Design and Evaluation of a Silent Speech-Based Selection Method for Eye-Gaze Pointing

LAXMI PANDEY, Inclusive Interaction Lab, University of California, Merced, United States
AHMED SABBIR ARIF, Inclusive Interaction Lab, University of California, Merced, United States

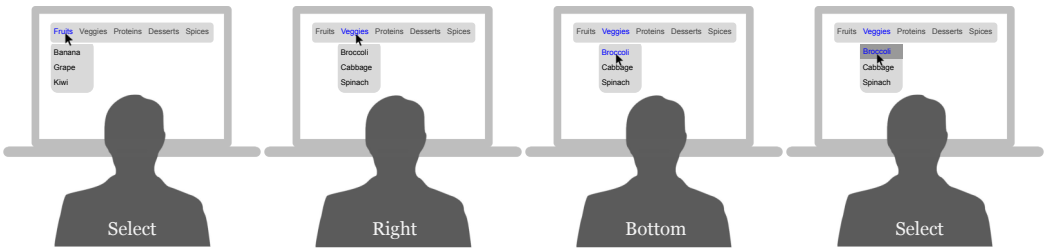


Fig. 1. A menu selection scenario with the proposed method. To select "Broccoli", the user starts scanning the horizontal menu from the left. The system locks the cursor on the first item when the gaze is within 10 pixels of the item. The user silently speaks the command "Select" to expand the current 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.

We investigate silent speech as a hands-free selection method in eye-gaze pointing. 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.

CCS Concepts: • **Human-centered computing** → **Pointing**; *Natural language interfaces*; User interface design.

Additional Key Words and Phrases: Fitts' law, multi-modal, speech, silent speech, lip reading, dwell, eye tracking, pointing, selection

ACM Reference Format:
Laxmi Pandey and Ahmed Sabbir Arif. 2022. Design and Evaluation of a Silent Speech-Based Selection Method for Eye-Gaze Pointing. *Proc. ACM Hum.-Comput. Interact.* 6, ISS, Article 570 (December 2022), 26 pages. https://doi.org/10.1145/3567723

Authors' addresses: Laxmi Pandey, Inclusive Interaction Lab, University of California, Merced, 5200 N. Lake Road, Merced, California, United States, 95343, lpandey@ucmerced.edu; Ahmed Sabbir Arif, Inclusive Interaction Lab, University of California, Merced, 5200 N. Lake Road, Merced, California, United States, 95343, asarif@ucmerced.edu.

2573-0142/2022/12-ART570
https://doi.org/10.1145/3567723

1 INTRODUCTION

Eye-gaze-based interaction is a promising modality for faster and seamless hands-free (also known as contactless or touchless) interaction [108]. It enables people with limited motor skills to interact with computer systems without using the hands [5, 16, 26, 57, 71]. It is also beneficial in Situationally-Induced Impairments and Disabilities (SIID) [104, 125], when the hands are incapacitated due to reasons such as performing a secondary task, minor injuries, or unavailability of a keyboard [101]. 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 [53].

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 [69, 127]. 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 [69, 106, 126]. The most common application of eye tracking is to direct control a mouse cursor using eye movements [127]. 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 [55], 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 [42] 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 [14, 42]. 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 [56]. 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) [30–32, 99]. Besides, speech does not work well with people with severe speech disorders since it relies on the sound produced by the users [5, 16].

In this work, we investigate silent speech as an alternative selection method for eye-gaze pointing. Silent speech is an image-based language processing method that interprets users' lip movements into text [89]. We envision several benefits of using silent speech commands as a selection method. First, it does not require the use of external hardware since both eye tracking and silent speech recognition can occur through the same webcam. Second, silent speech does not rely on acoustic features, thus can be used in noisy places or in places where people do not want to be disturbed [90]. Although outside the scope of this work, silent speech can also accommodate people with speech disorders. The contribution of our work is three-fold. 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 experiment. 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.

2 RELATED WORK

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., [18, 97, 102]) and physical keys, buttons, and controllers (e.g., [70, 71, 83, 121]). In this section, we only cover hands-free selection methods that are accessible to people with limited motor skills.

2.1 Hands-Free Selection Methods

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 [42]. 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 [14, 42]. To address this, several works have enabled users to adjust the dwell time [76] or automatically adjusted dwell time based on user experience [82, 128]. 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. Drewes and Schmidt [29] explored gaze gestures with eye tracking, where users performed specific eye movements for target selection. Studies suggested users can perform complex gaze gestures intentionally [29, 50]. A follow-up study showed that gaze gestures can enable people with motor impairments to play online games [54]. Some have used specific types of gaze gestures (e.g., reverse crossing [34] and single gaze gestures [84]) and blinking [8] for target selection. However, performing intentional gaze gestures and blinking are unnatural [56], thus can cause users irritation and fatigue. Several works, in contrast, studied target selection through voluntary facial muscle activation [77, 117], brain signals [47], and foot pedals [79]. These methods use external and invasive hardware, thus not yet scalable in practical situations. Some have also attempted head gestures for target selection [79, 109, 110], 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 [93, 112]. These works either use a single command to confirm selection [15] or multiple commands to facilitate both pointing and selection [81, 107]. Speech is promising but unreliable in noisy places and users are often hesitant to use speech in public places [30–32, 99]. Besides, speech does not work well with people with severe speech disorder [5, 16].

2.2 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 [12, 80, 111, 129]. 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. Murata and Karwowski [83] positions the cursor at the center of a target by suppressing cursor movements caused by involuntary eye movements. Kammerer et al. [61] enable users to select a target my making a “click” sound when the cursor is over it.

Murata and Karwowski [83] 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 [61, 119].

2.3 Interaction with Silent Speech

Silent speech has not been well explored in user interfaces, presumably due to technological limitations, as the existing recognition models use expensive, invasive, or non-portable hardware, including electromagnetic articulography (EMA) [33, 38, 44], real-time magnetic resonance imaging (rtMRI) [91], electroencephalogram (EEG) [96], electromyography (EMG) [58–60, 62, 75, 105, 123], ultrasound imaging [27, 28, 35, 38, 44, 48, 49, 66], vibrational sensors of glottal activity [86, 94, 100, 118], speech motor cortex implants [10], and non-audible murmur microphones [45, 46, 85]. Recently, there have been some attempts to recognize speech from videos of mouth and tongue movements [3, 6, 9, 13, 20, 21, 21–23, 89, 95, 115, 116]. But these models are slow (takes ~5 seconds to recognize one word) and error prone (4–47% error rate) [89]. To the best of our knowledge, none have explored the possibility of using silent speech with gaze pointing.

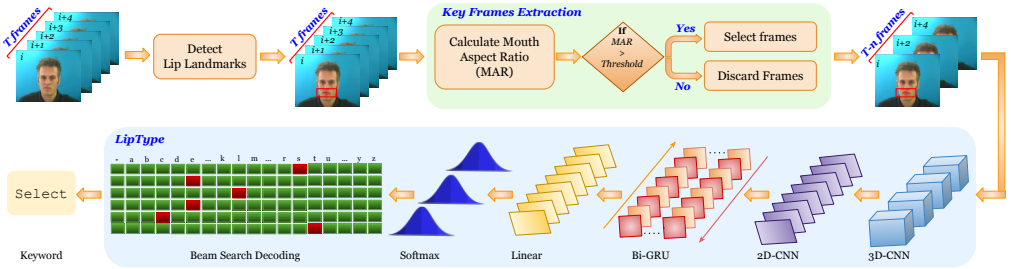


Fig. 2. The architecture of the proposed silent command recognition model. It pre-processes a sequence of T frames for mouth-centered cropped images to extract key frames. The key frames are fed to a 1-layer 3D CNN, followed by a 34-layer 2D SE-ResNet for spatiotemporal feature extraction. The features 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.

3 A MODEL FOR SILENT COMMAND RECOGNITION

We customized an existing silent speech recognition model LipType [89] 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 [25]). This increases the variability and ambiguity in lip movements (similar movements for different characters), which are disambiguated in post-processing using language models [9, 89]. 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% [89]. 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. 2 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 [67] and the iBug face landmark predictor [103] with 68 facial landmarks (L) and Kalman filtering (Fig. 3, 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. 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. 3, right).

$$MAR = \frac{\|L_{61} - L_{56}\| + \|L_{60} - L_{57}\| + \|L_{59} - L_{58}\|}{2 * \|L_{44} - L_{50}\|} \quad (1)$$

Module 2: Spatiotemporal feature extraction. This module passes the extracted key frames to a 3D-CNN with a kernel dimension of T:5 × W:7 × H:7, followed by Batch Normalization (BN) [52] and Rectified Linear Units (ReLU) [4]. 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) [19]. 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 [41]. The output is then decoded with a left-to-right beam search [24] using the Stanford-CTC decoder [72] to recognize spoken keywords.

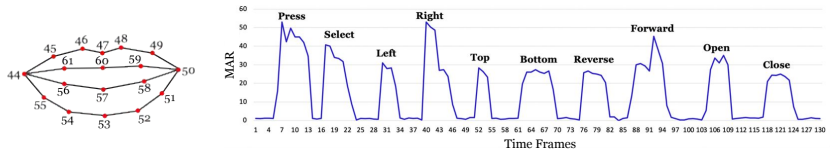


Fig. 3. From left, lip landmarks detected by DLib and iBug [67], and average mouth aspect ratios (MAR) of the ten keywords.

3.1 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

Table 1. Average recognition time (seconds) and accuracy rates (%) for the investigated models. "Sp." represents "Speech" and "Com." represents "Command".

Unit	Method	*Press*	*Select*	*Left*	*Right*	*Top*	*Bottom*	*Reverse*	*Forward*	*Open*	*Close*
Time	*Google Sp.*	1.73	1.64	1.65	1.54	1.69	1.82	1.82	1.68	1.65	1.61
	Kaldi Sp.	2.27	2.17	2.30	2.19	2.10	2.02	2.08	2.13	2.33	2.14
	Silent Com.	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 Sp.*	97.92	97.71	98.11	98.36	98.18	97.42	98.15	98.53	97.97	97.82
	Kaldi Sp.	88.05	88.65	90.19	87.48	89.04	88.41	85.83	88.04	89.62	88.02
	Silent Com.	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

zeros. We applied a channel-wise dropout [114] of 0.3. The model was trained end-to-end by the Adam optimizer [68] 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 [2] as the backend. Wll models were trained and tested on an NVIDIA GeForce 1080Ti GPU board. The source code[1] and the training dataset[2] are freely available to download.

3.2 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 [40], Kaldi (Api.ai) [98]) and silent speech (LipType [89]) 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 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.

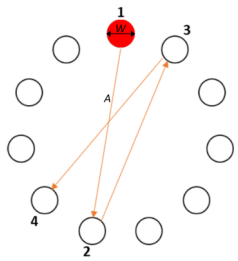
3.2.1 Results. On average, Google Speech-to-Text and Kaldi took 1.68 seconds (SD = 0.27) and 2.17 seconds (SD = 0.42), respectively, to recognize the keywords, whereas LipType and Silent command took 3.14 seconds (SD = 0.39) and 1.97 seconds (SD = 0.34), respectively. The differences were statistically 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 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

[1]Source code: https://github.com/theiilab/Eye-Gaze-Pointing
[2]Dataset: https://www.theiilab.com/resources/Keywords_Data.zip

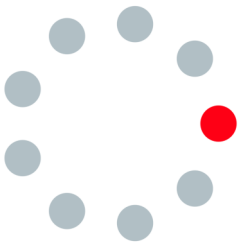
command. Silent command was almost as fast as Google Speech-to-Text (1.97 vs. 1.68 seconds) but was about 24% more error prone. However, this rate was recorded in a quiet room, while research showed that the accuracy rate of speech drops by 45–55% in presence of a background noise (42–58 db) [89]. The 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 [90]. 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.

4 EYE TRACKING

This work uses the GazeCloudAPI for real-time eye-tracking using a webcam [36]. 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 [39]. Afterward, a Haar-like feature classifier is used to classify the images into face and non-face regions [122]. 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 [65]. Finally, the point of gaze is estimated using the pupil location [37]. In an empirical evaluation [120], 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 [120].



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



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

Fig. 4. (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.

5 FITTS' LAW PROTOCOL

Fitts' law is a well-established method for evaluating target selection on computing systems [74]. 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 [113]. 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. 4a). 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 [51] 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.

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

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.

$$W_e = 4.133 * SD_x$$

6 EXPERIMENTAL SYSTEM

We developed a custom web application[3] with HTML5, CSS, PHP, and JavaScript for the Fitts' law experimental protocol (Section 5). It enables users to control a cursor with eye-gaze 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 4). We used it instead of other APIs [92, 126] due to its robustness [124]. 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 [18, 73, 82].
- **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.

7 USER STUDY 1: FITTS' LAW

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

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

7.1 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 6 (Fig. 4b) and the apparatus described in Section 3.2.

Fig. 5. Three participants taking part in the first user study.

7.2 Design

The experiment 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 [80], 70 pixels is the recommended width in eye tracking applications [121], while targets with widths over 220 pixels are unrealistic. The dependent variables in the experiment were as follows:

- **Throughput** (bps) as described in Section 5.
- **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.

7.3 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. 5), as recommended in eye tracking research [120]. 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).

7.4 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 7.2). We used a Friedman test for the questionnaire data [7]. We did not identify any effects of the between-subjects factors, namely age, gender, and the use of corrective eyeglasses or contact lenses.

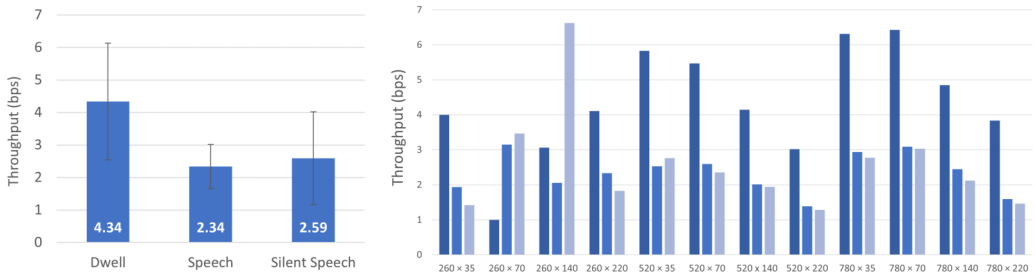
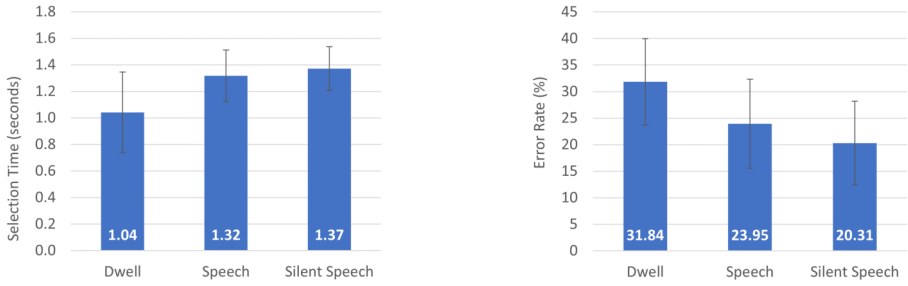


Fig. 6. Average throughput (bits/second) by (a) selection method and (b) selection method, amplitude, and width. Error bars represent ±1 standard deviation (SD).

7.4.1 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. 6a). 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. 6b illustrates average throughput by selection method, amplitude, and width.



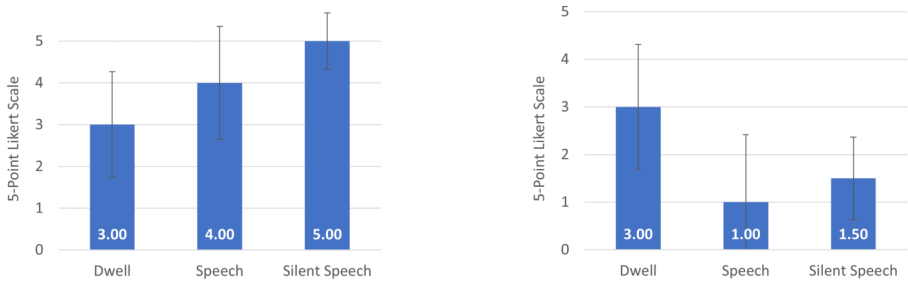
(a) Selection time (seconds)

(b) Error rate (%)

Fig. 7. Average selection time and error rate by selection method. Error bars represent ±1 standard deviation (SD).

7.4.2 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. 7a).

7.4.3 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. 7b).



(a) Willingness-to-use

(b) Physical and mental effort

Fig. 8. Median willingness-to-use and physical and mental effort. Error bars represent ±1 standard deviation (SD).

7.4.4 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 presents median willingness-to-use and perceived physical and mental effort ratings of the three methods.

7.5 Discussion

Results confirmed that target amplitude and width influence the selection methods in accordance to the Fitts' law (Fig. 6b), 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. 7a) and error rate (Fig. 7b). 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. 9 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). 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. 8b), but this effect was not statistically significant. These results identify silent speech as an effective selection method in eye-gaze pointing.

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

Fig. 9. Cursor trace examples for the three selection methods (A:520 × W:70 pixels).

8 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.

8.1 Participants

Twelve volunteers (M = 27.75 years, SD = 4.11) participated in the second study (Fig. 10b). 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.

8.2 Apparatus, Design, & Procedure

The study used the apparatus described in Section 3.2. 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. 10a). 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

(a) (b)

Fig. 10. (a) The twelve zones used in the second study and (b) two participants taking part in the study.

= 1,728 targets. The independent variable was "zone" and dependent variables were throughput, pointing time, and error rate (Section 7.2). This study used the same procedure as the first study (Section 7.3) 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. 10b). 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 (%)

Fig. 11. Average throughput, pointing time, and error rate per zone.

8.3 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. 11 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 [63]. 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"*.

9 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 [11, 88] and using six silent speech commands for cursor positioning and target selection. Target gravity uses a snap-to effect [88] 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 [83]. 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 2). Fig. 1 illustrates a menu selection scenario with the proposed system.

Table 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

10 USER STUDY 3: MENU SELECTION

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

10.1 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 3.2.

10.1.1 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

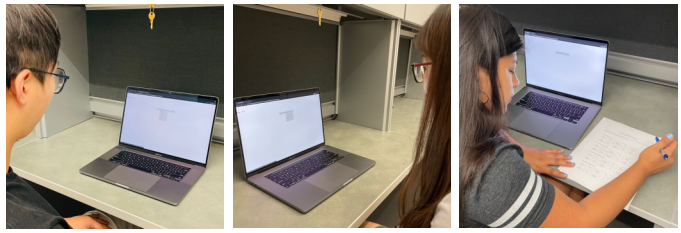


Fig. 12. Three participants taking part in the final user study.

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 [78]. 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 [17]. The menus were designed following the macOS guidelines [1] to provide a familiar look-and-feel. Each menu item was 150×38 pixels. Current items were highlighted in a blue font (Fig. 13) and selected items were highlighted in a dark gray background (Fig. 1).

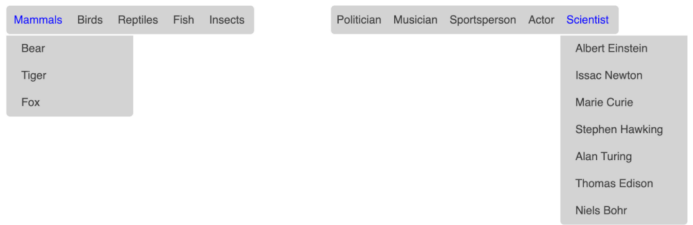


Fig. 13. Examples of two menus categorizing different types of animals and famous people.

10.2 Design & Procedure

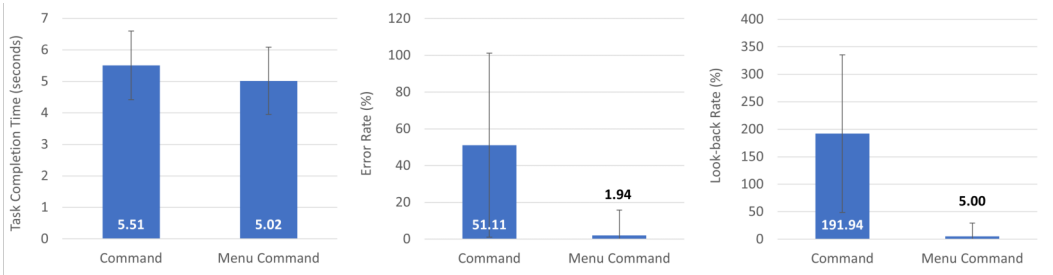
The study used the following within-subjects design: 12 participants × 2 methods (command, menu command, counterbalanced) × 2 unique menus per method × 15 tasks per menu = 720 menu selection tasks. The independent variable was "method" and dependent variables were as follows:

- **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 7.3). 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 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. 1, 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 [43] to rate the methods' perceived performance, usability, and workload.

10.3 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. [7] We did not identify any effects of the between-subjects factors, namely age, gender, and the use of corrective eyeglasses.



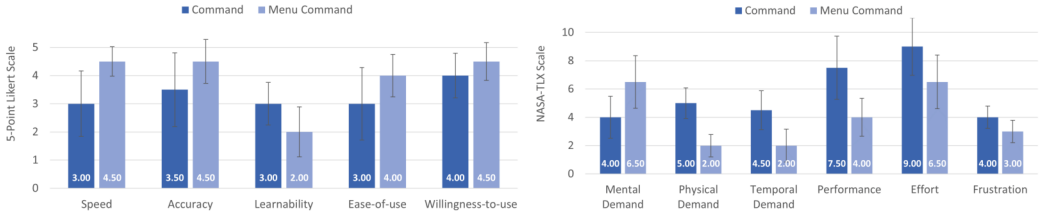
(a) Task completion time (seconds) (b) Error rate (%) (c) Look-back rate (%)

Fig. 14. Average task completion time, error rate, and look-back rate for the two investigated methods. Error bars represent ±1 standard deviation (SD).

10.3.1 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 5.51 (SD = 1.09) and 5.02 seconds (SD = 1.07), respectively (Fig. 14a).

10.3.2 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. 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. 14c).



(a) Usability questionnaire (b) NASA-TLX questionnaire

Fig. 15. Median willingness-to-use and physical and mental effort of the examined selection methods. Error bars represent ±1 standard deviation (SD).

10.3.3 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. 15a presents median perceived performance and usability ratings of both methods.

10.3.4 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. 15b presents median perceived workload ratings of both methods.

10.4 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. 14b). Menu command also yielded 97% lower look-back rate than the baseline (Fig. 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 experimental 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 [64]. 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 [87]. Our findings, however, are in line with a follow-up work that failed to replicate Nilsen [87]'s findings and argued that visual search and cursor movement strategies employed by actual users cannot be characterized easily [17].

Participants perceived the proposed method significantly faster and more accurate than the baseline (Fig. 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. 15b). They also felt that the method was better performed, required less effort, and caused less frustration than the baseline.

10.5 Limitations

Although the proposed approach is aimed at people that are unable use the hands due to a permanent or situational impairment, the studies recruited non-disabled people. While it is very likely that the quantitative findings are generalizable to the target population [130], it cannot be claimed with utmost certainty that the subjective feedback are also generalizable as people with disabilities might prefer a different method more due to lived experiences. Another limitation is that the study did not explore the effects of error correction on performance and preference. We decided not to force error correction in the study as it would have substantially increased the task completion time of the baseline condition, causing much frustration among the participants.

11 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 both 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.

12 CONCLUSION

In this work, we systematically studied the feasibility of using silent speech as a hands-free selection method in eye-gaze pointing. First, we proposed a stripped-down image-based model to recognize silent speech commands. An evaluation revealed that the model can recognize ten keywords almost as fast as a state-of-the-art speech recognition model. Second, we compared the method with other hands-free selection methods, namely dwell and speech, in a Fitts' law study, where both eye tracking and silent speech recognition used a webcam. Results showed that speech and silent speech are comparable in throughput and selection time, but the latter is significantly more accurate than the other methods. Besides, participants were significantly more enthusiastic about using silent speech than the other methods. We then conducted a follow-up study, which 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 presented a new method for selecting menu items with eye-gaze and silent speech commands. In a comparative evaluation, the

method was significantly faster and more accurate than the baseline. Participants also found the method significantly better in terms of performance, usability, and perceived workload.

13 FUTURE DIRECTIONS

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.

REFERENCES

[1] 2022. Menu Anatomy - Menus - macOS - Human Interface Guidelines - Apple Developer. https://developer.apple.com/design/human-interface-guidelines/macos/menus/menu-anatomy

[2] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mane, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viegas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv:1603.04467 [cs]* (March 2016). http://arxiv.org/abs/1603.04467 arXiv: 1603.04467.

[3] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. Deep Lip Reading: A Comparison of Models and an Online Application. *arXiv:1806.06053 [cs]* (June 2018). http://arxiv.org/abs/1806.06053 arXiv: 1806.06053.

[4] Abien Fred Agarap. 2019. Deep Learning using Rectified Linear Units (ReLU). (2019). http://arxiv.org/abs/1803.08375

[5] Ayush Agarwal, Dv JeevithaShree, Kamalpreet Singh Saluja, Atul Sahay, Pullikonda Mounika, Anshuman Sahu, Rahul Bhaumik, Vinodh Kumar Rajendran, and Pradipta Biswas. 2019. Comparing Two Webcam-Based Eye Gaze Trackers for Users with Severe Speech and Motor Impairment. In *Research into Design for a Connected World*, Amaresh Chakrabarti (Ed.). Vol. 135. Springer Singapore, Singapore, 641–652. https://doi.org/10.1007/978-981-13-5977-4_54 Series Title: Smart Innovation, Systems and Technologies.

[6] Ibrahim Almajai, Stephen Cox, Richard Harvey, and Yuxuan Lan. 2016. 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)*. 2722–2726. https://doi.org/10.1109/ICASSP.2016.7472172 ISSN: 2379-190X.

[7] Ahmed Sabbir Arif. 2021. Statistical Grounding. In *Intelligent Computing for Interactive System Design: Statistics, Digital Signal Processing, and Machine Learning in Practice* (1 ed.). Association for Computing Machinery, New York, NY, USA, 59–99. https://doi.org/10.1145/3447404.3447410

[8] Behrooz Ashtiani and I. Scott MacKenzie. 2010. BlinkWrite2: An Improved Text Entry Method Using Eye Blinks. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. Association for Computing Machinery, New York, NY, USA, 339–345. https://doi.org/10.1145/1743666.1743742

[9] Yannis M. Assael, Brendan Shillingford, Shimon Whiteson, and Nando de Freitas. 2016. LipNet: End-to-End Sentence-level Lipreading. *arXiv:1611.01599 [cs]* (Dec. 2016). http://arxiv.org/abs/1611.01599 arXiv: 1611.01599.

[10] Jess Bartels, D. Andreasen, P. Ehirim, Hui Mao, and P. Kennedy. 2008. Neurotrophic Electrode: Method of Assembly and Implantation into Human Motor Speech Cortex. *Journal of Neuroscience Methods* (2008). https://doi.org/10.1016/j.jneumeth.2008.06.030

[11] Scott Bateman, Regan Mandryk, Carl Gutwin, and Robert Xiao. 2009. *Investigation of Targeting-Assistance Techniques for Distant Pointing with Relative Ray Casting*. Technical Report 2009-03. University of Saskatchewan, Saskatoon, SK, Canada. 10 pages.

[12] Richard Bates and Howell Istance. 2002. Zooming Interfaces! Enhancing the Performance of Eye Controlled Pointing Devices. In *Proceedings of the fifth international ACM conference on Assistive technologies (Assets '02)*. Association for

Computing Machinery, New York, NY, USA, 119–126. https://doi.org/10.1145/638249.638272
[13] Helen L. Bear and Richard Harvey. 2019. Alternative Visual Units for an Optimized Phoneme-Based Lipreading System. 18 (2019), 3870. https://doi.org/10.3390/app9183870
[14] Roman Bednarik, Tersia Gowases, and Markku Tukiainen. 2009. Gaze Interaction Enhances Problem Solving: Effects of Dwell-Time Based, Gaze-Augmented, and Mouse Interaction on Problem-Solving Strategies and User Experience. *Journal of Eye Movement Research* 3, 1 (June 2009). https://doi.org/10.16910/jemr.3.1.3 Number: 1.
[15] T. R. Beelders and P. J. Blignaut. 2012. Measuring the Performance of Gaze and Speech for Text Input. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 337–340. https://doi.org/10.1145/2168556.2168631
[16] Maria Borgestig, Jan Sandqvist, Richard Parsons, Torbjörn Falkmer, and Helena Hemmingsson. 2016. Eye Gaze Performance for Children with Severe Physical Impairments Using Gaze-Based Assistive Technology—a Longitudinal Study. *Assistive Technology* 28, 2 (April 2016), 93–102. https://doi.org/10.1080/10400435.2015.1092182 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10400435.2015.1092182.
[17] Michael D. Byrne, John R. Anderson, Scott Douglass, and Michael Matessa. 1999. Eye Tracking the Visual Search of Click-down Menus. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '99)*. Association for Computing Machinery, New York, NY, USA, 402–409. https://doi.org/10.1145/302979.303118
[18] Ishan Chatterjee, Robert Xiao, and Chris Harrison. 2015. Gaze+Gesture: Expressive, Precise and Targeted Free-Space Interactions. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/2818346.2820752
[19] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. (2014). http://arxiv.org/abs/1412.3555
[20] Joon Son Chung and Andrew Zisserman. 2016. Out of Time: Automated Lip Sync in the Wild. In *ACCV Workshops*. https://doi.org/10.1007/978-3-319-54427-4_19
[21] Joon Son Chung and Andrew Zisserman. 2017. Lip Reading in Profile. In *BMVC*. https://doi.org/10.5244/C.31.155
[22] Joon Son Chung and Andrew Zisserman. 2017. Lip Reading in the Wild. In *Computer Vision – ACCV 2016 (Lecture Notes in Computer Science)*, Shang-Hong Lai, Vincent Lepetit, Ko Nishino, and Yoichi Sato (Eds.). Springer International Publishing, Cham, 87–103. https://doi.org/10.1007/978-3-319-54184-6_6
[23] Joon Son Chung and Andrew Zisserman. 2018. Learning to Lip Read Words by Watching Videos. *Computer Vision and Image Understanding* 173 (Aug. 2018), 76–85. https://doi.org/10.1016/j.cviu.2018.02.001
[24] Ronan Collobert, Awni Hannun, and Gabriel Synnaeve. 2019. A Fully Differentiable Beam Search Decoder. (2019). arXiv:1902.06022 http://arxiv.org/abs/1902.06022
[25] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition. *The Journal of the Acoustical Society of America* 120, 5 (Oct. 2006), 2421–2424. https://doi.org/10.1121/1.2229005 Publisher: Acoustical Society of America.
[26] F. Corno, L. Farinetti, and I. Signorile. 2002. A Cost-Effective Solution for Eye-Gaze Assistive Technology. In *Proceedings. IEEE International Conference on Multimedia and Expo*, Vol. 2. 433–436 vol.2. https://doi.org/10.1109/ICME.2002.1035632
[27] B. Denby, Y. Oussar, G. Dreyfus, and M. Stone. 2006. Prospects for a Silent Speech Interface Using Ultrasound Imaging. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, Vol. 1. I–I. https://doi.org/10.1109/ICASSP.2006.1660033 ISSN: 2379-190X.
[28] B. Denby and M. Stone. 2004. Speech Synthesis from Real Time Ultrasound Images of the Tongue. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1. I–685. https://doi.org/10.1109/ICASSP.2004.1326078 ISSN: 1520-6149.
[29] Heiko Drewes and Albrecht Schmidt. 2007. Interacting with the Computer Using Gaze Gestures. In *Human-Computer Interaction – INTERACT 2007 (Lecture Notes in Computer Science)*, Cécilia Baranauskas, Philippe Palanque, Julio Abascal, and Simone Diniz Junqueira Barbosa (Eds.). Springer, Berlin, Heidelberg, 475–488. https://doi.org/10.1007/978-3-540-74800-7_43
[30] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. 2014. Voice Activated Personal Assistant: Acceptability of Use in the Public Space. In *Human Interface and the Management of Information. Information and Knowledge in Applications and Services (Lecture Notes in Computer Science)*, Sakae Yamamoto (Ed.). Springer International Publishing, Cham, 324–334. https://doi.org/10.1007/978-3-319-07863-2_32
[31] Aarthi Easwara Moorthy and Kim-Phuong L. Vu. 2015. Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. *International Journal of Human–Computer Interaction* 31, 4 (April 2015), 307–335. https://doi.org/10.1080/10447318.2014.986642 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2014.986642.
[32] Christos Efthymiou and Martin Halvey. 2016. Evaluating the Social Acceptability of Voice Based Smartwatch Search. In *Information Retrieval Technology*, Shaoping Ma, Ji-Rong Wen, Yiqun Liu, Zhicheng Dou, Min Zhang, Yi Chang, and Xin Zhao (Eds.). Vol. 9994. Springer International Publishing, Cham, 267–278. https://doi.org/10.1007/978-3-319-

48051-0_20 Series Title: Lecture Notes in Computer Science.

[33] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman. 2008. Development of a (silent) Speech Recognition System for Patients Following Laryngectomy. *Medical Engineering & Physics* 30, 4 (May 2008), 419–425. https://doi.org/10.1016/j.medengphy.2007.05.003

[34] Wenxin Feng, Ming Chen, and Margrit Betke. 2014. 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)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/2674396.2674443

[35] Victoria M. Florescu, L. Crevier-Buchman, B. Denby, T. Hueber, Antonia Colazo-Simon, Claire Pillot-Loiseau, P. Roussel-Ragot, C. Gendrot, and S. Quattrocchi. 2010. Silent Vs Vocalized Articulation for a Portable Ultrasound-Based Silent Speech Interface. In *INTERSPEECH*.

[36] GazeRecorder. 2016. GazeCloudAPI: Real-Time online Eye-Tracking API. https://gazerecorder.com/gazecloudapi/

[37] Muhammad Usman Ghani, Sarah Chaudhry, Maryam Sohail, and Muhammad Nafees Geelani. [n.d.]. GazePointer: A real time mouse pointer control implementation based on eye gaze tracking. In *INMIC* (2013-12). 154–159. https://doi.org/10.1109/INMIC.2013.6731342

[38] J. M. Gilbert, S. I. Rybchenko, R. Hofe, S. R. Ell, M. J. Fagan, R. K. Moore, and P. Green. 2010. Isolated Word Recognition of Silent Speech Using Magnetic Implants and Sensors. *Medical Engineering & Physics* 32, 10 (Dec. 2010), 1189–1197. https://doi.org/10.1016/j.medengphy.2010.08.011

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

[40] Google. 2017. Speech-to-Text: Automatic Speech Recognition. https://cloud.google.com/speech-to-text

[41] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks. In *Proceedings of the 23rd international conference on Machine learning* (Pittsburgh, Pennsylvania, USA, 2006-06-25) *(ICML '06)*. Association for Computing Machinery, 369–376. https://doi.org/10.1145/1143844.1143891

[42] John Paulin Hansen, Anders Sewerin Johansen, Dan Witzner Hansen, Kenji Itoh, and Satoru Mashino. 2003. Command Without a Click: Dwell Time Typing by Mouse and Gaze Selections. In *The 10th International Conference on Human-Computer Interaction*, M. Rauterberg (Ed.). IOS, Crete, Greece, 121–128.

[43] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*. Vol. 52. Elsevier, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

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

[45] Panikos Heracleous, Tomomi Kaino, H. Saruwatari, and K. Shikano. 2007. Unvoiced Speech Recognition Using Tissue-Conductive Acoustic Sensor. *EURASIP J. Adv. Signal Process.* (2007). https://doi.org/10.1155/2007/94068

[46] Tatsuya Hirahara, Makoto Otani, Shota Shimizu, Tomoki Toda, Keigo Nakamura, Yoshitaka Nakajima, and Kiyohiro Shikano. 2010. Silent-Speech Enhancement Using Body-Conducted Vocal-Tract Resonance Signals. *Speech Communication* 52, 4 (April 2010), 301–313. https://doi.org/10.1016/j.specom.2009.12.001

[47] Baosheng James Hou, Per Bekgaard, Scott MacKenzie, John Paulin Paulin Hansen, and Sadasivan Puthusserypady. 2020. GIMIS: Gaze Input with Motor Imagery Selection. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3379157.3388932

[48] T. Hueber, G. Aversano, G. Chollet, B. Denby, G. Dreyfus, Y. Oussar, P. Roussel, and M. Stone. 2007. 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. I–1245–I–1248. https://doi.org/10.1109/ICASSP.2007.366140 ISSN: 2379-190X.

[49] Thomas Hueber, Elie-Laurent Benaroya, Gérard Chollet, Bruce Denby, Gérard Dreyfus, and Maureen Stone. 2010. Development of a Silent Speech Interface Driven by Ultrasound and Optical Images of the Tongue and Lips. *Speech Communication* 52, 4 (April 2010), 288–300. https://doi.org/10.1016/j.specom.2009.11.004

[50] Aulikki Hyrskykari, Howell Istance, and Stephen Vickers. 2012. Gaze Gestures or Dwell-Based Interaction?. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '12)*. Association for Computing Machinery, New York, NY, USA, 229–232. https://doi.org/10.1145/2168556.2168602

[51] International Organization for Standardization. 2012. ISO/TS 9241-411:2012. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/41/54106.html

[52] Sergey Ioffe and Christian Szegedy. 2015. 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 (Lille, France, 2015-07-06) *(ICML'15)*. JMLR.org, 448–456.

[53] Muhammad Zahid Iqbal and Abraham Campbell. 2020. The Emerging Need for Touchless Interaction Technologies. *Interactions* 27, 4 (July 2020), 51–52. https://doi.org/10.1145/3406100

[54] Howell Istance, Aulikki Hyrskykari, Lauri Immonen, Santtu Mansikkamaa, and Stephen Vickers. 2010. Designing Gaze Gestures for Gaming: An Investigation of Performance. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. Association for Computing Machinery, New York, NY, USA, 323–330. https://doi.org/10.1145/1743666.1743740

[55] Robert J. K. Jacob. 1991. The Use of Eye Movements in Human-Computer Interaction Techniques: What You Look at Is What You Get. *ACM Transactions on Information Systems* 9, 2 (April 1991), 152–169. https://doi.org/10.1145/123078.128728

[56] Robert J. K. Jacob. 1995. Eye Tracking in Advanced Interface Design. In *Virtual Environments and Advanced Interface Design*, W Barfield and T. A. Furness (Eds.). University Press, New York, NY, USA, 258–288.

[57] D. V. Jeevithashree, Kamalpreet Singh Saluja, and Pradipta Biswas. 2019. A Case Study of Developing Gaze Controlled Interface for Users with Severe Speech and Motor Impairment. *Technology and Disability* 31, 1-2 (Jan. 2019), 63–76. https://doi.org/10.3233/TAD-180206 Publisher: IOS Press.

[58] Charles Jorgensen and Sorin Dusan. 2010. Speech Interfaces Based Upon Surface Electromyography. *Speech Communication* 52, 4 (April 2010), 354–366. https://doi.org/10.1016/j.specom.2009.11.003

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

[60] S. Jou, Tanja Schultz, Matthias Walliczek, F. Kraft, and Alexander H. Waibel. 2006. Towards Continuous Speech Recognition Using Surface Electromyography. In *INTERSPEECH*.

[61] Yvonne Kammerer, Katharina Scheiter, and Wolfgang Beinhauer. 2008. 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)*. Association for Computing Machinery, New York, NY, USA, 213–220. https://doi.org/10.1145/1344471.1344522

[62] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. Alterego: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces (IUI '18)*. Association for Computing Machinery, New York, NY, USA, 43–53. https://doi.org/10.1145/3172944.3172977

[63] Anuradha Kar and Peter Corcoran. 2018. Performance Evaluation Strategies for Eye Gaze Estimation Systems with Quantitative Metrics and Visualizations. *Sensors* 18, 9 (Sept. 2018), 3151. https://doi.org/10.3390/s18093151 Number: 9 Publisher: Multidisciplinary Digital Publishing Institute.

[64] A.E. Kaufman, A. Bandopadhay, and B.D. Shaviv. 1993. An Eye Tracking Computer User Interface. In *Proceedings of 1993 IEEE Research Properties in Virtual Reality Symposium*. 120–121. https://doi.org/10.1109/VRAIS.1993.378254

[65] Carolyn Kimme, Dana Ballard, and Jack Sklansky. 1975. Finding Circles by an Array of Accumulators. *Commun. ACM* 18, 2 (feb 1975), 120–122. https://doi.org/10.1145/360666.360677

[66] Naoki Kimura, Michinari Kono, and Jun Rekimoto. 2019. 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)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376

[67] Davis E. King. 2009. Dlib-Ml: A Machine Learning Toolkit. *The Journal of Machine Learning Research* 10 (Dec. 2009), 1755–1758.

[68] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. (2017). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[69] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. 2016. Eye Tracking for Everyone. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 2176–2184. https://doi.org/10.1109/CVPR.2016.239

[70] Chandan Kumar, Ramin Hedeshy, I. Scott MacKenzie, and Steffen Staab. 2020. TAGSwipe: Touch Assisted Gaze Swipe for Text Entry. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376317

[71] Chandan Kumar, Raphael Menges, Daniel Müller, and Steffen Staab. 2017. 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)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 219–223. https://doi.org/10.1145/3041021.3054730

[72] Andrew L. Maas, Ziang Xie, Dan Jurafsky, and A. Ng. 2015. Lexicon-Free Conversational Speech Recognition with Neural Networks. In *HLT-NAACL*. https://doi.org/10.3115/v1/N15-1038

[73] I. Scott MacKenzie. 2012. Evaluating Eye Tracking Systems for Computer Input. , 205–225 pages. https://doi.org/10.4018/978-1-61350-098-9.ch015 ISBN: 9781613500989.

[74] I. Scott MacKenzie. 2018. Fitts' Law. In *The Wiley Handbook of Human Computer Interaction*. John Wiley & Sons, Ltd, Hoboken, NJ, USA, 347–370. https://doi.org/10.1002/9781118976005.ch17

[75] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. 2005. Session Independent Non-Audible Speech Recognition Using Surface Electromyography. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* 331–336. https://doi.org/10.1109/ASRU.2005.1566521

[76] Päivi Majaranta, Ulla-Kaija Ahola, and Oleg Špakov. 2009. Fast Gaze Typing with an Adjustable Dwell Time. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 357–360. https://doi.org/10.1145/1518701.1518758

[77] Julio C. Mateo, Javier San Agustin, and John Paulin Hansen. 2008. Gaze Beats Mouse: Hands-Free Selection by Combining Gaze and EMG. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 3039–3044. https://doi.org/10.1145/1358628.1358804

[78] George A. Miller. 1956. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review* 63, 2 (1956), 81–97. https://doi.org/10.1037/h0043158 Place: US Publisher: American Psychological Association.

[79] Katsumi Minakata, John Paulin Hansen, I. Scott MacKenzie, Per Bækgaard, and Vijay Rajanna. 2019. 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)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3317956.3318150

[80] Darius Miniotas, Oleg Špakov, and I. Scott MacKenzie. 2004. Eye Gaze Interaction with Expanding Targets. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems (CHI EA '04)*. Association for Computing Machinery, New York, NY, USA, 1255–1258. https://doi.org/10.1145/985921.986037

[81] Darius Miniotas, Oleg Špakov, Ivan Tugoy, and I. Scott MacKenzie. 2006. Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets. In *Proceedings of the 2006 symposium on Eye tracking research & applications (ETRA '06)*. Association for Computing Machinery, New York, NY, USA, 67–72. https://doi.org/10.1145/1117309.1117345

[82] Martez E. Mott, Shane Williams, Jacob O. Wobbrock, and Meredith Ringel Morris. 2017. Improving Dwell-Based Gaze Typing with Dynamic, Cascading Dwell Times. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2558–2570. https://doi.org/10.1145/3025453.3025517

[83] Atsuo Murata and Waldemar Karwowski. 2019. Automatic Lock of Cursor Movement: Implications for an Efficient Eye-Gaze Input Method for Drag and Menu Selection. *IEEE Transactions on Human-Machine Systems* 49, 3 (June 2019), 259–267. https://doi.org/10.1109/THMS.2018.2884737 Conference Name: IEEE Transactions on Human-Machine Systems.

[84] Emilie Møllenbach, Martin Lillholm, Alastair Gail, and John Paulin Hansen. 2010. Single Gaze Gestures. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. Association for Computing Machinery, New York, NY, USA, 177–180. https://doi.org/10.1145/1743666.1743710

[85] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. 2003. 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. V–708. https://doi.org/10.1109/ICASSP.2003.1200069 ISSN: 1520-6149.

[86] L.C. Ng, G.C. Burnett, J.F. Holzrichter, and T.J. Gable. 2000. 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. 229–232 vol.1. https://doi.org/10.1109/ICASSP.2000.861925 ISSN: 1520-6149.

[87] Erik Lloyd Nilsen. 1991. *Perceptual-Motor Control in Human-Computer Interaction.* Ph.D. University of Michigan, Ann Arbor, MI, USA. https://www.proquest.com/docview/303945464/abstract/683DEEF3C2344476PQ/1

[88] Ian Oakley, Marilyn Rose McGee, Stephen Brewster, and Philip Gray. 2000. Putting the Feel in 'Look and Feel'. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '00)*. Association for Computing Machinery, New York, NY, USA, 415–422. https://doi.org/10.1145/332040.332467

[89] Laxmi Pandey and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Yokohama, Japan, 19 pages. https://doi.org/10.1145/3411764.3445565

[90] 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 CHI Conference on Human Factors in Computing Systems (CHI '21)*. ACM, New York, NY, USA, Yokohama, Japan, 13 pages. https://doi.org/10.1145/3411764.3445430

[91] 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, 1–2. https://doi.org/10.1145/3450618.3469176
[92] Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. [n.d.]. WebGazer: Scalable Webcam Eye Tracking Using User Interactions. ([n. d.]), 7.
[93] Mohsen Parisay, Charalambos Poullis, and Marta Kersten. 2021. EyeTAP: A Novel Technique using Voice Inputs to Address the Midas Touch Problem for Gaze-based Interactions. *International Journal of Human-Computer Studies* 154 (Oct. 2021), 102676. https://doi.org/10.1016/j.ijhcs.2021.102676 arXiv: 2002.08455.
[94] Sanjay A. Patil and John H. L. Hansen. 2010. The Physiological Microphone (pmic): A Competitive Alternative for Speaker Assessment in Stress Detection and Speaker Verification. *Speech Communication* 52, 4 (April 2010), 327–340. https://doi.org/10.1016/j.specom.2009.11.006
[95] Stavros Petridis and Maja Pantic. 2016. Deep Complementary Bottleneck Features for Visual Speech Recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2304–2308. https://doi.org/10.1109/ICASSP.2016.7472088 ISSN: 2379-190X.
[96] Anne Porbadnigk, Marek Wester, Jan Calliess, and Tanja Schultz. 2009. EEG-based Speech Recognition - Impact of Temporal Effects. In *BIOSIGNALS*. https://doi.org/10.5220/0001554303760381
[97] Matti Pouke, Antti Karhu, Seamus Hickey, and Leena Arhippainen. 2012. 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)*. Association for Computing Machinery, New York, NY, USA, 505–512. https://doi.org/10.1145/2414536.2414614
[98] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit. https://infoscience.epfl.ch/record/192584 Conference Name: IEEE 2011 Workshop on Automatic Speech Recognition and Understanding Number: CONF Publisher: IEEE Signal Processing Society.
[99] S. Prabhakar, S. Pankanti, and A.K. Jain. 2003. Biometric Recognition: Security and Privacy Concerns. *IEEE Security Privacy* 1, 2 (March 2003), 33–42. https://doi.org/10.1109/MSECP.2003.1193209 Conference Name: IEEE Security Privacy.
[100] T.F. Quatieri, K. Brady, D. Messing, J.P. Campbell, W.M. Campbell, M.S. Brandstein, C.J. Weinstein, J.D. Tardelli, and P.D. Gatewood. 2006. Exploiting Nonacoustic Sensors for Speech Encoding. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 2 (March 2006), 533–544. https://doi.org/10.1109/TSA.2005.855838 Conference Name: IEEE Transactions on Audio, Speech, and Language Processing.
[101] Vijay Rajanna and Tracy Hammond. 2018. 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)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3204493.3208344
[102] David Rozado, Jeremy Hales, and Diako Mardanbegi. 2013. Interacting with Objects in the Environment by Gaze and Hand Gestures. In *Proceedings of the 3rd International Workshop on Pervasive Eye Tracking and Mobile Eye-Based Interaction*. ECEM, New York, NY, USA, 1–9.
[103] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2013. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. In *2013 IEEE International Conference on Computer Vision Workshops*. 397–403. https://doi.org/10.1109/ICCVW.2013.59
[104] Zhanna Sarsenbayeva, Vassilis Kostakos, and Jorge Goncalves. 2019. Situationally-Induced Impairments and Disabilities Research. *arXiv:1904.06128 [cs]* (April 2019). http://arxiv.org/abs/1904.06128 arXiv: 1904.06128.
[105] Tanja Schultz and Michael Wand. 2010. Modeling Coarticulation in Emg-Based Continuous Speech Recognition. *Speech Communication* 52, 4 (April 2010), 341–353. https://doi.org/10.1016/j.specom.2009.12.002
[106] Kilian Semmelmann and Sarah Weigelt. 2018. Online Webcam-Based Eye Tracking in Cognitive Science: A First Look. *Behavior Research Methods* 50, 2 (April 2018), 451–465. https://doi.org/10.3758/s13428-017-0913-7
[107] Korok Sengupta, Min Ke, Raphael Menges, Chandan Kumar, and Steffen Staab. 2018. Hands-Free Web Browsing: Enriching the User Experience with Gaze and Voice Modality. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications (ETRA '18)*. Association for Computing Machinery, New York, NY, USA, 1–3. https://doi.org/10.1145/3204493.3208338
[108] Linda E. Sibert and Robert J. K. Jacob. 2000. Evaluation of Eye Gaze Interaction. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems (CHI '00)*. Association for Computing Machinery, New York, NY, USA, 281–288. https://doi.org/10.1145/332040.332445
[109] Ludwig Sidenmark and Hans Gellersen. 2019. 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)*. Association for Computing Machinery, New York, NY, USA, 1161–1174. https://doi.org/10.1145/3332165.3347921
[110] Ludwig Sidenmark, Diako Mardanbegi, Argenis Ramirez Gomez, Christopher Clarke, and Hans Gellersen. 2020. BimodalGaze: Seamlessly Refined Pointing with Gaze and Filtered Gestural Head Movement. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers)*. Association for Computing Machinery, New York, NY,

USA, 1–9. https://doi.org/10.1145/3379155.3391312

[111] Henrik Skovsgaard, Julio C. Mateo, John M. Flach, and John Paulin Hansen. 2010. Small-Target Selection with Gaze Alone. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. Association for Computing Machinery, New York, NY, USA, 145–148. https://doi.org/10.1145/1743666.1743702

[112] Malcolm Slaney, Rahul Rajan, Andreas Stolcke, and Partha Parthasarathy. 2014. Gaze-Enhanced Speech Recognition. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 3236–3240. https://doi.org/10.1109/ICASSP.2014.6854198 ISSN: 2379-190X.

[113] R. William Soukoreff and I. Scott MacKenzie. 2004. Towards a Standard for Pointing Device Evaluation, Perspectives on 27 Years of Fitts' Law Research in HCI. *International Journal of Human-Computer Studies* 61, 6 (Dec. 2004), 751–789. https://doi.org/10.1016/j.ijhcs.2004.09.001

[114] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The Journal of Machine Learning Research* 15, 1 (Jan. 2014), 1929–1958.

[115] Themos Stafylakis and Georgios Tzimiropoulos. 2017. Combining Residual Networks with LSTMs for Lipreading. *INTERSPEECH* (2017). https://doi.org/10.21437/INTERSPEECH.2017-85

[116] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. 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)*. Association for Computing Machinery, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599

[117] Veikko Surakka, Marko Illi, and Poika Isokoski. 2004. Gazing and Frowning as a New Human–Computer Interaction Technique. *ACM Transactions on Applied Perception* 1, 1 (July 2004), 40–56. https://doi.org/10.1145/1008722.1008726

[118] Ingo R. Titze, Brad H. Story, Gregory C. Burnett, John F. Holzrichter, Lawrence C. Ng, and Wayne A. Lea. 1999. Comparison Between Electroglottography and Electromagnetic Glottography. *The Journal of the Acoustical Society of America* 107, 1 (Dec. 1999), 581–588. https://doi.org/10.1121/1.428324 Publisher: Acoustical Society of America.

[119] Mario H. Urbina, Maike Lorenz, and Anke Huckauf. 2010. 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)*. Association for Computing Machinery, New York, NY, USA, 93–96. https://doi.org/10.1145/1743666.1743689

[120] SIMPLY USER. 2013. *The Comparison of Accuracy and Precision of Eye Tracking: GazeFlow vs. SMI RED 250*. Technical Report. SIMPLY USER, User Experience Lab, Kraków, Poland. 29 pages. https://gazerecorder.com/webcam-eye-tracking-accuracy

[121] Roel Vertegaal. 2008. 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)*. Association for Computing Machinery, New York, NY, USA, 241–248. https://doi.org/10.1145/1452392.1452443

[122] P. Viola and M. Jones. [n.d.]. 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* (2001-12), Vol. 1. I–I. https://doi.org/10.1109/CVPR.2001.990517 ISSN: 1063-6919.

[123] Michael Wand and Tanja Schultz. 2011. Session-Independent Emg-Based Speech Recognition. In *BIOSIGNALS*. SciTePress, Setúbal, Portugal, 295–300. https://doi.org/10.5220/0003169702950300

[124] William Wang. 2020. Integrating GazeCloudAPI, a High Accuracy Webcam Based Eye-Tracking Solution, into Your Own Web-App. https://medium.com/@williamwang15/integrating-gazecloudapi-a-high-accuracy-webcam-based-eye-tracking-solution-into-your-own-web-app-2d8513bb9865

[125] Jacob O. Wobbrock. 2019. Situationally Aware Mobile Devices for Overcoming Situational Impairments. In *Proceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems (EICS '19)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3319499.3330292

[126] Pingmei Xu, Krista A. Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. 2015. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking. *arXiv:1504.06755 [cs]* (May 2015). http://arxiv.org/abs/1504.06755 arXiv: 1504.06755.

[127] Xuebai Zhang, Xiaolong Liu, Shyan-Ming Yuan, and Shu-Fan Lin. 2017. Eye Tracking Based Control System for Natural Human-Computer Interaction. *Computational Intelligence and Neuroscience* 2017 (Dec. 2017), e5739301. https://doi.org/10.1155/2017/5739301 Publisher: Hindawi.

[128] Oleg Špakov and Darius Miniotas. 2004. On-Line Adjustment of Dwell Time for Target Selection by Gaze. In *Proceedings of the third Nordic conference on Human-computer interaction (NordiCHI '04)*. Association for Computing Machinery, New York, NY, USA, 203–206. https://doi.org/10.1145/1028014.1028045

[129] Oleg Špakov and Darius Miniotas. 2005. Gaze-Based Selection of Standard-Size Menu Items. In *Proceedings of the 7th international conference on Multimodal interfaces (ICMI '05)*. Association for Computing Machinery, New York, NY, USA, 124–128. https://doi.org/10.1145/1088463.1088486

[130] Boštjan Šumak, Matic Špindler, Mojca Debeljak, Marjan Heričko, and Maja Pušnik. 2019. An Empirical Evaluation of a Hands-Free Computer Interaction for Users with Motor Disabilities. *Journal of Biomedical Informatics* 96 (Aug. 2019), 103249. https://doi.org/10.1016/j.jbi.2019.103249