated that the variances of populations were equal. Hence, we used a repeated-measures ANOVA for the quantitative within-subjects factors. We used a Wilcoxon Signed-Rank test for the questionnaire data. [18] We did not identify any effects of the between-subjects factors, namely age, gender, and the use of corrective eyeglasses.

8.8.3.1 Task Completion Time

An ANOVA identified a significant effect of method on task completion time ($F_{1,11} = 18.84, p < .005$). Average task completion time for command and menu command were



(a) Task completion time (seconds)

(b) Error rate (%)

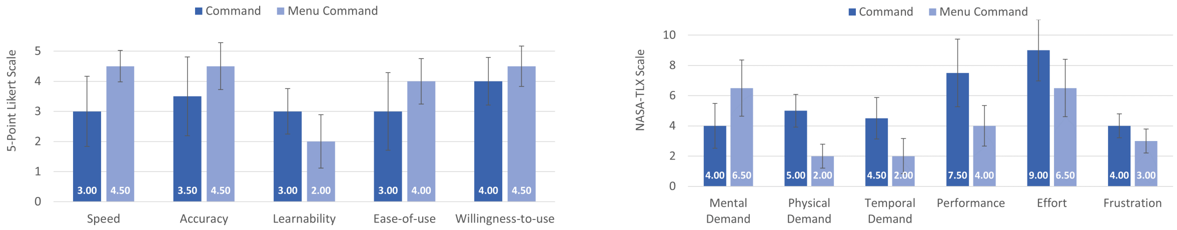
(c) Look-back rate (%)

Figure 8.14: Average task completion time, error rate, and look-back rate for the two investigated methods. Error bars represent ±1 standard deviation (SD).

5.51 (SD = 1.09) and 5.02 seconds (SD = 1.07), respectively (Fig. 8.14a).

8.8.3.2 Error & Look-Back Rates

An ANOVA identified a significant effect of method on error rate ($F_{1,11} = 265.30, p < .0001$). Average error rate for command and menu command were 51.11% (SD = 50.06) and 1.94% (SD = 13.83), respectively (Fig. 8.14b). An ANOVA also identified a significant effect on look-back rate ($F_{1,11} = 1113.35, p < .0001$). Average look-back rate for command and menu command were 191.94% (SD = 143.25) and 5.00% (SD = 24.24), respectively (Fig. 8.14c).



(a) Usability questionnaire

(b) NASA-TLX questionnaire

Figure 8.15: Median willingness-to-use and physical and mental effort of the examined selection methods. Error bars represent ±1 standard deviation (SD).

8.8.3.3 User Feedback

A Wilcoxon Signed-Rank test identified a significant effect of method on perceived speed ($z = -2.45, p < .05$), perceived accuracy ($z = -2.16, p < .05$), and ease-of-use ($z = -2.22, p <$

.05). However, there was no significant effect on learnability ($z = -1.06, p = .29$) and willingness-to-use ($z = -1.3, p < .19$). Fig. 8.15a presents median perceived performance and usability ratings of both methods.

8.8.3.4 Perceived Workload

A Wilcoxon Signed-Rank test identified a significant effect of method on mental demand ($z = -2.61, p < .01$), physical demand ($z = -2.82, p < .01$), temporal demand ($z = -2.83, p < .01$), performance ($z = -2.62, p < .01$), effort ($z = -2.95, p < .005$), and frustration ($z = -2.98, p < .005$). Fig. 8.15b presents median perceived workload ratings of both methods.

8.9 Key Findings and Design Recommendations

Below, we summarize the key findings of this work and make design recommendations.

- Silent command is a fast and effective alternative to dwell and speech-based selection methods in eye-gaze pointing, especially when the vocabulary is relatively small. We recommend designers to present a small number of options at a time to limit the total number of possible user responses to ten or less.

- We recommend against using dwell for tasks that require using the eyes for extended period of time since it tend to affect user performance, preference, and comfort.

- When designing eye-gaze-based interactive systems, we recommend placing the most important and frequently used interactive elements at the center or around the two sides of the display. Avoiding the top corners and the bottom is recommended as they are usually the slowest and the most error prone.

- We recommend using silent command for menu selection with eye-gaze pointing as it is a more private and secure option and significantly increases users' confidence in selecting the correct option. Besides, vertical and horizontal menus are equally effective in eye-gaze pointing with silent speech.

8.10 Discussion

Eye-gaze with menu command yielded about 9% faster task completion time than the baseline (the method without menu command). Most impressively, it reduced error rates by 96%. The baseline's 51% error rate (compared to menu command's 2%) suggests that roughly one in every two targets were incorrectly selected (Fig. 8.14b). Menu command also yielded 97% lower look-back rate than the baseline (Fig. 8.14c). The baseline yielded a 192% look-back rate, which suggests that most of the times participants were not confident that they were in the correct sub-menu, thus left to explore the other sub-menus. This behavior is particularly interesting since the study tasks did not require participants to explore the sub-menus to locate a target, instead displayed the exact path. The fact that participants did not look-back as much while using the menu command suggests that it increased their confidence in performing the tasks. A deeper analysis failed to identify an effect of horizontal and vertical (sub-)menu items on performance. This contradicts a prior work that found horizontal pointing to be about 18% more error prone than vertical pointing [152]. We also failed to identify any relationship between target distance and performance. This contradicts a prior finding that users' response time is an approximately linear function of serial position in the menu [214]. Our findings, however, are in line with a follow-up work that failed to replicate [214]'s findings and argued that visual search and cursor movement strategies employed by actual users cannot be characterized easily [48].

Participants perceived the proposed method significantly faster and more accurate than the baseline (Fig. 8.15a). A participant (female, 25 years) commented, *"I think with commands [gaze-based menu selection] is more reliable"*. They also found the method significantly easier to use. They felt that both methods were easy to learn. Interestingly, their ratings were also comparable in terms of willingness to use. We believe, the exclusion of error correction from the study protocol influenced this—their response could have been different if they were forced to correct all incorrect selections. Participants found the proposed method mentally, physically, and temporally less demanding than the baseline (Fig. 8.15b). They also felt that the method was better performed, required less effort, and caused less frustration than the baseline.

Chapter 9

Conclusion and Future Scope

This dissertation focused on investigating users' impression towards using silent speech input method on mobile devices from social acceptance, error tolerance, and feedback design perspectives, followed by the investigation of technical challenges associated with existing silent speech recognition models and solutions. Towards this, we first conducted an online survey to explore users' attitudes towards the speech and silent speech input methods with a particular focus on social acceptance. In a survey, we found out that in general people preferred using silent speech input over the traditional speech input. We also observed that users were more comfortable using silent speech input in different public and private locations but expressed their concerns about input recognition, privacy, and security issues.

Consequently, we conducted a study examining users' error tolerance with speech and silent speech input methods. Results reveal that users are willing to tolerate more errors with silent speech input than speech input as it offers a higher degree of privacy and security. Inspired by the findings, we further investigate suitable feedback method for silent speech input. Results show that users find both a commonly used video and an abstract (a blinking dot) feedback effective but the latter significantly more private, more secure, and less intrusive than the video feedback. We learned that designing solutions for silent speech input requires careful consideration of various factors and privacy concerns as well as people's tolerance towards using it on computer systems.

As a step forward, we attempt to address the technological limitations of existing silent speech recognition models. Towards this end, we develop LipType, an optimized silent speech recognition model for improved speed and accuracy. In an evaluation, LipType reduced the word error rate by 47% compared to the state-of-the-art silent speech recognition model.

We then develop an independent repair model that processes video input for poor lighting conditions, when applicable, and corrects potential errors in output for increased accuracy. In an evaluation, the repair model demonstrated it's effectiveness with various speech and silent speech recognition models. On average, speech and silent speech models showed 32% and 57% reduction in word error rates, respectively, without compromising the overall computation time. The findings confirm that the model can be used independently with a range of recognizers.

We conducted another study to explore how users interact with silent speech-based methods. Results revealed that native users speak about 8% faster than non-native users, but both groups slow down at comparable rates (34–40%) when interacting with these methods, mostly to increase their accuracy rates. A follow-up study confirms that slowing down does improve the accuracy of these methods. Both methods yield the best accuracy rates when speaking at 0.75x of the actual speaking rate. A post-hoc error analysis revealed that speech and silent speech methods and native and non-native speakers are susceptible to different types of errors. Native speakers committed 59% lower insertion errors, 11% higher deletion errors, and 38% higher substitution errors than non-native speakers with speech. Whereas with silent speech, they committed 81% lower insertion errors, 42% higher deletion errors, and 16% higher substitution errors than non-native speakers. The findings of this study highlight the importance of considering speaking rate in speech and silent speech-based interfaces. While designing interfaces for these methods, the recognition algorithms must be optimized for varying speaking rates and the characteristics of native and non-native speakers. Error analysis presented in this work could be used to identify areas that require extra effort to increase the respective method's accuracy rates. The findings could also provide guidance to users on improving speech and silent speech input performance.

Finally, we studied the feasibility of using silent speech as a hands-free selection method in eye-gaze pointing on computer systems. We first propose a stripped-down image-based model that can recognize a small number of silent commands almost as fast as state-of-the-art speech recognition models. We then compare it with other hands-free selection methods (dwell, speech) in a Fitts' law study. Results revealed that speech and silent speech are comparable in throughput and selection time, but the latter is significantly more accurate than the other methods. A follow-up study revealed that target selection around the center of a display is significantly faster and more accurate, while around the top corners and the bottom are slower and error prone. We then present a method for selecting menu items with eye-gaze and silent speech. A study revealed that it significantly reduces task completion time and error rate.

In the future, we will extend the work to support more than ten silent speech commands. We will also investigate the possibility of using targeted commands, where the user silently speaks a specific menu item to select it rather than using directional commands. Finally, we will explore different error correction mechanisms to enhance the usability of the method. We envision numerous opportunities for future extension of this work. The proposed mouth aspect ratio-based model could be trained with people with muteness and speech disorders to enable hands-free interaction with computer systems using a set of custom commands or even lip gestures. The model could also be used with conversational agents, e.g., chatbots. Since they usually ask close-ended questions to limit the number of possible answers, the system has to disambiguate the input from a small number of samples at a time, comparable to the menu selection concept presented here. Eye tracking and silent commands could also be used in other application domains, such as in virtual reality or in automotive user interfaces.

Bibliography

[1] Martín Abadi et al. "Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems". In: *arXiv:1603.04467 [cs]* (Mar. 2016). arXiv: 1603.04467. URL: http://arxiv.org/abs/1603.04467 (visited on 09/10/2020).

[2] Ossama Abdel-Hamid et al. "Convolutional Neural Networks for Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22.10 (Oct. 2014). Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1533–1545. ISSN: 2329-9304. DOI: 10.1109/TASLP.2014.2339736.

[3] Martine Adda-Decker and Lori Lamel. "Do Speech Recognizers Prefer Female Speakers?" In: *Ninth European Conference on Speech Communication and Technology*. 2005.

[4] Mahmoud Afifi et al. "Learning to Correct Overexposed and Underexposed Photos". en. In: *arXiv:2003.11596 [cs, eess]* (Mar. 2020). arXiv: 2003.11596. URL: http://arxiv.org/abs/2003.11596 (visited on 09/09/2020).

[5] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. "Deep Lip Reading: A Comparison of Models and an Online Application". In: *arXiv:1806.06053 [cs]* (June 2018). arXiv: 1806.06053. URL: http://arxiv.org/abs/1806.06053 (visited on 09/10/2020).

[6] Abien Fred Agarap. "Deep Learning using Rectified Linear Units (ReLU)". In: *ArXiv* (Feb. 7, 2019). URL: http://arxiv.org/abs/1803.08375 (visited on 08/05/2020).

[7] Ayush Agarwal et al. "Comparing Two Webcam-Based Eye Gaze Trackers for Users with Severe Speech and Motor Impairment". en. In: *Research into Design for a Connected World*. Ed. by Amaresh Chakrabarti. Vol. 135. Series Title: Smart Innovation, Systems and Technologies. Singapore: Springer Singapore, 2019, pp. 641–652. DOI: 10.1007/978-981-13-5977-4_54. (Visited on 08/24/2021).

[8] David Ahlström, Khalad Hasan, and Pourang Irani. "Are You Comfortable Doing That? Acceptance Studies of around-Device Gestures in and for Public Settings". In: *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices & Services*. MobileHCI '14. Toronto, ON, Canada: Association for Computing Machinery, 2014, pp. 193–202. ISBN: 9781450330046. DOI: 10.1145/2628363.2628381.

[9] Fouad Alallah et al. "Crowdsourcing vs Laboratory-Style Social Acceptability Studies? Examining the Social Acceptability of Spatial User Interactions for Head-Worn Displays". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–7. ISBN: 9781450356206. DOI: 10.1145/3173574.3173884.

[10] Fouad Alallah et al. “Performer vs. Observer: Whose Comfort Level Should We Consider When Examining the Social Acceptability of Input Modalities for Head-Worn Display?” In: *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*. VRST ’18. Tokyo, Japan: Association for Computing Machinery, 2018. ISBN: 9781450360869. DOI: 10.1145/3281505.3281541.

[11] Ohoud Alharbi et al. “WiseType: A Tablet Keyboard with Color-Coded Visualization and Various Editing Options for Error Correction”. In: *Proceedings of the 45th Graphics Interface Conference on Proceedings of Graphics Interface 2019*. GI’19. Kingston, Canada: Canadian Human-Computer Communications Society, 2019. ISBN: 9780994786845. DOI: 10.20380/GI2019.04.

[12] Alexandre Allauzen. “Error Detection in Confusion Network”. In: *INTERSPEECH*. 2007.

[13] Cyril Allauzen and Michael Riley. “Bayesian Language Model Interpolation for Mobile Speech Input”. In: *Interspeech 2011*. 2011, pp. 1429–1432.

[14] Ibrahim Almajai et al. “Improved Speaker Independent Lip Reading Using Speaker Adaptive Training and Deep Neural Networks”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. Mar. 2016, pp. 2722–2726. DOI: 10.1109/ICASSP.2016.7472172.

[15] Anastasios Anastasakos, Richard Schwartz, and Han Shu. “Duration modeling in large vocabulary speech recognition”. In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. IEEE. 1995, pp. 628–631.

[16] Aries Arditi and Jianna Cho. “Serifs and Font Legibility”. In: *Vision research* 45.23 (2005), pp. 2926–2933.

[17] Tarik Arici, Salih Dikbas, and Yucel Altunbasak. “A Histogram Modification Framework and Its Application for Image Contrast Enhancement”. In: *IEEE Transactions on Image Processing* 18.9 (Sept. 2009). Conference Name: IEEE Transactions on Image Processing, pp. 1921–1935. ISSN: 1941-0042. DOI: 10.1109/TIP.2009.2021548.

[18] Ahmed Sabbir Arif. “Statistical Grounding”. In: *Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice*. 1st ed. New York, NY, USA: Association for Computing Machinery, 2021, pp. 59–99. ISBN: 978-1-4503-9029-3. URL: https://doi.org/10.1145/3447404.3447410.

[19] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. “Analysis of Text Entry Performance Metrics”. In: *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)*. Sept. 2009, pp. 100–105. DOI: 10.1109/TIC-STH.2009.5444533.

[20] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. “Predicting the Cost of Error Correction in Character-Based Text Entry Technologies”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’10. Atlanta, Georgia, USA: Association for Computing Machinery, 2010, pp. 5–14. ISBN: 9781605589299. DOI: 10.1145/1753326.1753329.

[21] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. “User Adaptation to a Faulty Unistroke-Based Text Entry Technique by Switching to an Alternative Gesture Set”. In: *Proceedings of Graphics Interface 2014*. GI ’14. event-place: Montreal, Quebec, Canada. Toronto, Ont., Canada, Canada: Canadian Information Processing Society, 2014, pp. 183–192. ISBN: 978-1-4822-6003-8. URL: http://dl.acm.org/citation.cfm?id=2619648.2619679 (visited on 11/02/2019).

[22] Ahmed Sabbir Arif and Wolfgang Stuerzlinger. “User Adaptation to a Faulty Unistroke-Based Text Entry Technique by Switching to an Alternative Gesture Set”. In: *Proceedings of Graphics Interface 2014*. GI ’14. Montreal, Quebec, Canada: Canadian Information Processing Society, 2014, pp. 183–192. ISBN: 9781482260038.

[23] 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. New York, NY, USA: Association for Computing Machinery, Mar. 2010, pp. 339–345. ISBN: 978-1-60558-994-7. DOI: 10.1145/1743666.1743742. (Visited on 09/01/2021).

[24] Tim Ashwell and Jesse R Elam. “How Accurately Can the Google Web Speech API Recognize and Transcribe Japanese L2 English Learners’ Oral Production?.” In: *Jalt Call Journal* 13.1 (2017), pp. 59–76.

[25] Yannis M. Assael et al. “Lipnet: End-to-End Sentence-Level Lipreading”. In: *arXiv:1611.01599 [cs]* (Dec. 2016). arXiv: 1611.01599. URL: http://arxiv.org/abs/1611.01599 (visited on 09/10/2020).

[26] Mauro Avila Soto and Markus Funk. “Look, a guidance drone! Assessing the Social Acceptability of Companion Drones for Blind Travelers in Public Spaces”. In: *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. ASSETS ’18. New York, NY, USA: Association for Computing Machinery, Oct. 8, 2018, pp. 417–419. ISBN: 978-1-4503-5650-3. DOI: 10.1145/3234695.3241019. (Visited on 09/06/2020).

[27] Melissa M. Baese-Berk and Tuuli H. Morrill. “Speaking Rate Consistency in Native and Non-Native Speakers of English”. In: *The Journal of the Acoustical Society of America* 138.3 (Sept. 2015). Publisher: Acoustical Society of America, EL223–EL228. ISSN: 0001-4966. DOI: 10.1121/1.4929622. URL: https://asa.scitation.org/doi/full/10.1121/1.4929622 (visited on 07/04/2021).

[28] Dzmitry Bahdanau et al. “End-to-End Attention-Based Large Vocabulary Speech Recognition”. In: *arXiv:1508.04395 [cs]* (Mar. 2016). arXiv: 1508.04395. URL: http://arxiv.org/abs/1508.04395 (visited on 09/10/2020).

[29] L. Bahl and F. Jelinek. “Decoding for Channels with Insertions, Deletions, and Substitutions with Applications to Speech Recognition”. In: *IEEE Transactions on Information Theory* 21.4 (July 1975). Conference Name: IEEE Transactions on Information Theory, pp. 404–411. ISSN: 1557-9654. DOI: 10.1109/TIT.1975.1055419.

[30] Monique Faye Baier and Michael Burmester. “Not Just About the User: Acceptance of Speech Interaction in Public Spaces”. In: *Proceedings of Mensch Und Computer 2019*. MuC’19. Hamburg, Germany: Association for Computing Machinery, 2019, pp. 349–359. ISBN: 9781450371988. DOI: 10.1145/3340764.3340801.

[31] Brandon Ballinger et al. "On-Demand Language Model Interpolation for Mobile Speech Input". In: *Interspeech.* 2010, pp. 1812–1815.

[32] Jess Bartels et al. "Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex". In: *Journal of Neuroscience Methods* (2008). DOI: 10.1016/j.jneumeth.2008.06.030.

[33] Youssef Bassil and Paul Semaan. "Asr Context-Sensitive Error Correction Based on Microsoft N-Gram Dataset". In: *arXiv:1203.5262 [cs]* (Mar. 2012). arXiv: 1203.5262. URL: http://arxiv.org/abs/1203.5262 (visited on 09/10/2020).

[34] Scott Bateman et al. *Investigation of Targeting-Assistance Techniques for Distant Pointing with Relative Ray Casting.* Tech. rep. 2009-03. Saskatoon, SK, Canada: University of Saskatchewan, 2009, p. 10.

[35] Richard Bates and Howell Istance. "Zooming Interfaces! Enhancing the Performance of Eye Controlled Pointing Devices". In: *Proceedings of the fifth international ACM conference on Assistive technologies.* Assets '02. New York, NY, USA: Association for Computing Machinery, July 2002, pp. 119–126. ISBN: 978-1-58113-464-3. DOI: 10.1145/638249.638272. (Visited on 08/30/2021).

[36] Helen L. Bear and Richard Harvey. "Alternative Visual Units for an Optimized Phoneme-Based Lipreading System". In: *Applied Sciences* 18 (2019), p. 3870. DOI: 10.3390/app9183870.

[37] Roman Bednarik, Tersia Gowases, and Markku Tukiainen. "Gaze Interaction Enhances Problem Solving: Effects of Dwell-Time Based, Gaze-Augmented, and Mouse Interaction on Problem-Solving Strategies and User Experience". en. In: *Journal of Eye Movement Research* 3.1 (June 2009). Number: 1. ISSN: 1995-8692. DOI: 10.16910/jemr.3.1.3. (Visited on 08/26/2021).

[38] T. R. Beelders and P. J. Blignaut. "Measuring the Performance of Gaze and Speech for Text Input". In: *Proceedings of the Symposium on Eye Tracking Research and Applications.* ETRA '12. New York, NY, USA: Association for Computing Machinery, Mar. 2012, pp. 337–340. ISBN: 978-1-4503-1221-9. DOI: 10.1145/2168556.2168631. (Visited on 09/01/2021).

[39] Tara S. Behrend et al. "The Viability of Crowdsourcing for Survey Research". en. In: *Behavior Research Methods* 43.3 (Mar. 2011), p. 800. ISSN: 1554-3528. DOI: 10.3758/s13428-011-0081-0. (Visited on 09/16/2020).

[40] Linda Bell and Joakim Gustafson. "Interaction with an Animated Agent in a Spoken Dialogue System". In: *Sixth European Conference on Speech Communication and Technology.* 1999.

[41] Andre-Pierre Benguerel and Margaret Kathleen Pichora-Fuller. "Coarticulation Effects in Lipreading". In: *Journal of Speech, Language, and Hearing Research* 25.4 (1982), pp. 600–607.

[42] Godfred O Boateng et al. "Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer". In: *Frontiers in public health* 6 (2018), p. 149.

[43] Maria Borgestig et al. “Eye Gaze Performance for Children with Severe Physical Impairments Using Gaze-Based Assistive Technology—a Longitudinal Study”. In: *Assistive Technology* 28.2 (Apr. 2016), pp. 93–102. ISSN: 1040-0435. DOI: 10.1080/10400435.2015.1092182. (Visited on 08/24/2021).

[44] Holly P Branigan et al. “Linguistic Alignment Between People and Computers”. In: *Journal of pragmatics* 42.9 (2010), pp. 2355–2368.

[45] John Brooke. “SUS: A Quick and Dirty Usability Scale”. In: *Usability evaluation in industry* (1996), p. 189.

[46] Christopher Ralph Brown. “Automatic Pruning of Grammars in a Multi-Application Speech Recognition Interface”. en. US20080059195A1. Mar. 2008. URL: https://patents.google.com/patent/US20080059195/en (visited on 09/10/2020).

[47] Jonathan S. Brumberg et al. “Brain-Computer Interfaces for Speech Communication”. In: *Speech Communication* 52.4 (Apr. 2010), pp. 367–379. ISSN: 0167-6393. DOI: 10.1016/j.specom.2010.01.001. (Visited on 09/10/2020).

[48] Michael D. Byrne et al. “Eye Tracking the Visual Search of Click-down Menus”. In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. CHI ’99. New York, NY, USA: Association for Computing Machinery, May 1999, pp. 402–409. ISBN: 978-0-201-48559-2. DOI: 10.1145/302979.303118. (Visited on 08/30/2021).

[49] Ronald P Carver. “Word Length, Prose Difficulty, and Reading rate”. In: *Journal of Reading Behavior* 8.2 (1976), pp. 193–203.

[50] Turgay Celik and Tardi Tjahjadi. “Contextual and Variational Contrast Enhancement”. In: *IEEE Transactions on Image Processing* 20.12 (Dec. 2011). Conference Name: IEEE Transactions on Image Processing, pp. 3431–3441. ISSN: 1941-0042. DOI: 10.1109/TIP.2011.2157513.

[51] Ishan Chatterjee, Robert Xiao, and Chris Harrison. “Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions”. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI ’15. New York, NY, USA: Association for Computing Machinery, Nov. 2015, pp. 131–138. ISBN: 978-1-4503-3912-4. DOI: 10.1145/2818346.2820752. (Visited on 08/28/2021).

[52] Chen Chen et al. “Learning to See in the Dark”. In: *arXiv:1805.01934 [cs]* (May 2018). arXiv: 1805.01934. URL: http://arxiv.org/abs/1805.01934 (visited on 09/10/2020).

[53] Stanley F. Chen and Joshua Goodman. “An Empirical Study of Smoothing Techniques for Language Modeling”. In: *Proceedings of the 34th annual meeting on Association for Computational Linguistics*. ACL ’96. USA: Association for Computational Linguistics, June 1996, pp. 310–318. DOI: 10.3115/981863.981904. (Visited on 09/10/2020).

[54] Wei Chen et al. “Asr Error Detection in a Conversational Spoken Language Translation System”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. ISSN: 2379-190X. May 2013, pp. 7418–7422. DOI: 10.1109/ICASSP.2013.6639104.

[55] Yi Cheng et al. "Why Doesn't It Work? Voice-Driven Interfaces and Young Children's Communication Repair Strategies". In: *Proceedings of the 17th ACM Conference on Interaction Design and Children*. IDC '18. Trondheim, Norway: Association for Computing Machinery, 2018, pp. 337–348. ISBN: 9781450351522. DOI: 10.1145/3202185.3202749.

[56] Joon Son Chung and Andrew Zisserman. "Learning to Lip Read Words by Watching Videos". en. In: *Computer Vision and Image Understanding* 173 (Aug. 2018), pp. 76–85. ISSN: 1077-3142. DOI: 10.1016/j.cviu.2018.02.001. (Visited on 09/10/2020).

[57] Joon Son Chung and Andrew Zisserman. "Lip Reading in Profile". In: *BMVC*. 2017. DOI: 10.5244/C.31.155.

[58] Joon Son Chung and Andrew Zisserman. "Lip Reading in the Wild". en. In: *Computer Vision – ACCV 2016*. Ed. by Shang-Hong Lai et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 87–103. ISBN: 978-3-319-54184-6. DOI: 10.1007/978-3-319-54184-6_6.

[59] Joon Son Chung and Andrew Zisserman. "Out of Time: Automated Lip Sync in the Wild". In: *ACCV Workshops*. 2016. DOI: 10.1007/978-3-319-54427-4_19.

[60] Joon Son Chung et al. "Lip Reading Sentences in the Wild". In: *arXiv:1611.05358 [cs]* (Jan. 2017). arXiv: 1611.05358. URL: http://arxiv.org/abs/1611.05358 (visited on 09/10/2020).

[61] Junyoung Chung et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *arXiv:1412.3555 [cs]* (Dec. 11, 2014). URL: http://arxiv.org/abs/1412.3555 (visited on 08/05/2020).

[62] C. Cieri, D. Miller, and K. Walker. "The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text". In: *LREC*. 2004.

[63] Leigh Clark et al. "The State of Speech in HCI: Trends, Themes and Challenges". In: *Interacting with Computers* 31.4 (2019), pp. 349–371. ISSN: 1873-7951. DOI: 10.1093/iwc/iwz016.

[64] Ronan Collobert, Awni Hannun, and Gabriel Synnaeve. "A Fully Differentiable Beam Search Decoder". In: *arXiv:1902.06022 [cs]* (Feb. 15, 2019). arXiv: 1902.06022. URL: http://arxiv.org/abs/1902.06022 (visited on 08/05/2020).

[65] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. "Wav2letter: An End-to-End Convnet-Based Speech Recognition System". In: *arXiv:1609.03193 [cs]* (Sept. 2016). arXiv: 1609.03193. URL: http://arxiv.org/abs/1609.03193 (visited on 09/10/2020).

[66] Martin Cooke et al. "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition". In: *The Journal of the Acoustical Society of America* 120.5 (Oct. 2006). Publisher: Acoustical Society of America, pp. 2421–2424. ISSN: 0001-4966. DOI: 10.1121/1.2229005. (Visited on 09/10/2020).

[67] F. Corno, L. Farinetti, and I. Signorile. "A Cost-Effective Solution for Eye-Gaze Assistive Technology". In: *Proceedings. IEEE International Conference on Multimedia and Expo*. Vol. 2. Aug. 2002, 433–436 vol.2. DOI: 10.1109/ICME.2002.1035632.

[68] Catia Cucchiarini, Helmer Strik, and Lou Boves. "Quantitative Assessment of Second Language Learners' Fluency by Means of Automatic Speech Recognition Technology". In: *The Journal of the Acoustical Society of America* 107.2 (2000), pp. 989–999.

[69] Catia Cucchiarini, Helmer Strik, and Lou Boves. "Quantitative Assessment of Second Language Learners' Fluency: Comparisons Between Read and Spontaneous Speech". In: *the Journal of the Acoustical Society of America* 111.6 (2002), pp. 2862–2873.

[70] Ronald Cumbal et al. ""You don't understand me!": Comparing ASR Results for L1 and L2 Speakers of Swedish". In: *Proc. Interspeech 2021*. 2021, pp. 4463–4467. DOI: 10.21437/Interspeech.2021-2140.

[71] Charles S. DaSalla et al. "Spatial Filtering and Single-Trial Classification of Eeg During Vowel Speech Imagery". In: *Proceedings of the 3rd International Convention on Rehabilitation Engineering & Assistive Technology*. i-CREATe '09. New York, NY, USA: Association for Computing Machinery, Apr. 2009, pp. 1–4. ISBN: 978-1-60558-792-9. DOI: 10.1145/1592700.1592731. (Visited on 09/10/2020).

[72] B. Denby and M. Stone. "Speech Synthesis from Real Time Ultrasound Images of the Tongue". In: *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. ISSN: 1520-6149. May 2004, pp. I–685. DOI: 10.1109/ICASSP.2004.1326078.

[73] B. Denby et al. "Prospects for a Silent Speech Interface Using Ultrasound Imaging". In: *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Vol. 1. ISSN: 2379-190X. May 2006, pp. I–I. DOI: 10.1109/ICASSP.2006.1660033.

[74] Li Deng and Xuedong Huang. "Challenges in Adopting Speech Recognition". In: *Communications of the ACM* 47.1 (Jan. 2004), pp. 69–75. ISSN: 0001-0782. DOI: 10.1145/962081.962108. (Visited on 09/10/2020).

[75] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. "In situ with bystanders of augmented reality glasses: perspectives on recording and privacy-mediating technologies". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, Apr. 26, 2014, pp. 2377–2386. ISBN: 978-1-4503-2473-1. URL: https://doi.org/10.1145/2556288.2557352 (visited on 09/06/2020).

[76] T Derwing and MJ Munro. "What Speaking Rates Do Non-Native Listeners Prefer?" In: *Applied Linguistics* 22.3 (Sept. 2001), pp. 324–337. ISSN: 0142-6001. DOI: 10.1093/applin/22.3.324. (Visited on 07/04/2021).

[77] Ali Diba et al. "Spatio-Temporal Channel Correlation Networks for Action Classification". In: *arXiv:1806.07754 [cs]* (Feb. 2019). arXiv: 1806.07754. URL: http://arxiv.org/abs/1806.07754 (visited on 09/10/2020).

[78] Alan Dix et al. *Human-Computer Interaction (3rd Edition)*. USA: Prentice-Hall, Inc., 2003. ISBN: 0130461091.

[79] Xuan Dong et al. "Fast Efficient Algorithm for Enhancement of Low Lighting Video". In: *2011 IEEE International Conference on Multimedia and Expo*. ISSN: 1945-788X. July 2011, pp. 1–6. DOI: 10.1109/ICME.2011.6012107.

[80] Heiko Drewes and Albrecht Schmidt. "Interacting with the Computer Using Gaze Gestures". en. In: *Human-Computer Interaction – INTERACT 2007*. Ed. by Cécilia Baranauskas et al. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2007, pp. 475–488. DOI: 10.1007/978-3-540-74800-7_43.

[81] Stefania Druga et al. “”Hey Google is It OK If I Eat You?”: Initial Explorations in Child-Agent Interaction”. In: *Proceedings of the 2017 Conference on Interaction Design and Children*. IDC ’17. Stanford, California, USA: Association for Computing Machinery, 2017, pp. 595–600. ISBN: 9781450349215. DOI: 10.1145/3078072.3084330.

[82] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. “Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space”. In: *International Journal of Human–Computer Interaction* 31.4 (Apr. 2015), pp. 307–335. ISSN: 1044-7318. DOI: 10.1080/10447318.2014.986642. (Visited on 08/26/2021).

[83] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. “Voice Activated Personal Assistant: Acceptability of Use in the Public Space”. en. In: *Human Interface and the Management of Information. Information and Knowledge in Applications and Services*. Ed. by Sakae Yamamoto. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 324–334. ISBN: 978-3-319-07863-2. DOI: 10.1007/978-3-319-07863-2_32.

[84] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. “Voice Activated Personal Assistant: Acceptability of Use in the Public Space”. en. In: *Human Interface and the Management of Information. Information and Knowledge in Applications and Services*. Ed. by Sakae Yamamoto. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 324–334. ISBN: 978-3-319-07863-2. DOI: 10.1007/978-3-319-07863-2_32.

[85] Christos Efthymiou and M. Halvey. “Evaluating the Social Acceptability of Voice Based Smartwatch Search”. In: *AIRS*. 2016. DOI: 10.1007/978-3-319-48051-0_20.

[86] Christos Efthymiou and Martin Halvey. “Evaluating the Social Acceptability of Voice Based Smartwatch Search”. en. In: *Information Retrieval Technology*. Ed. by Shaoping Ma et al. Vol. 9994. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016, pp. 267–278. ISBN: 978-3-319-48050-3. DOI: 10.1007/978-3-319-48051-0_20. (Visited on 08/26/2021).

[87] M. J. Fagan et al. “Development of a (silent) Speech Recognition System for Patients Following Laryngectomy”. eng. In: *Medical Engineering & Physics* 30.4 (May 2008), pp. 419–425. ISSN: 1350-4533. DOI: 10.1016/j.medengphy.2007.05.003.

[88] Wenxin Feng, Ming Chen, and Margrit Betke. “Target Reverse Crossing: A Selection Method for Camera-Based Mouse-Replacement Systems”. In: *Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA ’14. New York, NY, USA: Association for Computing Machinery, May 2014, pp. 1–4. ISBN: 978-1-4503-2746-6. DOI: 10.1145/2674396.2674443. (Visited on 09/01/2021).

[89] Xue Feng, Yaodong Zhang, and James Glass. “Speech Feature Denoising and Dereverberation Via Deep Autoencoders for Noisy Reverberant Speech Recognition”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. May 2014, pp. 1759–1763. DOI: 10.1109/ICASSP.2014.6853900.

[90] Victoria M. Florescu et al. “Silent Vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface”. In: *INTERSPEECH*. 2010.

[91] Eric Fosler-Lussier and Nelson Morgan. "Effects of Speaking Rate and Word Frequency on Pronunciations in Convertional Speech". In: *Speech Commun.* 29.2 (Nov. 1999), pp. 137–158. ISSN: 0167-6393.

[92] Xueyang Fu et al. "A Weighted Variational Model for Simultaneous Reflectance and Illumination Estimation". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2016, pp. 2782–2790. DOI: 10.1109/CVPR.2016.304.

[93] Masaaki Fukumoto. "Silentvoice: Unnoticeable Voice Input by Ingressive Speech". In: *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. UIST '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 237–246. ISBN: 978-1-4503-5948-1. DOI: 10.1145/3242587.3242603. (Visited on 09/10/2020).

[94] Yohei Fusayasu et al. "Word-Error Correction of Continuous Speech Recognition Based on Normalized Relevance Distance". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, July 2015, pp. 1257–1262. ISBN: 978-1-57735-738-4. (Visited on 09/10/2020).

[95] GazeRecorder. *GazeCloudAPI: Real-Time online Eye-Tracking API*. en-US. 2016. URL: https://gazerecorder.com/gazecloudapi/ (visited on 08/24/2021).

[96] Shabnam Ghaffarzadegan, Hynek Bořil, and John H. Hansen. "Deep Neural Network Training for Whispered Speech Recognition Using Small Databases and Generative Model Sampling". In: *International Journal of Speech Technology* 20.4 (Dec. 2017), pp. 1063–1075. ISSN: 1381-2416. DOI: 10.1007/s10772-017-9461-x. (Visited on 09/10/2020).

[97] Shabnam Ghaffarzadegan, Hynek Bořil, and John H. L. Hansen. "Generative Modeling of Pseudo-Whisper for Robust Whispered Speech Recognition". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.10 (Oct. 2016). Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 1705–1720. ISSN: 2329-9304. DOI: 10.1109/TASLP.2016.2580944.

[98] Muhammad Usman Ghani et al. "GazePointer: A real time mouse pointer control implementation based on eye gaze tracking". In: *INMIC*. INMIC. Dec. 2013, pp. 154–159. DOI: 10.1109/INMIC.2013.6731342.

[99] J. M. Gilbert et al. "Isolated Word Recognition of Silent Speech Using Magnetic Implants and Sensors". en. In: *Medical Engineering & Physics* 32.10 (Dec. 2010), pp. 1189–1197. ISSN: 1350-4533. DOI: 10.1016/j.medengphy.2010.08.011. (Visited on 09/10/2020).

[100] J.J. Godfrey, E.C. Holliman, and J. McDaniel. "Switchboard: Telephone Speech Corpus for Research and Development". In: *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. ISSN: 1520-6149. Mar. 1992, 517–520 vol.1. DOI: 10.1109/ICASSP.1992.225858.

[101] 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.

[102] Rafael C. Gonzalez and Richard E. Woods. *Digital Image Processing*. 4th. Upper Saddle River, NJ, USA: Pearson, 2018.

[103] Google. *Speech-to-Text: Automatic Speech Recognition*. en. Apr. 2017. URL: https://cloud.google.com/speech-to-text (visited on 08/27/2021).

[104] Anja S. Göritz, Kathrin Borchert, and Matthias Hirth. "Using Attention Testing to Select Crowdsourced Workers and Research Participants". en. In: *Social Science Computer Review* (June 2019). Publisher: SAGE Publications Inc, p. 0894439319848726. ISSN: 0894-4393. DOI: 10.1177/0894439319848726. (Visited on 09/16/2020).

[105] Alex Graves and Navdeep Jaitly. "Towards End-to-End Speech Recognition with Recurrent Neural Networks". In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, June 2014, pp. II–1764–II–1772. (Visited on 09/10/2020).

[106] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech Recognition with Deep Recurrent Neural Networks". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. ISSN: 2379-190X. May 2013, pp. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947.

[107] Alex Graves et al. "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks". In: *Proceedings of the 23rd international conference on Machine learning*. ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, June 25, 2006, pp. 369–376. ISBN: 978-1-59593-383-6. DOI: 10.1145/1143844.1143891. (Visited on 08/05/2020).

[108] Dorde T. Grozdic and Slobodan T. Jovicic. "Whispered Speech Recognition Using Deep Denoising Autoencoder and Inverse Filtering". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (Dec. 2017). Conference Name: IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp. 2313–2322. ISSN: 2329-9304. DOI: 10.1109/TASLP.2017.2738559.

[109] Susan G. Guion et al. "Age of Learning Effects on the Duration of Sentences Produced in a Second Language". en. In: *Applied Psycholinguistics* 21.2 (June 2000). Publisher: Cambridge University Press, pp. 205–228. ISSN: 1469-1817, 0142-7164. DOI: 10.1017/S0142716400002034. (Visited on 07/04/2021).

[110] Xiaojie Guo, Yu Li, and Haibin Ling. "Lime: Low-Light Image Enhancement Via Illumination Map Estimation". In: *IEEE Transactions on Image Processing* 26.2 (Feb. 2017). Conference Name: IEEE Transactions on Image Processing, pp. 982–993. ISSN: 1941-0042. DOI: 10.1109/TIP.2016.2639450.

[111] Awni Hannun et al. "Deep Speech: Scaling up End-to-End Speech Recognition". In: *arXiv:14 125567 [cs]* (Dec. 2014). URL: http://arxiv.org/abs/1412.5567 (visited on 09/10/2020).

[112] John Paulin Hansen et al. "Command Without a Click: Dwell Time Typing by Mouse and Gaze Selections". In: *The 10th International Conference on Human-Computer Interaction*. Ed. by M. Rauterberg. INTERACT '03. Crete, Greece: IOS, 2003, pp. 121–128.

[113] Sandra G Hart. "NASA-task Load Index (NASA-TLX); 20 Years Later". In: *Proceedings of the human factors and ergonomics society annual meeting*. Vol. 50. 9. Sage publications Sage CA: Los Angeles, CA. 2006, pp. 904–908.

[114] Sandra G. Hart and Lowell E. Staveland. "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research". en. In: *Advances in Psychology*. Vol. 52. Elsevier, 1988, pp. 139–183. ISBN: 978-0-444-70388-0. DOI: 10.1016/S0166-4115(08)62386-9. (Visited on 11/22/2020).

[115] Hideki Hashimoto et al. "Speech Recognition Interface System Suitable for Window Systems and Speech Mail Systems". US5632002A. May 1997. URL: https://patents.google.com/patent/US5632002/en (visited on 09/10/2020).

[116] Alexander G. Hauptmann and Alexander I. Rudnicky. "A Comparison of Speech and Typed Input". In: *Proceedings of the Workshop on Speech and Natural Language*. HLT '90. Hidden Valley, Pennsylvania: Association for Computational Linguistics, 1990, pp. 219–224. DOI: 10.3115/116580.116652.

[117] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. June 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[118] Yanzhang He et al. "Streaming End-to-End Speech Recognition for Mobile Devices". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. May 2019, pp. 6381–6385. DOI: 10.1109/ICASSP.2019.8682336.

[119] Kenneth Heafield et al. "Scalable Modified Kneser-Ney Language Model Estimation". In: *ACL*. 2013.

[120] Panikos Heracleous and Norihiro Hagita. "Automatic Recognition of Speech Without Any Audio Information". In: *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. May 2011, pp. 2392–2395. DOI: 10.1109/ICASSP.2011.5946965.

[121] Panikos Heracleous et al. "Unvoiced Speech Recognition Using Tissue-Conductive Acoustic Sensor". In: *EURASIP J. Adv. Signal Process.* (2007). DOI: 10.1155/2007/94068.

[122] Tatsuya Hirahara et al. "Silent-Speech Enhancement Using Body-Conducted Vocal-Tract Resonance Signals". In: *Speech Communication* 52.4 (Apr. 2010), pp. 301–313. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.12.001. (Visited on 09/10/2020).

[123] Julia Hirschberg, Diane Litman, and Marc Swerts. "Prosodic and Other Cues to Speech Recognition Failures". In: *Speech Communication* 43.1-2 (2004), pp. 155–175.

[124] Alain Horé and Djemel Ziou. "Image Quality Metrics: Psnr Vs. Ssim". In: *2010 20th International Conference on Pattern Recognition*. ISSN: 1051-4651. Aug. 2010, pp. 2366–2369. DOI: 10.1109/ICPR.2010.579.

[125] Baosheng James Hou et al. "GIMIS: Gaze Input with Motor Imagery Selection". In: *ACM Symposium on Eye Tracking Research and Applications*. ETRA '20 Adjunct. New York, NY, USA: Association for Computing Machinery, June 2020, pp. 1–10. ISBN: 978-1-4503-7135-3. DOI: 10.1145/3379157.3388932. (Visited on 09/01/2021).

[126] Davis Howes. "On the Interpretation of Word Frequency as a Variable Affecting Speed of Recognition." In: *Journal of Experimental Psychology* 48.2 (1954), p. 106.

[127] Matthew B. Hoy. “Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants”. In: *Medical Reference Services Quarterly* 37.1 (Jan. 2018). Publisher: Routledge _eprint: https://doi.org/10.1080/02763869.2018.1404391, pp. 81–88. ISSN: 0276-3869. DOI: 10.1080/02763869.2018.1404391. (Visited on 09/10/2020).

[128] Jie Hu et al. “Squeeze-and-Excitation Networks”. In: *arXiv:1709.01507 [cs]* (May 2019). arXiv: 1709.01507. URL: http://arxiv.org/abs/1709.01507 (visited on 09/10/2020).

[129] T. Hueber et al. “Eigentongue Feature Extraction for an Ultrasound-Based Silent Speech Interface”. In: *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*. Vol. 1. ISSN: 2379-190X. Apr. 2007, pp. I–1245–I–1248. DOI: 10.1109/ICASSP.2007.366140.

[130] Thomas Hueber et al. “Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips”. In: *Speech Communication* 52.4 (Apr. 2010), pp. 288–300. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.11.004. (Visited on 09/10/2020).

[131] Aulikki Hyrskykari, Howell Istance, and Stephen Vickers. “Gaze Gestures or Dwell-Based Interaction?” In: *Proceedings of the Symposium on Eye Tracking Research and Applications*. ETRA '12. New York, NY, USA: Association for Computing Machinery, Mar. 2012, pp. 229–232. ISBN: 978-1-4503-1221-9. DOI: 10.1145/2168556.2168602. (Visited on 09/01/2021).

[132] International Organization for Standardization. *ISO/TS 9241-411:2012*. en. May 2012. URL: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/41/54106.html (visited on 08/21/2021).

[133] Sergey Ioffe and Christian Szegedy. “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift”. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, July 6, 2015, pp. 448–456. (Visited on 08/05/2020).

[134] Muhammad Zahid Iqbal and Abraham Campbell. “The Emerging Need for Touchless Interaction Technologies”. en. In: *Interactions* 27.4 (July 2020), pp. 51–52. ISSN: 1072-5520, 1558-3449. DOI: 10.1145/3406100. (Visited on 08/24/2021).

[135] Howell Istance et al. “Designing Gaze Gestures for Gaming: An Investigation of Performance”. In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ETRA '10. New York, NY, USA: Association for Computing Machinery, Mar. 2010, pp. 323–330. ISBN: 978-1-60558-994-7. DOI: 10.1145/1743666.1743740. (Visited on 08/31/2021).

[136] E. Jacewicz, R. Fox, and L. Wei. “Between-speaker and within-speaker Variation in Speech Tempo of American English.” In: *The Journal of the Acoustical Society of America* 128 2 (2010), pp. 839–50.

[137] Ewa Jacewicz et al. “Articulation Rate Across Dialect, Age, and Gender”. In: *Language variation and change* 21.2 (2009), p. 233.

[138] Robert J. K. Jacob. “Eye Tracking in Advanced Interface Design”. In: *Virtual Environments and Advanced Interface Design*. Ed. by W Barfield and T. A. Furness. New York, NY, USA: University Press, 1995, pp. 258–288.

[139] Robert J. K. Jacob. “The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look at Is What You Get”. In: *ACM Transactions on Information Systems* 9.2 (Apr. 1991), pp. 152–169. ISSN: 1046-8188. DOI: 10.1145/123078.128728. URL: https://doi.org/10.1145/123078.128728 (visited on 08/25/2021).

[140] D. V. Jeevithashree, Kamalpreet Singh Saluja, and Pradipta Biswas. “A Case Study of Developing Gaze Controlled Interface for Users with Severe Speech and Motor Impairment”. en. In: *Technology and Disability* 31.1-2 (Jan. 2019). Publisher: IOS Press, pp. 63–76. ISSN: 1055-4181. DOI: 10.3233/TAD-180206. (Visited on 08/24/2021).

[141] Shuiwang Ji et al. “3D Convolutional Neural Networks for Human Action Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.1 (Jan. 2013). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 221–231. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2012.59.

[142] Jiepu Jiang, Wei Jeng, and Daqing He. “How Do Users Respond to Voice Input Errors? Lexical and Phonetic Query Reformulation in Voice Search”. In: *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR ’13. Dublin, Ireland: Association for Computing Machinery, 2013, pp. 143–152. ISBN: 9781450320344. DOI: 10.1145/2484028.2484092.

[143] D.J. Jobson, Z. Rahman, and G.A. Woodell. “A Multiscale Retinex for Bridging the Gap Between Color Images and the Human Observation of Scenes”. In: *IEEE Transactions on Image Processing* 6.7 (July 1997). Conference Name: IEEE Transactions on Image Processing, pp. 965–976. ISSN: 1941-0042. DOI: 10.1109/83.597272.

[144] Timothy R Jordan and Sharon M Thomas. “When Half a Face is as Good as a Whole: Effects of Simple Substantial Occlusion on Visual and Audiovisual Speech Perception”. In: *Attention, Perception, & Psychophysics* 73.7 (2011), p. 2270.

[145] C. Jorgensen, D.D. Lee, and S. Agabont. “Sub Auditory Speech Recognition Based on Emg Signals”. In: *Proceedings of the International Joint Conference on Neural Networks, 2003.* Vol. 4. ISSN: 1098-7576. July 2003, 3128–3133 vol.4. DOI: 10.1109/IJCNN.2003.1224072.

[146] Charles Jorgensen and Sorin Dusan. “Speech Interfaces Based Upon Surface Electromyography”. en. In: *Speech Communication*. Silent Speech Interfaces 52.4 (Apr. 2010), pp. 354–366. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.11.003. (Visited on 09/10/2020).

[147] S. Jou et al. “Towards Continuous Speech Recognition Using Surface Electromyography”. In: *INTERSPEECH*. 2006.

[148] Szu-Chen Jou and et al et. *Adaptation for Soft Whisper Recognition Using a Throat Microphone*. 2004.

[149] Yvonne Kammerer, Katharina Scheiter, and Wolfgang Beinhauer. “Looking My Way Through the Menu: The Impact of Menu Design and Multimodal Input on Gaze-Based Menu Selection”. In: *Proceedings of the 2008 symposium on Eye tracking research & applications*. ETRA ’08. New York, NY, USA: Association for Computing Machinery, Mar. 2008, pp. 213–220. ISBN: 978-1-59593-982-1. DOI: 10.1145/1344471.1344522. (Visited on 08/31/2021).

[150] Arnav Kapur, Shreyas Kapur, and Pattie Maes. “Alterego: A Personalized Wearable Silent Speech Interface”. In: *23rd International Conference on Intelligent User Interfaces*. IUI ’18. New York, NY, USA: Association for Computing Machinery, Mar. 2018, pp. 43–53. ISBN: 978-1-4503-4945-1. DOI: 10.1145/3172944.3172977. (Visited on 09/10/2020).

[151] Anuradha Kar and Peter Corcoran. “Performance Evaluation Strategies for Eye Gaze Estimation Systems with Quantitative Metrics and Visualizations”. en. In: *Sensors* 18.9 (Sept. 2018). Number: 9 Publisher: Multidisciplinary Digital Publishing Institute, p. 3151. DOI: 10.3390/s18093151. (Visited on 08/31/2021).

[152] A.E. Kaufman, A. Bandopadhay, and B.D. Shaviv. “An Eye Tracking Computer User Interface”. In: *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium*. Oct. 1993, pp. 120–121. DOI: 10.1109/VRAIS.1993.378254.

[153] Alan Kennedy et al. “Dialogue with Machines”. In: *Cognition* 30.1 (1988), pp. 37–72.

[154] Sara Kiesler, Jane Siegel, and Timothy W. McGuire. “Social Psychological Aspects of Computer-Mediated Communication”. In: *American Psychologist* 39.10 (1984). Place: US Publisher: American Psychological Association, pp. 1123–1134. ISSN: 1935-990X. DOI: 10.1037/0003-066X.39.10.1123.

[155] Carolyn Kimme, Dana Ballard, and Jack Sklansky. “Finding Circles by an Array of Accumulators”. In: *Commun. ACM* 18.2 (Feb. 1975), pp. 120–122. ISSN: 0001-0782. DOI: 10.1145/360666.360677.

[156] Naoki Kimura, Michinari Kono, and Jun Rekimoto. “SottoVoce: An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI ’19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 1–11. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300376. (Visited on 09/10/2020).

[157] Davis E. King. “Dlib-Ml: A Machine Learning Toolkit”. In: *The Journal of Machine Learning Research* 10 (Dec. 2009), pp. 1755–1758. ISSN: 1532-4435.

[158] Peter F. King. “Server Based Speech Recognition User Interface for Wireless Devices”. US6532446B1. Mar. 2003. URL: https://patents.google.com/patent/US6532446B1/en (visited on 09/10/2020).

[159] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *arXiv:1412.6980 [cs]* (Jan. 29, 2017). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980 (visited on 08/05/2020).

[160] Marion Koelle, Swamy Ananthanarayan, and Susanne Boll. “Social Acceptability in HCI: A Survey of Methods, Measures, and Design Strategies”. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI ’20. Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–19. ISBN: 9781450367080. DOI: 10.1145/3313831.3376162.

[161] Marion Koelle, Matthias Kranz, and Andreas Möller. "Don't Look at Me That Way! Understanding User Attitudes Towards Data Glasses Usage". In: *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '15. Copenhagen, Denmark: Association for Computing Machinery, 2015, pp. 362–372. ISBN: 9781450336529. DOI: 10.1145/2785830.2785842.

[162] Marion Koelle et al. "All about Acceptability? Identifying Factors for the Adoption of Data Glasses". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 295–300. ISBN: 9781450346559. DOI: 10.1145/3025453.3025749.

[163] Oscar Koller, Hermann Ney, and Richard Bowden. "Deep Learning of Mouth Shapes for Sign Language". In: *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. Dec. 2015, pp. 477–483. DOI: 10.1109/ICCVW.2015.69.

[164] Andreas Komninos, Emma Nicol, and Mark Dunlop. *Investigating Error Injection to Enhance the Effectiveness of Mobile Text Entry Studies of Error Behaviour*. 2020. arXiv: 2003.06318 [cs.HC].

[165] Kyle Krafka et al. "Eye Tracking for Everyone". en. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 2176–2184. ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.239. (Visited on 08/24/2021).

[166] Chandan Kumar et al. "Chromium Based Framework to Include Gaze Interaction in Web Browser". In: *Proceedings of the 26th International Conference on World Wide Web Companion*. WWW '17 Companion. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, Apr. 2017, pp. 219–223. ISBN: 978-1-4503-4914-7. DOI: 10.1145/3041021.3054730. (Visited on 08/24/2021).

[167] Chandan Kumar et al. "TAGSwipe: Touch Assisted Gaze Swipe for Text Entry". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, Apr. 2020, pp. 1–12. ISBN: 978-1-4503-6708-0. URL: https://doi.org/10.1145/3313831.3376317 (visited on 09/01/2021).

[168] Dounia Lahoual and Myriam Frejus. "When Users Assist the Voice Assistants: From Supervision to Failure Resolution". In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI EA '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–8. ISBN: 9781450359719. DOI: 10.1145/3290607.3299053.

[169] E. H. Land and J. McCann. "Lightness and Retinex Theory." In: *Journal of the Optical Society of America* (1971). DOI: 10.1364/JOSA.61.000001.

[170] Chulwoo Lee, Chul Lee, and Chang-Su Kim. "Contrast Enhancement Based on Layered Difference Representation of 2d Histograms". In: *IEEE Transactions on Image Processing* 22.12 (Dec. 2013). Conference Name: IEEE Transactions on Image Processing, pp. 5372–5384. ISSN: 1941-0042. DOI: 10.1109/TIP.2013.2284059.

[171] DoYoung Lee et al. "Designing Socially Acceptable Hand-to-Face Input". In: *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. UIST '18. Berlin, Germany: Association for Computing Machinery, 2018, pp. 711–723. ISBN: 9781450359481. DOI: 10.1145/3242587.3242642.

[172] Xin Lei et al. “Accurate and Compact Large Vocabulary Speech Recognition on Mobile Devices”. In: *INTERSPEECH*. 2013.

[173] Chongyi Li et al. “Lightennet: A Convolutional Neural Network for Weakly Illuminated Image Enhancement”. In: *Pattern Recognit. Lett.* (2018). DOI: 10.1016/j.patrec.2018.01.010.

[174] Ning Li et al. “Research on Human-Computer Interaction Mode of Speech Recognition Based on Environment Elements of Command and Control System”. In: *2019 5th International Conference on Big Data and Information Analytics (BigDIA)*. July 2019, pp. 170–175. DOI: 10.1109/BigDIA.2019.8802812.

[175] Yuan Liang, Koji Iwano, and Koichi Shinoda. “An Efficient Error Correction Interface for Speech Recognition on Mobile Touchscreen Devices”. In: *2014 IEEE Spoken Language Technology Workshop (SLT)*. Dec. 2014, pp. 454–459. DOI: 10.1109/SLT.2014.7078617.

[176] M. Lohse et al. ““Try Something Else!” — When Users Change Their Discursive Behavior in Human-robot Interaction”. In: *2008 IEEE International Conference on Robotics and Automation*. 2008, pp. 3481–3486. DOI: 10.1109/ROBOT.2008.4543743.

[177] Gustavo López, Luis Quesada, and Luis A. Guerrero. “Alexa Vs. Siri Vs. Cortana Vs. Google Assistant: A Comparison of Speech-Based Natural User Interfaces”. en. In: *Advances in Human Factors and Systems Interaction*. Ed. by Isabel L. Nunes. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing, 2018, pp. 241–250. ISBN: 978-3-319-60366-7. DOI: 10.1007/978-3-319-60366-7_23.

[178] X. Lu et al. “Speech Enhancement Based on Deep Denoising Autoencoder”. In: *INTERSPEECH*. 2013.

[179] Paul A Luce and David B Pisoni. “Recognizing Spoken Words: The Neighborhood Activation Model”. In: *Ear and hearing* 19.1 (1998), p. 1.

[180] Andrés Lucero and Akos Vetek. “NotifEye: using interactive glasses to deal with notifications while walking in public”. In: *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology*. ACE ’14. New York, NY, USA: Association for Computing Machinery, Nov. 11, 2014, pp. 1–10. ISBN: 978-1-4503-2945-3. DOI: 10.1145/2663806.2663824. (Visited on 09/06/2020).

[181] Ewa Luger and Abigail Sellen. “”Like Having a Really Bad PA”: The Gulf between User Expectation and Experience of Conversational Agents”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: Association for Computing Machinery, 2016, pp. 5286–5297. ISBN: 9781450333627. DOI: 10.1145/2858036.2858288.

[182] Patrick J. Lynch. “Visual Design for the User Interface, Part 1: Design Fundamentals”. In: *Journal of Biocommunication* 21 (1994), pp. 22–22. URL: http://trantor.sheridanc.on.ca/sys32a1/manual/appendix/gui1.html.

[183] Andrew L. Maas et al. “Lexicon-Free Conversational Speech Recognition with Neural Networks”. In: *HLT-NAACL*. 2015. DOI: 10.3115/v1/N15-1038.

[184] Ritch Macefield. “Usability Studies and the Hawthorne Effect”. In: *Journal of Usability Studies* 2.3 (May 2007), pp. 145–154. ISSN: 1931-3357.

[185] I. Scott MacKenzie. *Evaluating Eye Tracking Systems for Computer Input.* en. 2012. DOI: 10.4018/978-1-61350-098-9.ch015. (Visited on 08/29/2021).

[186] I. Scott MacKenzie. "Fitts' Law". en. In: *The Wiley Handbook of Human Computer Interaction.* Hoboken, NJ, USA: John Wiley & Sons, Ltd, 2018, pp. 347–370. ISBN: 978-1-118-97600-5. DOI: 10.1002/9781118976005.ch17. (Visited on 06/11/2021).

[187] 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. New York, NY, USA: Association for Computing Machinery, Apr. 2003, pp. 754–755. ISBN: 978-1-58113-637-1. DOI: 10.1145/765891.765971. (Visited on 09/10/2020).

[188] 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. New York, NY, USA: Association for Computing Machinery, Apr. 5, 2003, pp. 754–755. ISBN: 978-1-58113-637-1. DOI: 10.1145/765891.765971. (Visited on 09/06/2020).

[189] L. Maier-Hein et al. "Session Independent Non-Audible Speech Recognition Using Surface Electromyography". In: *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* Nov. 2005, pp. 331–336. DOI: 10.1109/ASRU.2005.1566521.

[190] 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.* New York, NY, USA: Association for Computing Machinery, Apr. 2009, pp. 357–360. ISBN: 978-1-60558-246-7. URL: https://doi.org/10.1145/1518701.1518758 (visited on 09/01/2021).

[191] Fernando Martinez et al. "Characteristics of Slow, Average and Fast speech and Their Effects in Large Vocabulary Continuous Speech Recognition". In: *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997).* 1997, pp. 469–472.

[192] Sarah C Mason. "Is There a Correlation Between Oral Reading Rate and Social Conversational Speaking Rate?" In: (2019).

[193] Julio C. Mateo, Javier San Agustin, and John Paulin Hansen. "Gaze Beats Mouse: Hands-Free Selection by Combining Gaze and EMG". In: *CHI '08 Extended Abstracts on Human Factors in Computing Systems.* New York, NY, USA: Association for Computing Machinery, Apr. 2008, pp. 3039–3044. ISBN: 978-1-60558-012-8. URL: https://doi.org/10.1145/1358628.1358804 (visited on 09/01/2021).

[194] Ian McGraw et al. "Personalized Speech Recognition on Mobile Devices". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* ISSN: 2379-190X. Mar. 2016, pp. 5955–5959. DOI: 10.1109/ICASSP.2016.7472820.

[195] I. Mcloughlin, J. Li, and Yan Song. "Reconstruction of Continuous Voiced Speech from Whispers". In: *INTERSPEECH.* 2013.

[196] *Menu Anatomy - Menus - macOS - Human Interface Guidelines - Apple Developer.* 2022. URL: https://developer.apple.com/design/human-interface-guidelines/macos/menus/menu-anatomy (visited on 05/24/2022).

[197] George A. Miller. “The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information”. In: *Psychological Review* 63.2 (1956). Place: US Publisher: American Psychological Association, pp. 81–97. ISSN: 1939-1471. DOI: 10.1037/h0043158.

[198] Katsumi Minakata et al. “Pointing by Gaze, Head, and Foot in a Head-Mounted Display”. In: *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*. ETRA ’19. New York, NY, USA: Association for Computing Machinery, June 2019, pp. 1–9. ISBN: 978-1-4503-6709-7. DOI: 10.1145/3317956.3318150. (Visited on 09/01/2021).

[199] Darius Miniotas, Oleg Špakov, and I. Scott MacKenzie. “Eye Gaze Interaction with Expanding Targets”. In: *CHI ’04 Extended Abstracts on Human Factors in Computing Systems*. CHI EA ’04. New York, NY, USA: Association for Computing Machinery, Apr. 2004, pp. 1255–1258. ISBN: 978-1-58113-703-3. DOI: 10.1145/985921.986037. (Visited on 08/29/2021).

[200] Darius Miniotas et al. “Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets”. In: *Proceedings of the 2006 symposium on Eye tracking research & applications*. ETRA ’06. New York, NY, USA: Association for Computing Machinery, Mar. 2006, pp. 67–72. ISBN: 978-1-59593-305-8. DOI: 10.1145/1117309.1117345. (Visited on 09/01/2021).

[201] Nikki Mirghafori, Eric Foster, and Nelson Morgan. “Fast speakers in large vocabulary continuous speech recognition: analysis & antidotes”. In: *Fourth European Conference on Speech Communication and Technology*. 1995.

[202] Emilie Møllenbach et al. “Single Gaze Gestures”. In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ETRA ’10. New York, NY, USA: Association for Computing Machinery, Mar. 2010, pp. 177–180. ISBN: 978-1-60558-994-7. DOI: 10.1145/1743666.1743710. (Visited on 09/01/2021).

[203] Calkin S. Montero et al. “Would you do that? understanding social acceptance of gestural interfaces”. In: *Proceedings of the 12th international conference on Human computer interaction with mobile devices and services*. MobileHCI ’10. New York, NY, USA: Association for Computing Machinery, Sept. 7, 2010, pp. 275–278. ISBN: 978-1-60558-835-3. DOI: 10.1145/1851600.1851647. (Visited on 09/06/2020).

[204] Tuuli Morrill, Melissa Baese-Berk, and Ann Bradlow. “Speaking rate consistency and variability in spontaneous speech by native and non-native speakers of English”. In: *Proceedings of the International Conference on Speech Prosody*. Vol. 2016. 2016, pp. 1119–1123.

[205] Martez E. Mott et al. “Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, May 2017, pp. 2558–2570. ISBN: 978-1-4503-4655-9. URL: https://doi.org/10.1145/3025453.3025517 (visited on 08/25/2021).

[206] Atsuo Murata and Waldemar Karwowski. “Automatic Lock of Cursor Movement: Implications for an Efficient Eye-Gaze Input Method for Drag and Menu Selection”. In: *IEEE Transactions on Human-Machine Systems* 49.3 (June 2019). Conference Name: IEEE Transactions on Human-Machine Systems, pp. 259–267. ISSN: 2168-2305. DOI: 10.1109/THMS.2018.2884737.

[207] Chelsea Myers et al. "Patterns for How Users Overcome Obstacles in Voice User Interfaces". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–7. ISBN: 9781450356206. DOI: 10.1145/3173574.3173580.

[208] Y. Nakajima et al. "Non-Audible Murmur Recognition Input Interface Using Stethoscopic Microphone Attached to the Skin". In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*. Vol. 5. ISSN: 1520-6149. Apr. 2003, pp. V–708. DOI: 10.1109/ICASSP.2003.1200069.

[209] Y. Nakajima et al. "Non-Audible Murmur Recognition Input Interface Using Stethoscopic Microphone Attached to the Skin". In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)*. Vol. 5. ISSN: 1520-6149. Apr. 2003, pp. V–708. DOI: 10.1109/ICASSP.2003.1200069.

[210] NASA. *NASA Task Load Index (TLX) V 1.0: Paper and Pencil Package*. Tech. rep. Moffett Field, CA, USA: Human Performance Research Group, NASA Ames Research Center, 1986.

[211] Chalapathy Neti et al. "Audio-Visual Speech Recognition". en. In: (2000), p. 86.

[212] Meghan Neumer. "The Relationship Between Natural Speech Rate and Oral Reading Fluency Rate and Reading Comprehension Among Third Grade Students". In: (2013).

[213] L.C. Ng et al. "Denoising of Human Speech Using Combined Acoustic and Em Sensor Signal Processing". In: *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. Vol. 1. ISSN: 1520-6149. June 2000, 229–232 vol.1. DOI: 10.1109/ICASSP.2000.861925.

[214] Erik Lloyd Nilsen. "Perceptual-Motor Control in Human-Computer Interaction". English. Ph.D. Ann Arbor, MI, USA: University of Michigan, 1991. URL: https://www.proquest.com/docview/303945464/abstract/683DEEF3C2344476PQ/1 (visited on 09/01/2021).

[215] K. Noda et al. "Lipreading Using Convolutional Neural Network". In: *INTERSPEECH*. 2014.

[216] Ian Oakley et al. "Putting the Feel in 'Look and Feel'". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. CHI '00. New York, NY, USA: Association for Computing Machinery, Apr. 2000, pp. 415–422. ISBN: 978-1-58113-216-8. DOI: 10.1145/332040.332467. URL: https://doi.org/10.1145/332040.332467 (visited on 09/02/2021).

[217] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. "Linguistic Adaptations During Spoken and Multimodal Error Resolution". In: *Language and speech* 41.3-4 (1998), pp. 419–442.

[218] Vassil Panayotov et al. "Librispeech: An Asr Corpus Based on Public Domain Audio Books". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. Apr. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[219] Laxmi Pandey and Ahmed Sabbir Arif. "Silent Speech and Emotion Recognition from Vocal Tract Shape Dynamics in Real-Time MRI". In: *ACM SIGGRAPH 2021 Posters*. SIGGRAPH '21. Virtual Event, USA: Association for Computing Machinery, 2021. ISBN: 9781450383714. DOI: 10.1145/3450618.3469176.

[220] Alexandra Papoutsaki et al. "WebGazer: Scalable Webcam Eye Tracking Using User Interactions". en. In: (), p. 7.

[221] Mohsen Parisay, Charalambos Poullis, and Marta Kersten. "EyeTAP: A Novel Technique using Voice Inputs to Address the Midas Touch Problem for Gaze-based Interactions". In: *International Journal of Human-Computer Studies* 154 (Oct. 2021). arXiv: 2002.08455, p. 102676. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2021.102676. (Visited on 09/01/2021).

[222] Seongjin Park and John Culnan. "A Comparison Between Native and Non-native Speech for Automatic Speech Recognition". In: *The Journal of the Acoustical Society of America* 145.3 (2019), pp. 1827–1827.

[223] Sanjay A. Patil and John H. L. Hansen. "The Physiological Microphone (pmic): A Competitive Alternative for Speaker Assessment in Stress Detection and Speaker Verification". In: *Speech Communication* 52.4 (Apr. 2010), pp. 327–340. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.11.006. (Visited on 09/10/2020).

[224] Douglas B. Paul and Janet M. Baker. "The Design for the Wall Street Journal-Based Csr Corpus". In: *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*. 1992. URL: https://www.aclweb.org/anthology/H92-1073 (visited on 09/13/2020).

[225] Hannah R.M. Pelikan and Mathias Broth. "Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2016, pp. 4921–4932. ISBN: 9781450333627. URL: https://doi.org/10.1145/2858036.2858478.

[226] T. Pellegrini and I. Trancoso. "Improving Asr Error Detection with Non-Decoder Based Features". In: *INTERSPEECH*. 2010.

[227] Thomas Pellegrini and Isabel Trancoso. "Error Detection in Broadcast News Asr Using Markov Chains". In: *Proceedings of the 4th conference on Human language technology: challenges for computer science and linguistics*. LTC'09. Berlin, Heidelberg: Springer-Verlag, Nov. 2009, pp. 59–69. ISBN: 978-3-642-20094-6. (Visited on 09/10/2020).

[228] Dr Marta Perez Garcia, Sarita Saffon Lopez, and Hector Donis. "Everybody is Talking About Virtual Assistants, But How are People Really Using Them?" In: *Proceedings of the 32nd International BCS Human Computer Interaction Conference 32*. 2018, pp. 1–5.

[229] Stavros Petridis and Maja Pantic. "Deep Complementary Bottleneck Features for Visual Speech Recognition". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. Mar. 2016, pp. 2304–2308. DOI: 10.1109/ICASSP.2016.7472088.

[230] J.W. Picone. "Signal Modeling Techniques in Speech Recognition". In: *Proceedings of the IEEE* 81.9 (Sept. 1993). Conference Name: Proceedings of the IEEE, pp. 1215–1247. ISSN: 1558-2256. DOI: 10.1109/5.237532.

[231] Anne Porbadnigk et al. "EEG-Based Speech Recognition - Impact of Temporal Effects". In: *BIOSIGNALS*. 2009. DOI: 10.5220/0001554303760381.

[232] G. Potamianos et al. “Recent Advances in the Automatic Recognition of Audiovisual Speech”. In: *Proceedings of the IEEE* 91.9 (Sept. 2003). Conference Name: Proceedings of the IEEE, pp. 1306–1326. ISSN: 1558-2256. DOI: 10.1109/JPROC.2003.817150.

[233] Matti Pouke et al. “Gaze Tracking and Non-Touch Gesture Based Interaction Method for Mobile 3d Virtual Spaces”. In: *Proceedings of the 24th Australian Computer-Human Interaction Conference*. OzCHI ’12. New York, NY, USA: Association for Computing Machinery, Nov. 2012, pp. 505–512. ISBN: 978-1-4503-1438-1. DOI: 10.1145/2414536.2414614. (Visited on 09/01/2021).

[234] Daniel Povey et al. *The Kaldi Speech Recognition Toolkit*. en. Conference Name: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding Number: CONF Publisher: IEEE Signal Processing Society. 2011. URL: https://infoscience.epfl.ch/record/192584 (visited on 09/07/2020).

[235] S. Prabhakar, S. Pankanti, and A.K. Jain. “Biometric Recognition: Security and Privacy Concerns”. In: *IEEE Security Privacy* 1.2 (Mar. 2003). Conference Name: IEEE Security Privacy, pp. 33–42. ISSN: 1558-4046. DOI: 10.1109/MSECP.2003.1193209.

[236] Halley Profita et al. “The AT Effect: How Disability Affects the Perceived Social Acceptability of Head-Mounted Display Use”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. CHI ’16. San Jose, California, USA: Association for Computing Machinery, 2016, pp. 4884–4895. ISBN: 9781450333627. DOI: 10.1145/2858036.2858130.

[237] Halley P. Profita. “Designing wearable computing technology for acceptability and accessibility”. In: *ACM SIGACCESS Accessibility and Computing* 114 (Mar. 16, 2016), pp. 44–48. ISSN: 1558-2337. DOI: 10.1145/2904092.2904101. (Visited on 09/06/2020).

[238] T.F. Quatieri et al. “Exploiting Nonacoustic Sensors for Speech Encoding”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 14.2 (Mar. 2006). Conference Name: IEEE Transactions on Audio, Speech, and Language Processing, pp. 533–544. ISSN: 1558-7924. DOI: 10.1109/TSA.2005.855838.

[239] Vijay Rajanna and Tracy Hammond. “A Gaze Gesture-Based Paradigm for Situational Impairments, Accessibility, and Rich Interactions”. In: *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*. ETRA ’18. New York, NY, USA: Association for Computing Machinery, June 2018, pp. 1–3. ISBN: 978-1-4503-5706-7. DOI: 10.1145/3204493.3208344. (Visited on 08/24/2021).

[240] K Sreenivasa Rao and Shashidhar G Koolagudi. “Robust Emotion Recognition using Speaking Rate Features”. In: *Robust Emotion Recognition using Spectral and Prosodic Features*. Springer, 2013, pp. 85–94.

[241] Stuart Reeves et al. “Designing the Spectator Experience”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’05. Portland, Oregon, USA: Association for Computing Machinery, 2005, pp. 741–750. ISBN: 1581139985. DOI: 10.1145/1054972.1055074.

[242] Julie Rico and Stephen Brewster. “Gestures all around us: user differences in social acceptability perceptions of gesture based interfaces”. In: *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI ’09. New York, NY, USA: Association for Computing Machinery, Sept. 15, 2009, pp. 1–2. ISBN: 978-1-60558-281-8. DOI: 10.1145/1613858.1613936. (Visited on 09/06/2020).

[243] Julie Rico and Stephen Brewster. “Usable Gestures for Mobile Interfaces: Evaluating Social Acceptability”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’10. Atlanta, Georgia, USA: Association for Computing Machinery, 2010, pp. 887–896. ISBN: 9781605589299. DOI: 10.1145/1753326.1753458.

[244] Michael Riley, Cyril Allauzen, and Martin Jansche. “Openfst: An Open-Source, Weighted Finite-State Transducer Library and Its Applications to Speech and Language”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT) 2009 conference, Tutorials*. 2009. URL: http://aclweb.org/anthology/N09-4005 (visited on 09/10/2020).

[245] Judy Robertson and Maurits Kaptein. *Modern Statistical Methods for HCI*. Springer.

[246] Sami Ronkainen et al. “Tap Input as an Embedded Interaction Method for Mobile Devices”. In: *Proceedings of the 1st International Conference on Tangible and Embedded Interaction*. TEI ’07. Baton Rouge, Louisiana: Association for Computing Machinery, 2007, pp. 263–270. ISBN: 9781595936196. DOI: 10.1145/1226969.1227023.

[247] David Rozado, Jeremy Hales, and Diako Mardanbegi. “Interacting with Objects in the Environment by Gaze and Hand Gestures”. en. In: *Proceedings of the 3rd International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*. New York, NY, USA: ECEM, 2013, pp. 1–9.

[248] Sherry Ruan et al. “Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.4 (Jan. 2018), 159:1–159:23. DOI: 10.1145/3161187. (Visited on 09/09/2020).

[249] A. Rubin et al. “Laryngeal Hyperfunction During Whispering: Reality or Myth?” In: *Journal of voice : official journal of the Voice Foundation* (2006). DOI: 10.1016/J.JVOICE.2004.10.007.

[250] Marit Ruitenberg et al. “Post-error Slowing in Sequential Action: An Aging Study”. In: *Frontiers in psychology* 5 (2014), p. 119.

[251] Christos Sagonas et al. “300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge”. In: *2013 IEEE International Conference on Computer Vision Workshops*. Dec. 2013, pp. 397–403. DOI: 10.1109/ICCVW.2013.59.

[252] Arup Sarma and David D. Palmer. “Context-Based Speech Recognition Error Detection and Correction”. In: *Proceedings of HLT-NAACL 2004: Short Papers*. HLT-NAACL-Short ’04. USA: Association for Computational Linguistics, May 2004, pp. 85–88. ISBN: 978-1-932432-24-4. (Visited on 09/10/2020).

[253] Zhanna Sarsenbayeva, Vassilis Kostakos, and Jorge Goncalves. “Situationally-Induced Impairments and Disabilities Research”. In: *arXiv:1904.06128 [cs]* (Apr. 2019). URL: http://arxiv.org/abs/1904.06128 (visited on 08/24/2021).

[254] Tanja Schultz and Michael Wand. “Modeling Coarticulation in Emg-Based Continuous Speech Recognition”. In: *Speech Communication* 52.4 (Apr. 2010), pp. 341–353. ISSN: 0167-6393. DOI: 10.1016/j.specom.2009.12.002. (Visited on 09/10/2020).

[255] Mike Schuster. “Speech Recognition for Mobile Devices at Google”. en. In: *PRICAI 2010: Trends in Artificial Intelligence*. Ed. by Byoung-Tak Zhang and Mehmet A. Orgun. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2010, pp. 8–10. ISBN: 978-3-642-15246-7. DOI: 10.1007/978-3-642-15246-7_3.

[256] Kilian Semmelmann and Sarah Weigelt. “Online Webcam-Based Eye Tracking in Cognitive Science: A First Look”. en. In: *Behavior Research Methods* 50.2 (Apr. 2018), pp. 451–465. ISSN: 1554-3528. DOI: 10.3758/s13428-017-0913-7. (Visited on 08/25/2021).

[257] Korok Sengupta et al. “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. New York, NY, USA: Association for Computing Machinery, June 2018, pp. 1–3. ISBN: 978-1-4503-5706-7. DOI: 10.1145/3204493.3208338. (Visited on 09/01/2021).

[258] Marcos Serrano, Barrett M. Ens, and Pourang P. Irani. “Exploring the Use of Hand-to-Face Input for Interacting with Head-Worn Displays”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI ’14. Toronto, Ontario, Canada: Association for Computing Machinery, 2014, pp. 3181–3190. ISBN: 9781450324731. DOI: 10.1145/2556288.2556984.

[259] A.R. Setlur, R.A. Sukkar, and J. Jacob. “Correcting Recognition Errors Via Discriminative Utterance Verification”. In: *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP ’96*. Vol. 2. Oct. 1996, 602–605 vol.2. DOI: 10.1109/ICSLP.1996.607433.

[260] R. V. Shannon et al. “Speech Recognition with Primarily Temporal Cues”. In: *Science* (1995). DOI: 10.1126/science.270.5234.303.

[261] T. Shinozaki and S. Furui. “Error Analysis Using Decision Trees in Spontaneous Presentation Speech Recognition”. In: *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU ’01*. 2001, pp. 198–201. DOI: 10.1109/ASRU.2001.1034621.

[262] 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. New York, NY, USA: Association for Computing Machinery, Apr. 2000, pp. 281–288. ISBN: 978-1-58113-216-8. DOI: 10.1145/332040.332445. (Visited on 08/24/2021).

[263] Ludwig Sidenmark and Hans Gellersen. “Eye&Head: Synergetic Eye and Head Movement for Gaze Pointing and Selection”. In: *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. UIST ’19. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 1161–1174. ISBN: 978-1-4503-6816-2. DOI: 10.1145/3332165.3347921. (Visited on 09/01/2021).

[264] Ludwig Sidenmark et al. “BimodalGaze: Seamlessly Refined Pointing with Gaze and Filtered Gestural Head Movement”. In: *ACM Symposium on Eye Tracking Research and Applications*. ETRA ’20 Full Papers. New York, NY, USA: Association for Computing Machinery, June 2020, pp. 1–9. ISBN: 978-1-4503-7133-9. DOI: 10.1145/3379155.3391312. (Visited on 09/01/2021).

[265] M.A. Siegler and R.M. Stern. “On The Effects of Speech Rate in Large Vocabulary Speech Recognition Systems”. In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1. 1995, 612–615 vol.1. DOI: 10.1109/ICASSP.1995.479672.

[266] Matthew A Siegler and Richard M Stern. “On the Effects of Speech Rate in Large Vocabulary Speech Recognition Systems”. In: *1995 international conference on acoustics, speech, and signal processing*. Vol. 1. IEEE. 1995, pp. 612–615.

[267] Henrik Skovsgaard et al. “Small-Target Selection with Gaze Alone”. In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ETRA ’10. New York, NY, USA: Association for Computing Machinery, Mar. 2010, pp. 145–148. ISBN: 978-1-60558-994-7. DOI: 10.1145/1743666.1743702. (Visited on 08/30/2021).

[268] Malcolm Slaney et al. “Gaze-Enhanced Speech Recognition”. In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. May 2014, pp. 3236–3240. DOI: 10.1109/ICASSP.2014.6854198.

[269] R. William Soukoreff and I. Scott MacKenzie. “Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts’ Law Research in HCI”. en. In: *International Journal of Human-Computer Studies*. Fitts’ law 50 years later: applications and contributions from human-computer interaction 61.6 (Dec. 2004), pp. 751–789. ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2004.09.001. (Visited on 08/28/2021).

[270] 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. New York, NY, USA: Association for Computing Machinery, Oct. 2005, pp. 124–128. ISBN: 978-1-59593-028-6. DOI: 10.1145/1088463.1088486. (Visited on 08/30/2021).

[271] 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. New York, NY, USA: Association for Computing Machinery, Oct. 2004, pp. 203–206. ISBN: 978-1-58113-857-3. DOI: 10.1145/1028014.1028045. (Visited on 09/01/2021).

[272] Brent Spehar, Stacey Goebel, and Nancy Tye-Murray. “Effects of Context Type on Lipreading and Listening Performance and Implications for Sentence Processing”. In: *Journal of speech, language, and hearing research* 58.3 (2015), pp. 1093–1102.

[273] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *The Journal of Machine Learning Research* 15.1 (Jan. 1, 2014), pp. 1929–1958. ISSN: 1532-4435.

[274] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *The Journal of Machine Learning Research* 15.1 (Jan. 2014), pp. 1929–1958. ISSN: 1532-4435.

[275] Themos Stafylakis and Georgios Tzimiropoulos. “Combining Residual Networks with Lstms for Lipreading”. In: *INTERSPEECH* (2017). DOI: 10.21437/INTERSPEECH.2017-85.

[276] Ke Sun et al. “Lip-Interact: Improving Mobile Device Interaction with Silent Speech Commands”. In: *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology*. UIST ’18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 581–593. ISBN: 978-1-4503-5948-1. DOI: 10.1145/3242587.3242599. (Visited on 09/10/2020).

[277] P. Suppes, B. Han, and Z. Lu. “Brain-Wave Recognition of Sentences.” In: *Proceedings of the National Academy of Sciences of the United States of America* (1998). DOI: 10.1073/pnas.95.26.15861.

[278] P. Suppes, Z. Lu, and B. Han. “Brain Wave Recognition of Words.” In: *Proceedings of the National Academy of Sciences of the United States of America* (1997). DOI: 10.1073/pnas.94.26.14965.

[279] Veikko Surakka, Marko Illi, and Poika Isokoski. “Gazing and Frowning as a New Human–Computer Interaction Technique”. In: *ACM Transactions on Applied Perception* 1.1 (July 2004), pp. 40–56. ISSN: 1544-3558. DOI: 10.1145/1008722.1008726. (Visited on 09/01/2021).

[280] Satoshi Tamura et al. “Audio-Visual Speech Recognition Using Deep Bottleneck Features and High-Performance Lipreading”. In: *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. Dec. 2015, pp. 575–582. DOI: 10.1109/APSIPA.2015.7415335.

[281] Ingo R. Titze et al. “Comparison Between Electroglottography and Electromagnetic Glottography”. In: *The Journal of the Acoustical Society of America* 107.1 (Dec. 1999). Publisher: Acoustical Society of America, pp. 581–588. ISSN: 0001-4966. DOI: 10.1121/1.428324. (Visited on 09/10/2020).

[282] Ying-Chao Tung et al. “User-Defined Game Input for Smart Glasses in Public Space”. 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, pp. 3327–3336. ISBN: 9781450331456. DOI: 10.1145/2702123.2702214.

[283] Naoya Ukai et al. “Gif-Lr: GA-Based Informative Feature for Lipreading”. In: *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. Dec. 2012, pp. 1–4.

[284] Mario H. Urbina, Maike Lorenz, and Anke Huckauf. “Pies with EYEs: The Limits of Hierarchical Pie Menus in Gaze Control”. In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. ETRA ’10. New York, NY, USA: Association for Computing Machinery, Mar. 2010, pp. 93–96. ISBN: 978-1-60558-994-7. DOI: 10.1145/1743666.1743689. (Visited on 08/31/2021).

[285] SIMPLY USER. *The Comparison of Accuracy and Precision of Eye Tracking: GazeFlow vs. SMI RED 250*. Tech. rep. Kraków, Poland: SIMPLY USER, User Experience Lab, Aug. 2013, p. 29. URL: https://gazerecorder.com/webcam-eye-tracking-accuracy.

[286] Ashish Vaswani et al. “Attention Is All You Need”. In: *arXiv:1706.03762 [cs]* (Dec. 2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762 (visited on 09/10/2020).

[287] Roel Vertegaal. “A Fitts’ Law Comparison of Eye Tracking and Manual Input in the Selection of Visual Targets”. In: *Proceedings of the 10th international conference on Multimodal interfaces*. ICMI ’08. New York, NY, USA: Association for Computing Machinery, Oct. 2008, pp. 241–248. ISBN: 978-1-60558-198-9. DOI: 10.1145/1452392.1452443. (Visited on 08/25/2021).

[288] Pascal Vincent et al. “Extracting and Composing Robust Features with Denoising Autoencoders”. In: *Proceedings of the 25th international conference on Machine learning*. ICML ’08. New York, NY, USA: Association for Computing Machinery, July 2008, pp. 1096–1103. ISBN: 978-1-60558-205-4. DOI: 10.1145/1390156.1390294. (Visited on 09/10/2020).

[289] P. Viola and M. Jones. “Rapid object detection using a boosted cascade of simple features”. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Vol. 1. ISSN: 1063-6919. Dec. 2001, pp. I–I. DOI: 10.1109/CVPR.2001.990517.

[290] Michael Wand, Jan Koutník, and Jürgen Schmidhuber. “Lipreading with Long Short-Term Memory”. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. Mar. 2016, pp. 6115–6119. DOI: 10.1109/ICASSP.2016.7472852.

[291] Michael Wand and Tanja Schultz. “Session-Independent Emg-Based Speech Recognition”. In: *BIOSIGNALS*. 2011. DOI: 10.5220/0003169702950300.

[292] Lijun Wang et al. “Slowing After Observed Error Transfers Across Tasks”. In: *PloS one* 11.3 (2016), e0149836.

[293] Shuhang Wang et al. “Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images”. In: *IEEE Transactions on Image Processing* 22.9 (Sept. 2013). Conference Name: IEEE Transactions on Image Processing, pp. 3538–3548. ISSN: 1941-0042. DOI: 10.1109/TIP.2013.2261309.

[294] Wenjing Wang et al. “Gladnet: Low-Light Enhancement Network with Global Awareness”. In: *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*. May 2018, pp. 751–755. DOI: 10.1109/FG.2018.00118.

[295] William Wang. *Integrating GazeCloudAPI, a High Accuracy Webcam Based Eye-Tracking Solution, into Your Own Web-App*. en. Nov. 2020. URL: https://medium.com/williamwang/integrating-gazecloudapi-a-high-accuracy-webcam-based-eye-tracking-solution-into-your-own-web-app-2d8513bb9865 (visited on 08/24/2021).

[296] Zhirong Wang, T. Schultz, and A. Waibel. “Comparison of Acoustic Model Adaptation Techniques on Non-native Speech”. In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03)*. Vol. 1. 2003, pp. I–I. DOI: 10.1109/ICASSP.2003.1198837.

[297] Dean Weber. “Interactive User Interface Using Speech Recognition and Natural Language Processing”. en. Apr. 2001. URL: https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2001026093 (visited on 09/10/2020).

[298] Dean Weber. "Object Interactive User Interface Using Speech Recognition and Natural Language Processing". US6434524B1. Aug. 2002. URL: https://patents.google.com/patent/US6434524B1/en (visited on 09/10/2020).

[299] Chen Wei et al. "Deep Retinex Decomposition for Low-Light Enhancement". In: *arXiv:1808.04560 [cs]* (Aug. 2018). URL: http://arxiv.org/abs/1808.04560 (visited on 09/10/2020).

[300] Jacob O. Wobbrock. "Situationally Aware Mobile Devices for Overcoming Situational Impairments". In: *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems*. EICS '19. New York, NY, USA: Association for Computing Machinery, June 2019, pp. 1–18. ISBN: 978-1-4503-6745-5. DOI: 10.1145/3319499.3330292. (Visited on 08/24/2021).

[301] Kai Xu et al. "LCANet: End-to-end Lipreading with Cascaded Attention-CTC". In: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE. 2018, pp. 548–555.

[302] Pingmei Xu et al. "TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking". In: *arXiv:1504.06755 [cs]* (May 2015). arXiv: 1504.06755. URL: http://arxiv.org/abs/1504.06755 (visited on 08/25/2021).

[303] Jiahong Yuan, Mark Liberman, and Christopher Cieri. "Towards an Integrated Understanding of Speaking Rate in Conversation". In: *Ninth International Conference on Spoken Language Processing*. 2006.

[304] D. Zaykovskiy. "Survey of the Speech Recognition Techniques for Mobile Devices". In: 2006.

[305] Xiangyu Zeng, Shi Yin, and Dong Wang. "Learning Speech Rate in Speech Recognition". In: *arXiv preprint arXiv:1506.00799* (2015).

[306] Xuebai Zhang et al. "Eye Tracking Based Control System for Natural Human-Computer Interaction". en. In: *Computational Intelligence and Neuroscience* 2017 (Dec. 2017). Publisher: Hindawi, e5739301. ISSN: 1687-5265. DOI: 10.1155/2017/5739301. (Visited on 08/24/2021).

[307] You Zhang et al. "Automobile Speech-Recognition Interface". en. US7826945B2. Nov. 2010. URL: https://patents.google.com/patent/US7826945B2/en (visited on 09/10/2020).

[308] Zhilu Zhang and Mert R. Sabuncu. "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels". In: *arXiv:1805.07836 [cs, stat]* (Nov. 2018). arXiv: 1805.07836. URL: http://arxiv.org/abs/1805.07836 (visited on 09/10/2020).

[309] Hang Zhao et al. "Loss Functions for Neural Networks for Image Processing". In: *arXiv:1511.08861 [cs]* (Apr. 2018). URL: http://arxiv.org/abs/1511.08861 (visited on 09/10/2020).

[310] Yu Zhong et al. "Justspeak: Enabling Universal Voice Control on Android". In: *W4A 2014*. 2014. URL: http://dl.acm.org/citation.cfm?id=2596720 (visited on 09/10/2020).

[311] Lina Zhou et al. "Data Mining for Detecting Errors in Dictation Speech Recognition". In: *IEEE Transactions on Speech and Audio Processing* 13.5 (Sept. 2005). Conference Name: IEEE Transactions on Speech and Audio Processing, pp. 681–688. ISSN: 1558-2353. DOI: 10.1109/TSA.2005.851874.

[312] Ziheng Zhou et al. “A Review of Recent Advances in Visual Speech Decoding”. en. In: *Image and Vision Computing* 32.9 (Sept. 2014), pp. 590–605. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2014.06.004. (Visited on 09/10/2020).

Appendix A

Test Dataset: Pretrained Model

In this appendix, we provide details about the selected phrases for seen and unseen data. For seen data, we randomly selected 30 phrases from each pretrained models' training dataset. For unseen data, we randomly selected 30 phrases from MacKenzie and Soukoreff [187] dataset, which is common for all models.

A.1 LipNet: Seen Data (Grid Data [66])

1. bin blue at c one again
2. set blue in f four soon
3. set blue with l eight now
4. bin green by a four soon
5. place blue by t nine soon
6. place red with a four please
7. place green by p five again
8. lay red with k seven soon
9. set blue in g four now
10. set red with m zero soon
11. bin white at q six soon
12. place blue at n six now
13. bin white with f seven soon
14. place blue at m seven now
15. bin red by j five please
16. bin white at a one please
17. set red by b one now
18. place blue at i one soon
19. place blue in n two please
20. lay red by d seven please
21. bin white by z three now

22. place white with e three again
23. bin red in j four now
24. set blue in e four again
25. lay green at p four again
26. bin red with z eight please
27. place red with n two please
28. lay blue with b two please
29. set green with v eight now
30. bin white at j nine now

A.2 LipType: Seen Data (Grid Data [66])

1. set green at f four soon
2. bin green by h zero please
3. bin white at f zero again
4. place blue with j five again
5. place white in g two please
6. bin white by d eight again
7. bin blue in q seven please
8. lay red with f zero again
9. place white at p eight now
10. lay red with k three now
11. lay red in j one soon
12. lay white at j nine soon
13. lay red at v eight again
14. place green in u zero now
15. lay red with c eight again
16. place green at u two now
17. place white by v four now
18. bin red in x one now
19. bin green at e zero again
20. lay white by p six please
21. bin red with x nine again
22. place red at c three now
23. set green at o seven please
24. bin red at s eight again
25. place red in s three please
26. bin green by n four again
27. place green by y two please
28. place green by k one please
29. lay blue at c one please

30. place red by n one please

A.3 Transformer: Seen Data (LRS Data [5])

1. the whole gardens are extraordinary and
2. like hundreds of thousands of people do every year
3. but now there is more protection
4. we have a lot less atmosphere above us
5. and a couple of weeks ago
6. enjoy the summer
7. are they relatives of yours
8. no longer dependent on the sun
9. but the waldorf astoria
10. they would be able to go back
11. not just a hotel
12. not just in this town
13. every september this place would be transformed into what
14. now they are gathering
15. so from his vantage point
16. there is no air so there is no sound
17. with one of the rooms upstairs
18. maybe more of steel and iron
19. we have run out of time
20. before we all get too excited about that prospect
21. it can be quite expensive
22. in the form of a dessert plate
23. on the face of it
24. it could be your passport to a small fortune
25. some issues with potential damp
26. a great place for him to be
27. we have to pay for that
28. so rather than just relying on this information
29. all of the brain is combining all the different senses
30. he ordered them back inside

[noitemsep]

A.4 DeepSpeech: Seen Data (Fisher English-Conversational [62])

1. can you hear me okay by the way
2. oh good as long as you can hear me
3. yeah i can hear you
4. yeah that would be interesting
5. like ten minutes with a head set on i might as well exercise
6. yeah thats great
7. listening to the music anyway so um
8. i actually think its actually going out
9. fifth wheel dating show
10. i also watch that show the fifth wheel third and fourth wheel
11. and i have seen i remember when survivor first started
12. i saw that like a couple things
13. cause my roommate where watching it
14. yeah my roommates are you in college too
15. i am in graduate school
16. oh yeah okay i just graduated from um
17. first time graduate last year
18. and how about what school are you in
19. that was great performance tonight
20. it would be it would be cool to be on it
21. thats very cool
22. popular everyone talks about it
23. somebody from my high school one something too
24. he won he was like on that
25. greatest bachelor show
26. it was before these millionaire the millionaire guy ones
27. it was like a pageant for men
28. i didnt see it but i think i know what you were talking about
29. yeah he was in my old high school
30. going to rat on the other one

A.5 Kaldi: Seen Data (LIBRISPEECH Audiobooks [218])

1. he was in a mood for music was he not
2. give not so earnest a mind to these mummeries child
3. a golden fortune and a happy life

4. he was like my father in a way and yet was not my father
5. also there was a stripling page who turned into a maid
6. this was so sweet a lady sir and in some manner i do think
7. but then the picture was gone as quickly as it came
8. sister nell do you hear these marvels
9. take your place and let us see what the crystal can show you
10. like as not young master though i am an old man
11. he was going home after victory
12. it was almost buried now in flowers and foliage
13. But I wrestled with this fellow
14. but he saw nothing that moved no signal lights twinkled
15. and why should that disturb me let him enter
16. there was not a single note of gloom
17. boats put out both from the fort and the shore
18. his excellency madam the prefect
19. so i did push this fellow
20. what do i care for food
21. shame on you citizens cried he i blush for my fellows
22. surely we can submit with good grace
23. fine for you to talk old man answered the lean
24. at the same time every avenue of the throne was assaulted
25. vintage years have much to do with the quality of wines
26. come to me men here here he raised his voice still louder
27. dry and of magnificent bouquet
28. pour mayonnaise over all chill and serve
29. set into a cold place to chill and become firm
30. when thickened strain and cool

A.6 Wave2Letter: Seen Data (LIBRISPEECH Audiobooks [218])

1. last two days of the voyage bartley found almost intolerable
2. i never dreamed it would be you bartley
3. the cuisine is the best and the chefs rank at the top of the art
4. he pulled up a window as if the air were heavy
5. it it hasnt always made you miserable has it
6. always but its worse now
7. it's unbearable it tortures me every minute
8. i get nothing but misery out of either
9. there is this deception between me and everything

10. he dropped back heavily into his chair by the fire
11. i have thought about it until i am worn out
12. after the very first
13. we never planned to meet and when we met
14. i dont know what becomes of the ladies
15. but now it doesnt seem to matter very much
16. presently it stole back to his coat sleeve
17. yes hilda i know that he said simply
18. i understand bartley i was wrong
19. season with salt and pepper and a little sugar to taste
20. you want me to say it she whispered
21. what alternative was there for her
22. its got to be a clean break hilda
23. oh bartley what am i to do
24. you ask me to stay away from you because you want me
25. i will ask the least imaginable but i must have something
26. you see the treatment is a trifle fanciful
27. he protected her and she strengthened him
28. and then you came back not caring very much
29. dont cry dont cry he whispered
30. a little attack of nerves possibly

A.7 Common Unseen Data (MacKenzie & Soukoreff Dataset [187])

1. my watch fell in the water
2. prevailing wind from the east
3. never too rich and never too thin
4. breathing is difficult
5. I can see the rings on Saturn
6. physics and chemistry are hard
7. my bank account is overdrawn
8. elections bring out the best
9. you are a wonderful example
10. do not squander your time
11. do not drink too much
12. take a coffee break
13. popularity is desired by all
14. the music is better than it sounds
15. I agree with you

16. do not say anything
17. play it again Sam
18. the force is with you
19. we went grocery shopping
20. the assignment is due today
21. what you see is what you get
22. for your information only
23. a quarter of a century
24. the store will close at ten
25. head shoulders knees and toes
26. always cover all the bases
27. this is a very good idea
28. can we play cards tonight
29. get rid of that immediately
30. public transit is much faster