SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography

Naoki Kimura
The University of Tokyo
Tokyo, Japan
kimura-naoki@g.ecc.u-tokyo.ac.jp

Tan Gemicioglu, Jonathan Womack, Yuhui Zhao
Georgia Institute of Technology
Atlanta, GA, USA

Richard Li
University of Washington
Seattle, WA, USA

Abdelkareem Bedri
Carnegie Mellon University
Pittsburgh, PA, USA

Zixiong Su
The University of Tokyo
Tokyo, Japan

Alex Olwal
Google Research
Mountain View, CA, USA

Jun Rekimoto
The University of Tokyo
Tokyo, Japan

Thad Starner
Georgia Institute of Technology
Atlanta, USA

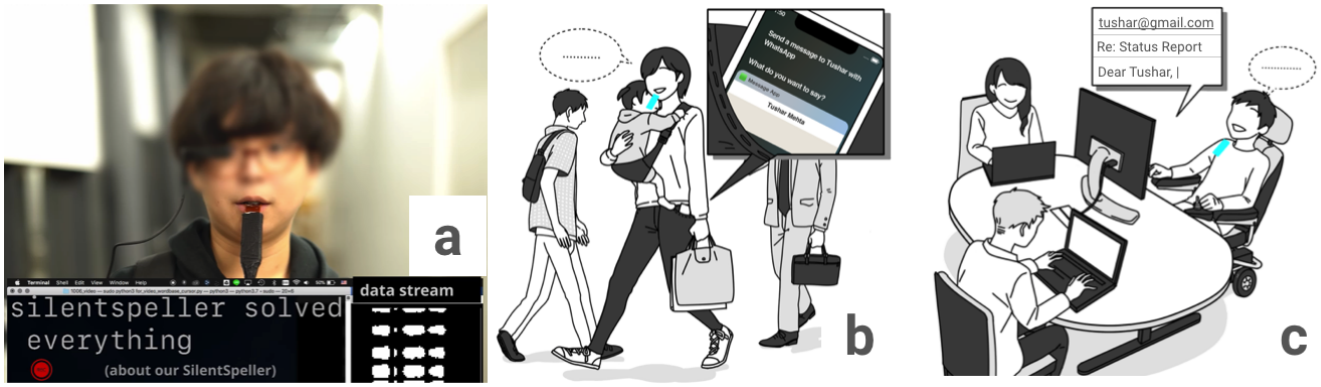

Figure 1: a) A SilentSpeller user wears the SmartPalate retainer whose 124 electrodes sense the position of the tongue at 100 Hz. Applications include b) hands-busy situations where speech is socially inappropriate and c) users with low manual dexterity working in an open office.

ABSTRACT

Speech is inappropriate in many situations, limiting when voice control can be used. Most unvoiced speech text entry systems can not be used while on-the-go due to movement artifacts. Using a dental retainer with capacitive touch sensors, SilentSpeller tracks tongue movement, enabling users to type by spelling words without voicing. SilentSpeller achieves an average 97% character accuracy in offline isolated word testing on a 1164-word dictionary. Walking has little effect on accuracy; average offline character accuracy was roughly equivalent on 107 phrases entered while walking (97.5%) or seated (96.5%). To demonstrate extensibility, the system was tested on 100 unseen words, leading to an average 94% accuracy. Live text entry speeds for seven participants averaged 37 words per minute at 87% accuracy. Comparing silent spelling to current practice suggests that SilentSpeller may be a viable alternative for silent mobile text entry.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Interaction devices**;

KEYWORDS

wearable computing, silent speech interface, text entry

ACM Reference Format:
Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Yuhui Zhao, Richard Li, Abdelkareem Bedri, Zixiong Su, Alex Olwal, Jun Rekimoto, and Thad Starner. 2022. SilentSpeller: Towards mobile, hands-free, silent speech text entry using electropalatography. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3491102.3502015

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3502015

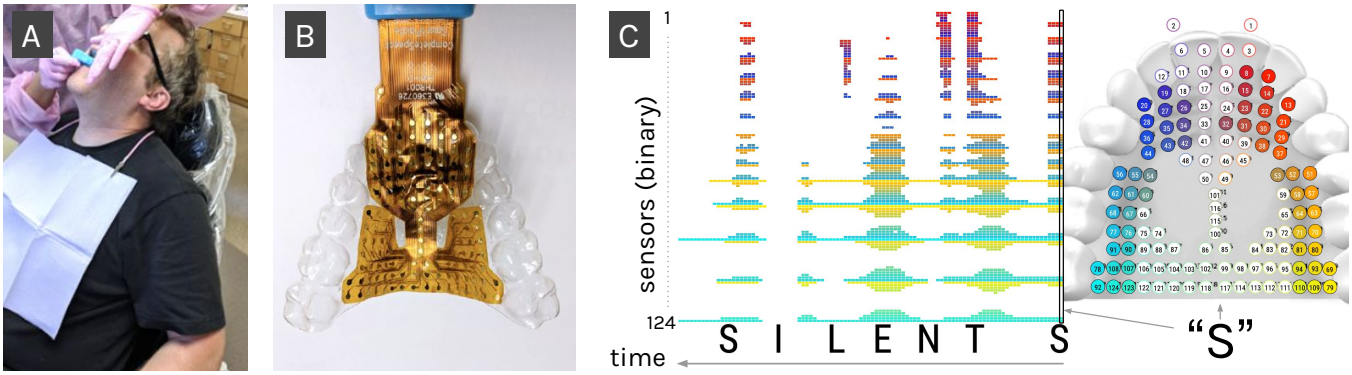

Figure 2: a) Dental impression needed for custom-fit SmartPalate, b) resulting SmartPalate, and c) Palatogram and electrode map. Note that individual letters are not recognized in real-time but are added to the image for illustration purposes.

1 INTRODUCTION

The following real-life scenario, articulated to one of the authors while consulting, initially motivated SilentSpeller:

Kim is a project manager for a large software company that maintains an open office plan (i.e., many workers in the same room with adjacent desks). Kim enjoys the camaraderie, and her team ensures that the aisles are clear so she can navigate her power wheelchair to her desk in the middle of the team (Figure 1c). Unfortunately, Kim's muscular dystrophy has weakened her hands so that she can no longer type her emails and documents. Speech recognition could help, but it would distract her colleagues at neighboring desks and raise privacy and security concerns given her managerial role. Kim is now searching for a text entry method which will maintain her efficacy and allow her to sit with her team.

Conditions such as amyotrophic lateral sclerosis (ALS), cerebral palsy, stroke, multiple sclerosis (MS), Parkinson's disease, essential tremor, and arthritis can limit a computer user's manual dexterity. Silent speech, which recognizes text entry via non-voiced speech, offers an alternative. Silent speech is able to offer the convenience and learnability of a speech interface while maintaining confidentiality. However, silent speech recognition is a very difficult task [11] that is often limited to a vocabulary of around 100 words and requires the user to be stationary [10, 20, 23, 27, 30, 36, 38, 40, 53, 58]. In addition, none of the silent speech systems in the literature have been tested for use in live input using a standard text entry corpus. Most papers only report results from offline experiments, which can lead to overly optimistic results due to overfitting. The few systems that report live results are designed to recognize command phrases or words and do not allow arbitrary composition of the words in the dictionary as is required for live text entry.

We introduce silent spelling as an alternative for silent speech interaction (SSI). In silent spelling, the user spells words without voicing (i.e., instead of saying them audibly). This technique is more easily recognized than silent speech, allowing larger vocabularies (1164 words in this work) and on-the-go interaction. Although spelling is slower than speech, we show that the speed is still comparable to virtual mini-QWERTY text entry on smartphones. We choose to compare to text entry on smartphones, as users have clearly accepted it as sufficiently fast to communicate in many situations [43]. It is worth noting that spelling requires less training than other text entry methods for fast, silent, and gaze-free mobile text entry. Unlike other systems [33, 35, 59], silent spelling is fast to learn; SilentSpeller users achieve 30 wpm in their first 20-minute session using the interface.

SilentSpeller detects spelling using a device in the form factor of a dental retainer that tracks the tongue at 100 Hz using 124 capacitive touch sensors on the roof of the mouth (Figures 1a and 2). The sensor is very robust to motion artifacts, enabling on-the-go scenarios. We compare text entry accuracy while seated and walking; the results are almost the same. This property allows applications for silent speech systems beyond the desktop.

1.1 Contributions

We offer the following contributions:

- **A Wizard of Oz study comparing silent spelling to smartphone mini-QWERTY texting** and silent speech in terms of words per minute and workload (NASA TLX). Results suggest silent spelling, in the ideal situation, may be a viable text entry alternative.
- **Optimization experiments** determining the amount of training data needed for a user dependent SilentSpeller recognizer. 2328 words (1164 unique words twice) were spelled by two participants. SilentSpeller reaches a maximum average 97% character accuracy (93% word accuracy) within 1500 words of training.
- **Isolated word experiments** to test the generalizability of SilentSpeller to unseen vocabulary. We remove 100 words (200 examples) from the training data set to attempt recognition on unseen words. SilentSpeller achieves an average of 94% accuracy (86% word).
- **Walking versus seated experiments** establish that SilentSpeller is tolerant to user movement during input with little degradation of performance (97.5% character accuracy walking, 96.5% seated).
- **An interactive text entry system** that combines the spelling recognizers with gestures for editing.

Table 1: Experiments conducted

No.	Purpose	Partic.	Words Collected	Dictionary	Testing	Result
Pilot	Wizard of Oz experiment on speed & usability	6	N/A	N/A	N/A	Figure 6
1	Recognition feasibility and optimizing models	2	2328	1164	10-fold cross-validation	Table 3 Figure 11
2	Recognition on unseen words	2	2328	1164	test on 100 random words (200 examples) left out	Table 4
3	Determining robustness to to motion artifacts	2	2328 + 556x2	321	training on 2328 plus 556 words walking or seating; test on other	Table 5
4	Examining performance as a text entry system	7	500 + 556	321	Live text entry	Table 6

- **Live text entry experiments** comparing SilentSpeller to standard smartphone virtual QWERTY text entry using the standard MacKenzie-Soukoreff phrase set.
- **A public database** of SilentSpeller input that includes 37,633 letters in 6325 words produced by seven users.

Table 1 summarizes the experiments presented.

2 RELATED WORK

Computing interaction has historically revolved around inputting text into a computing system for storage (i.e., taking notes) and action (i.e., running commands), and much previous research has focused on text entry [35]. While the QWERTY keyboard remains the primary modality, the transition from mechanical keyboards for desktop computing to touchscreen smartphones has caused a resurgence in interest in alternative text entry techniques for on-the-go scenarios. In addition, people with disabilities often seek appropriate alternative and augmentative communication (AAC) aids as text entry requires manual dexterity and visual attention they may not be able to sustain. While there is extensive literature on these subjects [35], here we restrict our review of text entry in these domains to points that illustrate important concepts with SilentSpeller.

2.1 Silent text entry techniques for users with low dexterity

Edgewrite [64] is an excellent example of modifying a text entry method to address the needs of people with low dexterity (see Wobbrock et al. for more examples [67]). Like Unistroke and Graffiti [5], Edgewrite establishes a simplified and consistent writing method for pen-based text input. However, Edgewrite characters are drawn simply by going from one corner to another in a square. This attribute allows Edgewrite to be adapted for input by a trackball or joystick for users who have tremor or flail hands. However, speeds for the target users tend to be under 20 wpm [64]. Several research projects have explored using the tongue for controlling interfaces [4, 41, 42], which, while slow, could be used for text input. Other research has explored the use of gaze and brain waves as proxies for input. Gaze-based keyboards often involve tracking the eyes as they look at different keys on a visual keyboard and defining a selection gesture such as blinking for "tapping" that key [37, 52]. Similarly, most electroencephalogram (EEG) spellers rely on the electrical signals emanating from the brain as a result of visual stimuli to determine the key being selected [1]. Non-invasive gaze and brain computer interfaces (BCI) rarely exceed 20 wpm [6, 39, 60, 69] and are highly susceptible to body movements.

2.2 Mobile text entry techniques

Palin et al. [43] report mini-QWERTY typing methods (gesture, two-thumbs, completion, autocorrect, etc.) and speeds (average 36wpm) on smartphones given the contributions of 37,000 volunteers. Clawson et al. report mini-QWERTY expert rates of 57 wpm and 95% character accuracy when seated and 53 wpm and 94% when walking [9]. Seated mini-QWERTY typists that could not see the keyboard or the characters they typed averaged 53 wpm and 91% accuracy [8]. Ruan et al. [49] report expert iPhone mini-QWERTY virtual keyboard users can sustain 52 wpm with 95% accuracy while seated, and using English speech for text entry averages 153 wpm with 96% accuracy. These studies use variants of the MacKenzie-Soukoreff text entry phrase set [34] and metrics [35, 65] used in our experiments. Many other mobile text entry systems exist, such as gesture-based keyboards [68, 72], specialty devices [33, 59], or soft keyboards that use alternative sensors in smartphones [17, 18]. However, these methods often require significant learning or focused visuomanual attention that SilentSpeller seeks to avoid.

2.3 Silent Speech

Silent speech interfaces circumvent normal acoustic sensing by measuring other parts of the speech production pipeline, including physical vibrations of the vocal cords, movements of the jaw and tongue, and byproducts of speech production, such as muscle-activation electrical signals or non-audible acoustics [11, 14, 15, 22, 40]. Such interfaces can enable speech signal amplification as well as speech recognition. Measuring the surface electromyography (sEMG) signals created by muscle activation has been used to detect neck, jaw, and cheek movements [27, 36, 38, 53]. Optical and magnetic sensors have been used to track movements of the jaw, face, head and tongue [2, 3, 12, 19]. Ultrasound intraoral images have also been used for silent speech recognition or speech synthesis [10, 20, 23, 24, 30, 58].

While Kapur et al.'s AlterEgo sEMG-based silent speech recognizer [27] was originally tested for subtle stationary input, a recent extension tests three users with movement impairments and dysphonia due to MS [28]. Participants with dysphonia have wildly varying speech patterns such that user-dependent training is often required for any speech-based system. Unfortunately, these

Table 2: Silent Speech Research (expanded from Li et al. [32])

Interface	Modality	Proxy	Dictionary	Accuracy (char/word/phr)	Entry Rate	Subtle Form Factor?	Walking?	Unseen Vocabulary?
Bedri 2015 [2]	Optical & Mag.	Jaw and tongue	11 phrases	P90%	Live, N/A	Google Glass and earpiece	No	Trained only
Kapur 2018 [27]	sEMG	Jaw and cheek	10 words	W92%	Live, N/A	visible electrodes	No	Trained only
Meltzner 2018 [38]	sEMG	Face and neck	65 words	W90%	Offline	visible electrodes	No	Yes
Sun 2018 [56]	Camera	Lip Movement	44 phrases	P95%	38wpm	smartphone	No	Trained only
Fukumoto 2018 [15]	Audio	Ingressive speech	85 phrases	P98%	Live, N/A	smartphone or ring	No	Yes
Li 2019 [32]	Capacitive	Tongue	15 words	W97%	2.2bpm	in mouth, wired	Yes	Trained only
Wang 2019 [62]	RFID	Tongue and face	100 words	W86%	Offline	visible stickers	Potentially	Trained only
Kimura 2019 [30]	Ultrasound	Jaw and tongue	15 words	N/A	Offline	ultrasound probe under jaw	No	Trained only
Gao 2020 [16]	Acoustic	Lip movement	45 words	W91%	Offline	smartphone	No	Trained only
Stone 2020 [55]	EOS	Tongue & lip	30 words	W97%	Offline	in mouth (voiced)	No	Trained only
Zhang 2021 [74]	Acoustic	Lip Movement	90 phrases	P91%	Live, N/A	smartphone	No	Yes
Pandey 2021 [44]	Camera	Lip Movement	105 words	W97%	Off:6.4wpm	smartphone	No	Yes
SilentSpeller Off.	**Capacitive**	**Tongue**	**1164 words**	**W92%**	**Offline**	**in mouth, wireless proto.**	**Yes**	**Yes**
SilentSpeller Live	**Capacitive**	**Tongue**	**321 words**	**C87%**	**37wpm**	**in mouth, wireless proto.**	**Yes**	**Yes**

participants are easily tired giving example data. Collecting 10 repetitions of 15 sentences (660 words total in 30-40 minutes) from each participant, AlterEgo achieved 81% accuracy distinguishing one out of 15 phrases with 5-fold cross-validation. With further development, such a system might allow users to control home automation or send predefined messages to friends or caretakers. Unfortunately, EMG-based systems require precise placement of electrodes, making them difficult to don and doff. These systems are also sensitive to movement artifacts. However, this work highlights that a recognizer that can be quickly trained, even if to distinguish a low number of classes, might be of benefit to these populations. **With SilentSpeller, we seek to create a system that is similarly fast to train, easy to don and doff and wear for extended times, is tolerant of movement artifacts, and enables large vocabulary interaction with high accuracy.**

We are not the first to evaluate the potential of electropalatography (EPG) for silent speech [21]. After preliminary work in 2016, Stone and Birkholz [55] recently demonstrated 97% user dependent accuracy across four users on a 30-word command vocabulary using electro-optical stomatography (EOS), a combination of electrical contact sensors to measure the palato-lingual contact pattern and optical sensors to measure the distance between the tongue and palate and the lip opening and protrusion. User independent recognition averaged 56%. Li et al.'s TongueBoard [32] reported a live user dependent study on a vocabulary consisting of 15 common words, and achieves an average information transfer rate of 3.78 bits per decision (number of choices = 17, accuracy = 97.1%). TongueBoard's vocabulary was focused on the numerical digits and five operators such that the system's phrases were limited to calculator operations, and the system was not appropriate for general text entry. However, the study demonstrated the robustness of EPG to motion artifacts. Inspired by this work, we attempted recognizing the 26 letters of the alphabet with one user and confirmed that accuracies over 90% can be achieved. Encouraged by this initial result, we hypothesized that combining silent spelling with continuous speech recognition methods would result in a text entry system with a large vocabulary that was still fast enough to be useful. In comparison to Stone and Birkholz's study, we evaluate live text entry speed and accuracy, demonstrate robustness to motion artifacts, achieve a vocabulary size of over 1000 words, and evaluate the system's ability to recognize unseen words.

Meltzner et al.'s work with sEMG-based silent speech recognition [38] demonstrated up to 90% user-dependent offline isolated word accuracy on a 65 word dictionary using a hidden Markov model (HMM) based approach. Extensions explored recognizing phrases of silent speech using strict grammars and using a phonetic version of the system to recognize unseen words. SilentSpeller follows a similar development approach but, unlike Meltzner's system, is tolerant of body movement; can be quickly donned (as opposed to careful pasting of electrodes on the neck and jaw); is tested on live text entry (as opposed to completely off-line testing); and can be designed to be contained completely in the mouth. In addition, training data for SilentSpeller was collected in multiple sessions over days or months before the live text entry system was tested, demonstrating that the sensing system is stable and consistent over time.

LipType [44] is a computer vision system that reads the lips of the user. Offline results are reported on 30 MacKenzie phrases containing 105 unique words (compared to SilentSpeller's live text entry with 107 MacKenzie phrases with 321 unique words). LipType is less practical (and not tested) for on-the-go use and is restricted to 6.4 WPM due to computational costs. SoundLip [74] is an offline recognizer for 20 Chinese word commands and 70 sentence commands. We highlight the difference between SilentSpeller and previous work in Table 2. As seen in the table, few systems can generalize to unseen words; they must include examples in their training set. Accuracies are based on the units reported in the literature (characters, words or phrases). When available, text entry rates are included. Not included in the table is our CHI2021 Interactivity demonstration which used pilot results from this research from two participants [29].

3 SILENTSPELLER

We propose silent spelling as a practical alternative to silent speech. In silent spelling, instead of mouthing words, the user spells each letter in the words, one by one. For example, instead of saying "rapidly" (ra·puhd·lee), the user spells each letter "a:r ei pi: di: el wai" (Figure 3). Silent spelling increases the amount of signal available per word for recognition. It is also compositionable in that words that were never seen in training might still be recognized. Subparts of the word, such as three letter (triletter) blocks, might be combined to recognize unseen words. In this manner, large vocabulary

recognition might be possible with relatively little user training. Silent spelling also enables distinguishing between homophonic heterographs such as "I," "aye," and "eye" or "right," "write," and "wright." However, in order to be valuable, silent spelling should provide some advantage over the current dominant mobile silent text entry method, mini-QWERTY virtual touchscreen keyboards.

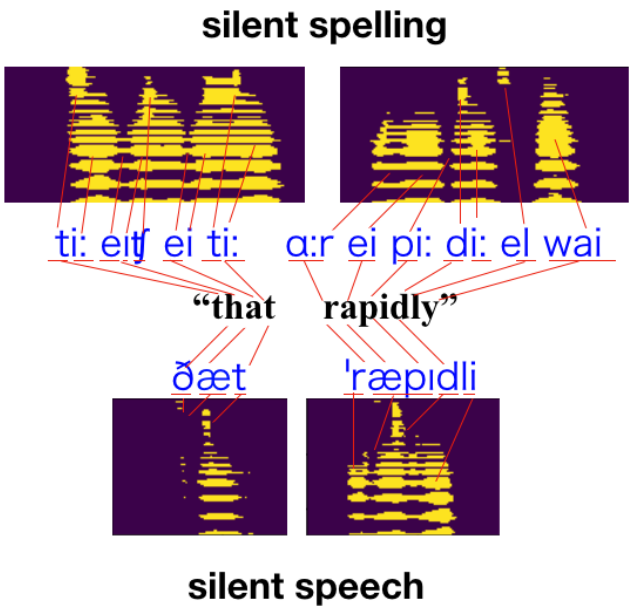

Figure 3: Visualization of samples when spelling and saying "that rapidly." Silent spelling increases the amount of signal available per word for recognition.

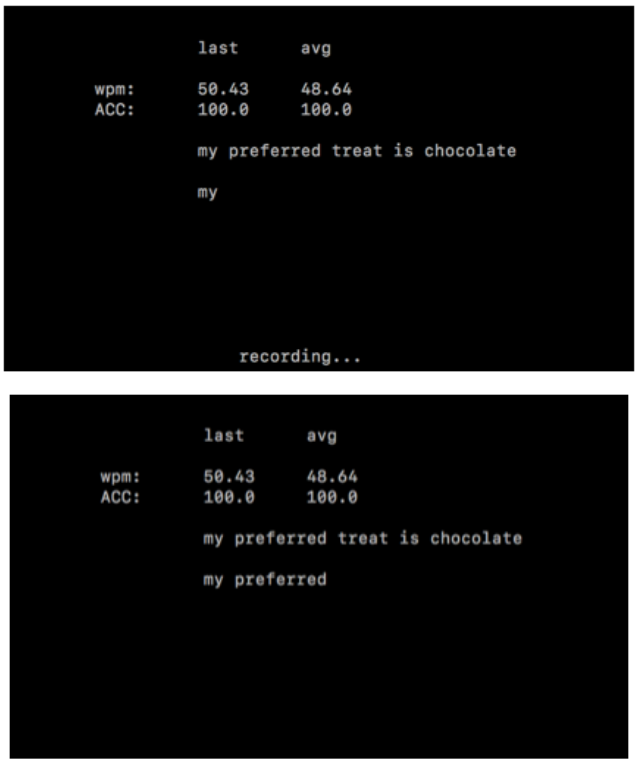

Figure 4: The interface application used in the Wizard of Oz experiment. User transcribes the presented texts and the "recognized" words appear below.

3.1 Pilot Experiment: Silent Spelling versus Smartphone Mini-QWERTY

One concern with silent spelling is speed, because each word is spelled one character at a time. We conducted Wizard of Oz experiments emulating ideal text entry to see if silent spelling provides acceptable speed and ease-of-use before actually building a system. Following previous work [7, 33, 54], we implemented a traditional interface application to test the speed of text entry (Figure 4). The application presents phrases to the participant who then transcribes them over the course of 10 minutes. The user presses the command button on the test computer while silently spelling or speaking (i.e., a push-to-talk interface). Since it is a Wizard of Oz experiment, the system does not actually have a recognition pipeline, and the correct word is always displayed in response to the user's input.

We prepared three conditions for comparison: silent speech input, silent spelling input, and the mini-QWERTY keyboard on the Apple iPhone. Here, when we use the term mini-QWERTY, we are referring to the small virtual touchscreen rendering of the desktop QWERTY keyboard typically used for text input on smartphones. The comparison with mini-QWERTY is important because one of the powerful aspects of our system is its mobile use. Mini-QWERTY, used in a myriad of different implementations, is the dominant form of text entry on mobile devices and is well studied in the literature [9, 13, 50]. We hoped that the results would show that silent spelling, while not as fast as silent speech, could compete with mini-QWERTY. If the speed and ease of use are comparable to mini-QWERTY, then we presume silent spelling may be a viable text entry method, at least for the ideal case (of no additional hardware and perfect recognition), and has the benefit of being hands-free.

For evaluating ease of use, we choose the NASA Task Load Index (TLX) as a metric. As a measure of speed, we use words per minute (WPM) using the formulas presented by Mackenzie, where T is the length of the transcribed text, and S is the time it takes to enter the entire phrase in seconds [35]. Since S is measured from the first keystroke to the last for the phrase, the number of letters is reduced by one. The constant 60 is used as the number of seconds in a minute, 1/5 is adopted because the average length of a word (including spaces) is 5.

$$WPM = \frac{T-1}{S} \cdot 60 \cdot \frac{1}{5} \tag{1}$$

3.1.1 Procedure. Six participants (five male, one female, ages 23-34, and four out of six were iPhone users) were first briefed about the experiment and each item of NASA-TLX. For the experiments using mini-QWERTY, all participants used the same Apple iPhoneX. After each experiment, the participants answered the NASA-TLX questions. After all the experiments were done, an interview was

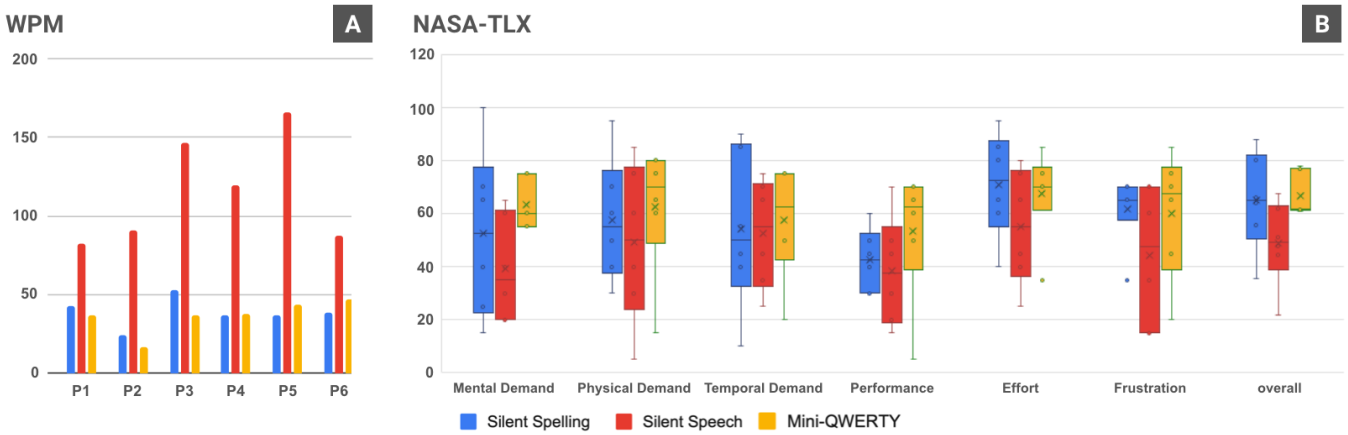

Figure 5: Wizard of Oz study. a) Word per minute rates for each input method. b) The NASA-TLX results averaged across users.

held to ask about the whole experiment. The order of the three conditions followed a balanced Latin square design.

3.1.2 Wizard of Oz Results. Figure 5a shows the words per minute (WPM) for each input method. A one-way within-subjects ANOVA shows a significant difference between the conditions ($F(2, 10) = 32.4$; $p < 0.001$). Post-hoc tests using Bonferroni correction for multiple hypotheses shows a difference between the silent speech and silent spelling conditions ($p = 0.007$; 95% CI [29.3, 124]) and the silent speech and mini-QWERTY conditions (p=0.006; 95% CI [31.2,126]). As we expected, silent speech was by far the fastest, with an average of 115 wpm, while silent spelling and mini-QWERTY were similar, with averages of 38.7 and 36.6, respectively. The 95% confidence interval for the difference in means results in spelling being between 7.54 wpm slower and 11.79 wpm faster than mini-QWERTY.

The NASA-TLX subscale results and the weighted overall workload are shown in Figure 5b. A one-way within subjects ANOVA shows a significant difference between the overall weighted workload scores ($F(2, 10) = 25.9$; $p = 0.022$). Post-hoc tests using Bonferroni correction for multiple hypotheses shows a difference between the silent spelling and the silent speech conditions ($p = 0.002$; 95% CI [8.24,23.32]). The difference in means between spelling and mini-QWERTY did not reach the level of significance for the overall weighted workload. The 95% confidence interval for the difference in means between spelling and mini-QWERTY is between -21.35 and 17.42 (out of a maximum difference of 100).

As expected, the workload for silent speech is lowest, as can be seen in Figure 5b. However, visualizing the results in this manner is misleading as there is large variability between users in how they scale the TLX. Instead, it is more informative to show the mean *differences* in component and weighted overall scores per user between conditions (see Figure 6). While the comparisons are post hoc, examining these graphs can inform future research and potential improvements. As expected, Figure 6a shows that silent speech is favored over silent spelling in almost all categories. Interestingly, the same trend can be seen comparing silent speech versus mini-QWERTY (Figure 6c), which, combined with the increase in text entry speed, suggests that silent speech interfaces may, indeed, find favor with users.

Figure 6b shows the difference between silent spelling and mini-QWERTY. Of note is the effort subscale. In their comments, several participants mentioned that the effort of breaking the word into letters was high, as the word would unconsciously come out of their mouths before they could break it into letters. While, anecdotally, performance seemed to improve during the 10-minute session, perhaps one improvement to reducing the effort of silent spelling is to add recognition for some common words that are easily distinguishable and are routine "slips" while spelling. Another improvement may be to add recognition of proper names tailored for each user. This hybrid approach may point to a method of gradual improvement of the silent spelling system toward silent speech.

It should be noted that all participants are daily users of mini-QWERTY text entry, but they were using silent spelling for the first time. The results suggest potential for silent spelling having an acceptable learning curve and could provide usable text entry speed for the workload compared to current practice. These attributes are promising for user acceptance, which encouraged us to develop a working prototype of SilentSpeller.

3.2 SmartPalate

SmartPalate is a dental retainer-type device with 124 binary capacitive sensors that line the user's palate and capture tongue movements (Figure 2b). Complete Speech originally developed SmartPalate for speech therapy to correct pronunciation. Data is sampled at 100 Hz and sent via a flex circuit ribbon cable to a data module external to the mouth. This module converts the signal to standard USB signals and transmits the data to a personal computer or smartphone via a USB cable. We expect SilentSpeller to be tolerant of body movements [32] as it fits firmly in the top of the mouth. Each user must obtain a dental impression so that the electrode array can be custom fit to each user's mouth (Figure 2a). *Due to Covid-19, the number of participants who could be fitted at this time was limited.*

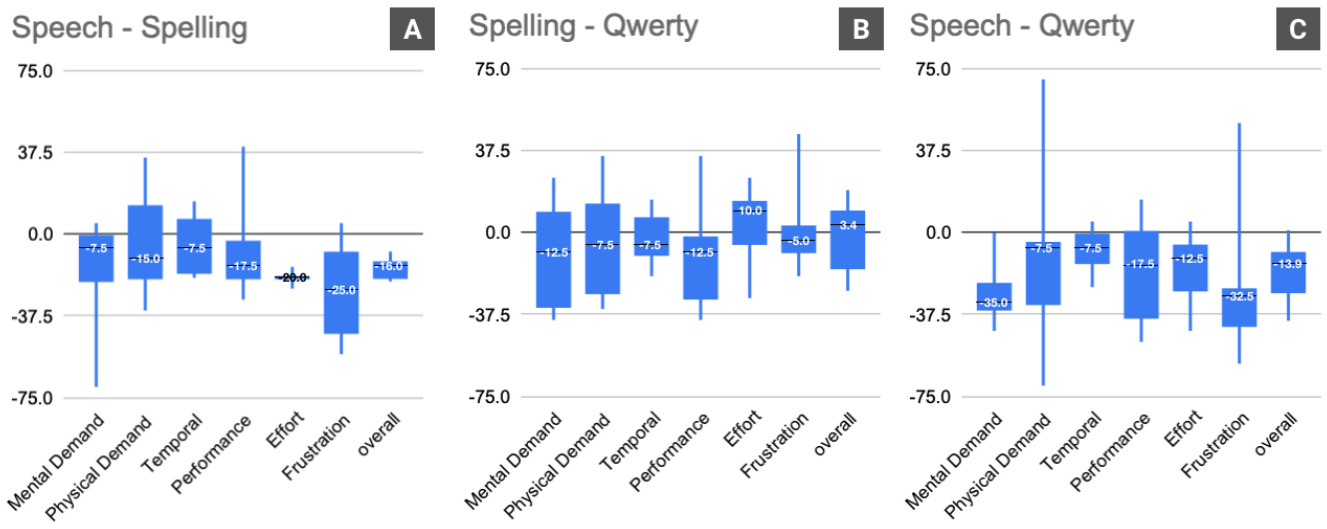

Figure 6: NASA-TLX results showing the difference between a) silent speech versus silent spelling b) silent spelling versus mini-QWERTY and c) silent speech versus mini-QWERTY. Negative values mean that the second input method has a worse rating. White numbers on box plots are median values.

3.3 Recognizer Pipeline

Silent speech and silent spelling share many of the same attributes for recognition. Silent speech often relies on recognizing approximately 44 phonemes in the context of words [11, 27, 38]. Accuracy is given at the word level as silent speech recognizers can leverage co-articulation effects in context to improve rates. Similarly, SilentSpeller focuses on recognizing the 26 letters of the alphabet, silently mouthed in the context of spelling a word. Following the path of early voiced and unvoiced speech recognition development [26, 38], we start with recognition on isolated word dictionaries and gradually increase complexity, transitioning to phrase input and more difficult usage environments.

Training SilentSpeller recognizers follows a consistent training and testing process, detailed here. Principal component analysis is performed on training data sets (which are kept independent from test data). Based on the results of tuning experiments (see the tuning results of Section 4 and Figure 11), we choose the top 16 components ("eigen-palates") for use in our recognition pipeline as the best compromise between accuracy and processing speed. Eigen-palates are never trained with test data.

Figure 7 shows an example of the components extracted for one of P1's tests. Reducing the number of features from 124 to 16 (by 100 times/sec) significantly improves the speed of the real-time recognizer. Another potential use of these components is to reduce the complexity of the hardware. Unused or redundant electrodes can be removed. Alternatively, fewer and larger electrodes that better match the shape of the components may be used. For a wireless system, this reduction in electrodes reduces the amount of data that needs to be transmitted, resulting in a more stable and power efficient system. While the higher order eigen-palate components can be quite complex, the first few show human understandable features. For example, component 1 mostly represents when the mouth is open and the tongue is flattened against the back of the palate as when saying the letter E. Component 2 shows the tongue in front of the mouth as when saying the letter T. Component 3 is representative of the first part of saying the letter J.

When each silently spelled word is collected, each data frame of 124 binary electrode values is projected to the top 16 principal components. Figure 8 shows a visualization of silently spelled data for the English alphabet. In the raw data visualization, plots on the upper area represent activity from the front area of the tongue. For example, in "L" and "T", where the tip of the tongue touches the palate when starting spelling, the area around 0 to 20 is activated.

The resulting 100 Hz 16-dimensional signal is then decoded using hidden Markov models (Figure 9). HMMs are well suited for this task because they have shown high performance in time series pattern recognition in early voiced and unvoiced speech recognition

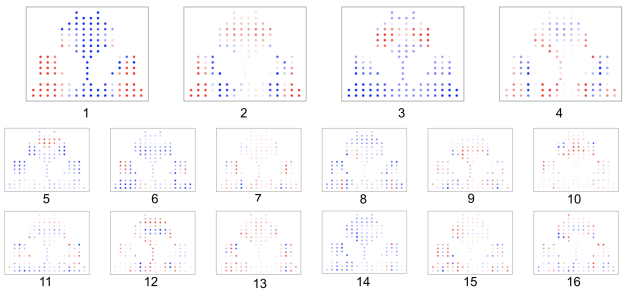

Figure 7: Examples of the top 16 "eigen-palates" (eigenvectors visualized on the SmartPalate's electrodes) extracted using principal component analysis from one fold of P1's 2328 isolated word training data. Dark red are high values; dark blue values are low.

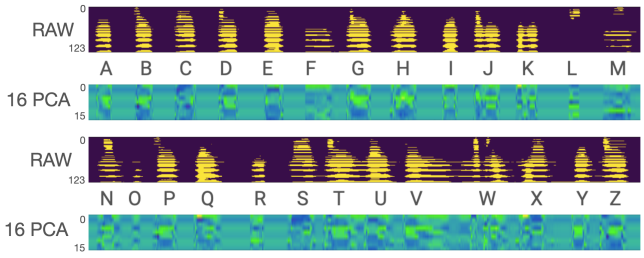

Figure 8: Palatogram of Participant 1 reciting the alphabet. The upper RAW visualizations show the 124 binary electrode values. The lower ones represent projection on the top 16 principal components. Some similar sounding letters, such as B/P and D/T/Z, have similar palatograms. Others, such as J/K, are distinct at some point during production. Longer palatograms for some letters (T/V) are the result of the user not fully opening the mouth after production such that the tongue remains in contact with the palate, which can lead to co-articulation effects.

[38, 46]. In addition, HMMs often require less training data than neural net techniques such as LSTMs and Transformers [51, 61]. Preliminary testing suggests HMMs outperform other methods for this data set. The Georgia Tech Gesture Toolkit (GT2K) [63], a wrapper around the HTK Speech Recognition Toolkit [71], is used for training and testing the HMMs. Training is provided in the form of words, not individual letters, so that co-articulation can be modeled. Initially, the word is artificially segmented into sections equal in number to the letters in the word. Viterbi alignment and Baum-Welch re-estimation refine these boundaries to converge on better boundaries for each letter. First the 26 letters are trained, progressing to triletters, akin to training phones and triphones in conventional speech recognition systems. For triletters with limited occurrence in the training dataset, tied-state triletters are used to reduce error. Based on early experiments, we choose a 12-state, left-to-right HMM topology with no skip transitions. The recognizer pipeline and the dataset are available on GitHub[1].

4 EXPERIMENT 1: ESTABLISHING FEASIBILITY AND TUNING MODELS

To determine the feasibility of an unvoiced spelling input system for mobile, on-the-go, silent, and hands-free text entry, we perform a series of experiments culminating in a live text entry experiment. For practical purposes, we collect words instead of simply the 26 letters of the alphabet. Just as co-articulation affects the pronunciation of phonemes when spoken in a word, letters spoken together affect each other. This effect is especially true when communicating quickly, and, in our experience, users who spell for text entry tend to spell quickly, even eliding (skipping) some letters in longer words. To be effective, SilentSpeller needs to recognize such words, even when users are not being precise.

[1]https://github.com/supernaiter/SilentSpeller

4.1 Text entry corpus and participants

To tune the parameters of the system, we collect 2328 isolated words (each unique word twice) for two participants. P1 and P2 are both male, ages 25 and 50. We use the Mackenzie-Soukoreff phrase set, which consists of 500 phrases, 1164 unique words, and 7048 letters [54]. Each phrase is about 28 characters words long and is designed to be memorable such that participants can read the phrase quickly, potentially memorize it, and enter it as if it was their own thought. While the corpus does not contain any special characters, punctuation, or capitalization, it has become a standard in the literature as it models the short and informal communication that has become commonplace in SMS and social media applications. It is also a reasonable surrogate for the short communication associated with alternative and augmentative communication (AAC) aids.

4.2 Isolated word capture system

We developed a push-to-talk style recording application (Figure 10) to collect samples of silent spelling. The user pushes and holds the command button on the keyboard while spelling each word, releasing the button between words. If the participant makes a mistake, they are required to re-record the word, but no real-time checks are provided. Participants are allowed to take a break when desired. An estimate of speed (wpm) is displayed after every word is recorded. Participants are asked to spell at a rate exceeding 30 wpm to imitate test conditions. The 2328-word data sets required approximately five hours of input for each of the two participants. Data from this experiment and all experiments in this paper can be found at the linked page in the footnote.

4.3 Tuning User Dependent Recognizers

Using the 2328-word data sets from P1 and P2, we use 10-fold cross-validation (i.e., independent training and test sets, random 10% for testing each fold) for each test. To be clear, we are creating user dependent recognizers where only one participant's data is used for training and testing at a time. Table 3 summarizes the top results in our experiments with HMMs and Transformers. The HMM-based recognizers performed exceedingly well, with an average 97% character accuracy and 92% word accuracy. Character accuracies are provided in context and will be higher than word accuracies. In other words, the recognizers attempt to select a word from the dictionary that best matches the silent utterance. While a word could be wrong, most of the letters could be correct (e.g., "cause" instead of "cars"). While deep learning techniques such as Transformers have recently shown much success in language tasks [25, 61], performance here was poor, suggesting that significantly more data would be required to train the neural net models. Given that we are trying to create user-dependent recognizers without requiring an onerous amount of training data, we decided to continue with an HMM-based approach. Future work will investigate data augmentation methods, such as SpecAugment [45], to supplement both approaches.

We wish to optimize HMM parameters on P1 and P2 before testing on P3-P5. We swept over two through 18 states and discovered that 12 states provided good overall accuracy and still worked on the most quickly articulated letters. Figure 11 shows the results of additional parameter tuning. While Figure 11 shows the results

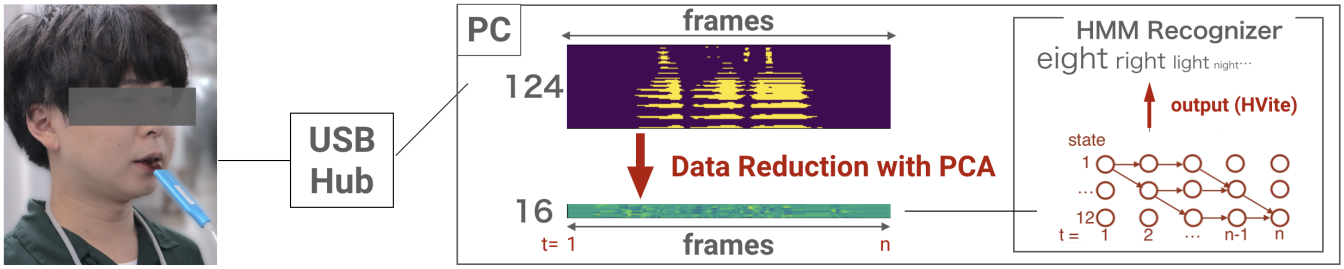

Figure 9: SilentSpeller's recognition pipeline. Captured data from SmartPalate is extracted using principal component analysis. The sequence of extracted features is sent to a hidden Markov model-based recognizer to be decoded into words.

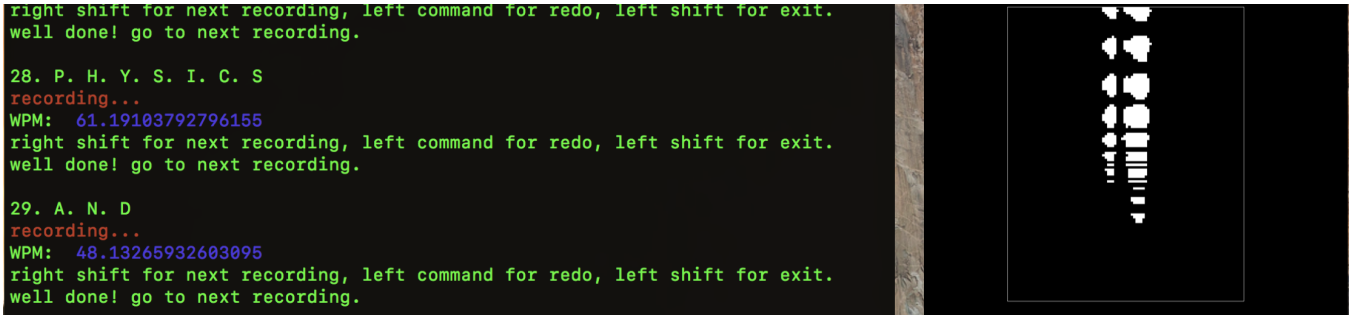

Figure 10: Isolated word recording system. The text interface (left) prompts the user and allows examples to be discarded and redone. The visualizer (right) provides a live trace of which of the 124 electrodes are activated, scrolling right to left.

Participant	1	2
Character (word) accuracy HMM	**97% (93%)**	**97% (91%)**
Character (word) accuracy Transformer	37% (9.1%)	34% (8.8%)

Table 3: Average 10-fold, cross-validation, user-dependent word accuracy on 2328 isolated words, 1164 unique, using HMMs and deep learning Transformers.

obtained for P1, both participants' results followed similar trends. After trying 4, 8, 16, 32, 64, and all 124 principal components, we discovered that 16 components were the fewest that did not adversely affect recognition. Attempting recognition using temporal subsampling at 20 Hz, 25 Hz, 33 Hz, 50 Hz, and 100 Hz showed that 50 Hz was sufficient for this isolated word task. However, testing with the live system below showed that participants could spell quickly enough such that accuracy declined. Thus, we retain the full 100 Hz rate for our system.

Figure 11 (center) shows a very interesting trend. With as little as 500 words (a random 10% are removed from training for each fold of 10-fold cross-validation), the recognizer achieves 90% accuracy and 98% 4-best accuracy. The results for 4-best are especially interesting if we model the SilentSpeller interface on current mobile phone gesture typing keyboards. These systems provide four options for each input. The user can select the top result by simply proceeding to the next word or tap one of three alternatives. By imitating this technique, SilentSpeller will be highly likely to provide a correct word from the 1148-word dictionary with just one hour of training for each participant. When entering phrases, adding a statistical word bigram should further improve the results. For the SilentSpeller use cases of silent text entry while mobile, or for people with movement disorders, one or two hours of training data is quite reasonable, especially since such use cases may often use a limited vocabulary [28, 38].

5 EXPERIMENT 2: GENERALIZATION TO UNSEEN WORDS

Different dictionaries are required for different text entry situations. In early speech recognition systems, a common dictionary across tasks might be trained to establish phonetic models, and then new words are added to the phonetic dictionary to tailor the recognizer to a given task without retraining. Dictionaries can even be swapped as the user changes between tasks.

Adding untrained dictionary words is also useful for recognizing proper nouns. One can imagine a SilentSpeller text entry system importing the top 100 most used entries from a user's contact list and adding them to a personal dictionary without the need for additional training. In this experiment, we test the resilience of the recognizer to words that are added to the dictionary without training examples.

We randomly choose 100 words from the 1164 word dictionary to remove from training. Since every word was spelled twice, we removed both examples, resulting in a training set of 2128 examples. Otherwise, training and recognition were performed as described in Experiment 1. Table 4 summarizes the results.

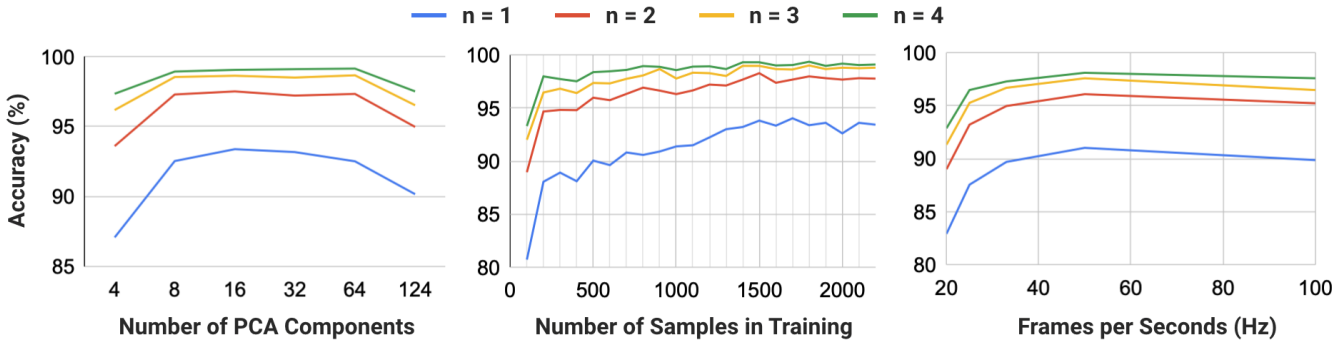

Figure 11: Word accuracy versus number of PCA components (left), training examples (center), and frames per second (right). n=4 means the correct word was returned in the top four most probable words by the recognizer. The number of PCA components were increasingly doubled starting at four and ending with the full 124 dimensions. These figures correspond to P1's data, though P2's graphs are similar.

Participant	P1	P2
Accuracy	93% (84%)	96% (87%)

Table 4: Character (word) accuracy when testing on 100 words (200 examples) removed from the training set.

As expected, there is a drop in performance. However, the system still averages 94.5% character accuracy and 85.5% word accuracy over the 100 unseen words for the two participants. These results are extraordinarily good and suggest that the recognizer can indeed generalize to words for which it has seen no examples.

6 EXPERIMENT 3: TOLERANCE TO ON-THE-GO INPUT

Most silent speech system to date have been limited to seated environments as motion artifacts caused by walking often overwhelm EMG, camera, and ultrasonic sensors. However, we expect SilentSpeller to be as accurate at recognizing silently spelled words when the user is walking as when seated. The experiment below investigates this hypothesis. We have reason to be optimistic: the tongue is relatively isolated from the mechanical shock of walking (otherwise, voiced speech while walking would not be possible) and SilentSpeller's electrode array fits snugly in the mouth such that there is little motion while walking.

6.1 Participants and corpus

P1 and P2 provided a total of 428 phrases (107 phrases for both the walking and seated conditions for each participant). In later experiments, P3–P7 will provide seated phrase data (107 phrases each) for training for the live text entry experiment. The 107 phrases are from the MacKenzie phrase set and consist of 556 words, 321 of which are unique. The most frequently used word "a" appears 24 times. Repetition of such short connector words is fortuitous. We want many examples of the most commonly used words so that the recognizer can be tuned for them.

We used the same capture system that collected the isolated dictionary words. Participants enter the isolated words in the order in which they occur in the phrases, which emulates entry with a live text entry system (but without the ability to see or edit the result). For the walking condition, participants walked continuously indoors in their homes (due to Covid-19) while capturing the 107 phrases. The SmartPalate and its external data recorder were connected to an Apple Macbook laptop that displayed the text entry interface and that was carried by the participant. The seated condition was identical but performed at a desk.

Note that **this phrase data was recorded weeks or months after the initial isolated words for each user.** The good results, seen below, suggest that the sensing system is robust, consistent, and reproducible between sessions.

6.2 Recognizer

As we wish to compare walking versus seated text input, we choose to use the most advantageous training that is reasonable for this study. The recognizer is trained on the two participants' 2328 isolated dictionary words (each) plus their 107 phrases from the condition not being tested. That is, the recognizer for the walking condition was trained with the 2328 dictionary words plus the 556 words from the 107 phrases collected during the seated condition. Similarly, the recognizer for the seated condition was trained with the 2328 dictionary words plus the 556 words from the 107 phrases collected during the walking condition. No training data is used in any test set. During recognition, the system uses a dictionary constructed from the 321 unique words from the 107 phrases. A bigram is constructed using the 107 phrases and Laplace smoothing (so that any word combination is possible). Results with a unigram and trigram are included for comparison.

6.3 Results and Discussion

Table 5 presents the results of the study. There is little difference in the accuracy between the seated and walking conditions, demonstrating the robustness of SilentSpeller to body motions. Note that there is less difference between the character accuracy and the word accuracy with this experiment than with the dictionary words. This difference can be explained by the higher prevalence of shorter words in phrases than in the dictionary. With higher representation

participant condition	unigram perplexity=226	**bigram perplexity=10.5**	trigram perplexity=3.25
1-seated	94% (89%)	**99% (98%)**	99% (99%)
1-walking	94% (88%)	**99% (97%)**	97% (99%)
2-seated	88% (79%)	**94% (93%)**	95% (93%)
2-walking	87% (76%)	**96% (93%)**	97% (95%)

Table 5: Comparing walking to seated text input. Character and (word) accuracy are shown using different n-grams.

of words such as "of," "an," "my," "a," etc., the average word length is shorter, and the rates are more similar.

In certain situations, such augmented control and communication for people with both low dexterity and severe dysphonia [28], a limited set of phrases might be used (resulting in very low perplexity and easier recognition). To determine if SilentSpeller might be suitable for such a situation, **we test the recognizer with a strict grammar that matches the input to one of the 500 MacKenzie phrases. Accuracy increases to 100% for all four conditions.** While silently spelling a phrase is slower than silently speaking it, SilentSpeller's increased reliability might be preferred in many scenarios.

Table 5 also compares results using a unigram, bigram, and trigram and shows the perplexity for each grammar, given the dictionary of 321 words and the 107 phrases. While strong grammars can be very useful in limited cases, as in the phrase selection experiment just described, one must choose a grammar that is appropriate for the task. Here we focus on a bigram, as it allows composition of many phrases while requiring less training data than unconstrained situations. In this case, the trigram severely limits composition and provides little improvement on the results.

7 LIVE TEXT ENTRY USING SILENTSPELLER VERSUS MINI-QWERTY

We test text entry using a live, interactive version of the SilentSpeller recognizer. We could find no comparable silent speech system in the literature to test against SilentSpeller for the mobile text entry task as they were only run offline [38, 44, 62], have too small a vocabulary [2, 27, 30, 32, 57, 74], are constructed for command phases as opposed to text entry [2, 15, 57], or some combination thereof. Instead, for reference we again resort to comparing to the most common form of English mobile text entry: two-thumb typing on virtual mini-QWERTY keyboards [43]. Comparing SilentSpeller's speed and accuracy to a commonly available reference system like mini-QWERTY helps establish whether SilentSpeller is viable and invites comparison by future silent speech text entry systems.

7.1 Participants and corpus

A total of seven participants participated in the experiment, including the two participants who had participated in the previous sections. Covid-19 restrictions for elective dental procedures limited recruitment efforts. Since the system is user dependent, ideally all users would collect the 2328 examples and 107 phrases describe previously. However, this process would have taken an impractical amount of time for volunteers. Therefore, we asked the participants to collect the 556 words contained in the 107 phrases and 500 additional randomly selected words from the dictionary. Participants required about two hours to collect this data, and, given the accuracy versus number of training examples curve in Experiment 1, we expected this amount of training to be sufficient. For the two participants who provided 2328 examples, 500 words were chosen at random and only the seated dataset from Experiment 3 was used. The live text entry experiment, which included six 20-minute sessions in total, took two hours. Thus, participation required a total of four hours.

P3-P7 are male, ranging from ages 23 to 45. P2 and P5 are native English speakers. Due to Covid-19 circumstances, all experiments were conducted in participants' respective homes using Apple MacBook Pro laptops and SmartPalates. Although P3 and P4 are non-native speakers, they have been living in primarily English-speaking countries for more than five years and have acquired advanced English oral skills. P1, P6 and P7 are non-native speakers, and have not stayed in English native countries for more than 10 months. Even so, these participants have a basic understanding of English (approximately 70 to 90 in TOEFL iBT score, though not all of them have taken the test).

For testing, the participants attempted to input the same 107 phrases again, as quickly and as accurately as possible while seated. We augment the HMM recognizer with the same bigram as described above.

7.2 Text Entry using SilentSpeller

Based on standard text entry practices established by previous work [7, 33, 54], we implemented an interface application to test the speed and accuracy of text entry using SilentSpeller. The application presents phrases from the MacKenzie-Soukoreff phrase set to the participant who then transcribes them over the course of 20 minutes. The SilentSpeller app mimics the user experience from the gesture keyboard [72] included on most smartphones. Interactions include INPUT (silent spelling), N-BEST-SELECT/TAP (produced by touching the front of the palate for more than 0.3 seconds and less than 1 second), and ERASE-WORD/STICK (pressing the tongue firmly on the entire palate between 0.3-1.0 seconds). While transcribing each phrase, the user presses a push-to-record button and inputs each word by silently spelling with the SmartPalate (Figure 12b). Upon button release, the captured data frames are recognized. In about a second, the interface displays a list of the five best word predictions in order of probability (Figure 12c). If the next input is started, the first candidate is assumed correct. If, instead, the correct answer is in the 5-best list, the user selects the best candidate with the TAP gesture (Figure 12d). If no correct answer is shown, the candidates are deleted with the STICK gesture; the system returns to the input state so the user can start over with that word. When the user has completed a phrase, the user presses the right shift to move to the next phrase.

STICK and TAP are distinguished from swallowing using simple thresholds. Gestures are triggered when 20 or more electrodes are activated. If the gesture takes longer than a second, the system recognizes and ignores a swallow. If less than 80 electrodes are activated by the gesture, the TAP gesture is recognized. If more than 80 electrodes are activated, the STICK gesture is recognized.

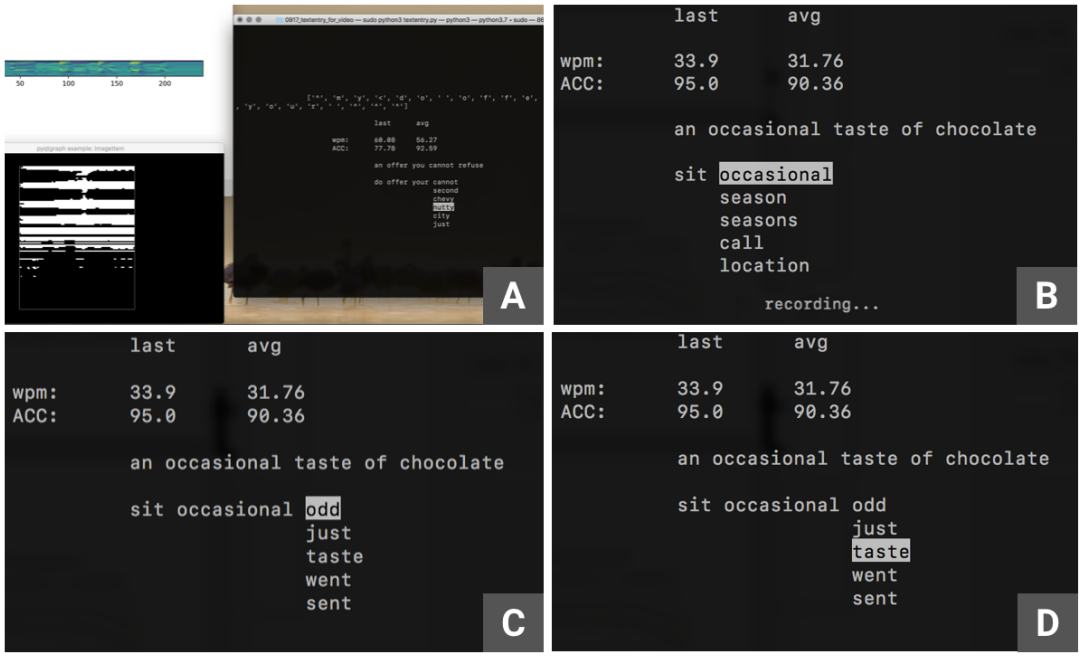

Figure 12: SilentSpeller live text entry. a) Three screens are displayed: a real-time palatogram, a palatogram of the latest recorded sample, and text prompts and results. b) User starts recording the next word by pushing the command button. c) The interface shows the user a list of the five best word predictions in order of probability. d) User can choose the words from the list using TAP gesture.

Activation means that the electrode was activated at one of the time frames during the gesture. These gestures can only be recognized when the user is not pressing the push-to-talk command button and cannot be confused with silent spelling.

7.3 Text Entry with mini-QWERTY

We ran the same text entry experiment using virtual QWERTY keyboards on smartphones for comparison. Participants used their personal smartphone for this experiment (some used iPhone, others used various Android models) in portrait mode using two thumbs for entry. Participants were seated and entered one character at a time without word prediction but with auto-correction to retroactively fix spelling after the user has entered a word. To be clear, the user typed one letter at a type (no gesture typing) and was **not** allow to chose from a list of suggested words to avoid typing the rest of the word (e.g. typing "mis" and selecting "mission" from a autocompletion list). However, if the user mistyped a word, the autocorrect on the keyboard had a good chance of correcting it (e.g., "misson" would be corrected as "mission"). This method of mini-QWERTY text entry was chosen as it is the fastest method reported (43wpm vs. 36wpm across all methods) in a survey of 37,000 smartphone users [43].

7.4 Metrics

We calculate words per minute (WPM) using the formulas presented by Mackenzie, as stated previously [35]. In this experiment, the time required for recognition (i.e., the time from the completion of the input to the return of the recognition result from the server), was removed from the overall time. The goal of this experiment is to measure the speed of silent spelling for text input, not the speed of a recognizer. This compromise probably underestimates the speed of silent spelling as the second or two delay in the recognizer interrupts the flow of the user. With focused development we expect the recognizer latency to decrease to under 20 ms and not be noticeable by the user.

Total Error Rate (TER) is a metric proposed by Soukoreff et al. [54] to measure text entry error. It is an alternative to using minimum string distance that considers all keypresses in the input stream, such as backspaces, as well as the final transcript. The input stream is divided into the following four classes:

- Correct (C) keystrokes – alphanumeric keystrokes that are not errors.
- Incorrect and Not Fixed (INF) keystrokes – errors that go unnoticed and appear in the transcribed text.
- Incorrect but Fixed (IF) keystrokes – erroneous keystrokes in the input stream that are later corrected.
- Fixes (F) – the keystrokes that perform the corrections (in this case, erase word and next n-best-candidate).

Using these classes, the TER is calculated by the following equation.

$$\text{TotalErrorRate\%} = \frac{INF + IF}{C + INF + IF} \cdot 100 \tag{2}$$

Since SilentSpeller recognizes letters in the context of whole words instead of letter-by-letter, we must adapt the TER. For example, suppose SilentSpeller recognizes a word incorrectly, suggesting a word that consists of five letters. The user triggers the ERASE-WORD gesture, erasing the suggested word. The gesture would count as a single fix (F) ("<"), and the five incorrect letters are counted as five incorrect but fixed (IF) keystrokes. In contrast, in the case of an N-BEST gesture (represented by "*" below), the gesture counts as a single fix (F) with no IF keystrokes. One can argue that SilentSpeller is a gesture recognition system that injects characters into the text entry stream only after the user confirms the recognition of the current word by continuing to the next word, selecting an alternative from the n-best list, or erasing the word and re-entering. In such a case, counting IF keystrokes after a ERASE-WORD or N-BEST gesture unfairly penalizes the system. An alternative argument is that SilentSpeller should inject the letters for each word edited with ERASE-WORD and N-BEST gestures and inject corresponding backspaces when an alternative word is selected or a word is erased (see Zhang and Wobbrock [73] for a discussion). The method outlined here seems a reasonable compromise.

We illustrate how the TER is calculated in our system using an example sentence "that is very unfortunate." We suppose the user entered "that" and "is" correctly but failed to enter "very" ("berry" was entered instead) and also did not get a good candidate in the n-best list. The user deletes the word. By redoing the input, the correct word appeared as the second candidate in the n-best list, so he did a N-BEST gesture to choose that candidate. The resulting input stream is

> Presented Text : [that is very unfortunate]
> Input Stream : [that is berry<*very fortunate] (29 in total)

These are classified as:

> C : [that is very fortunate](22)
> IF : [berry](5)
> F : [<*](2)
> INF : [](0)

The resulting TER is 18.5%. Similarly, when the sentence presented is "time to go shopping" and input stream is "time to her shopping", the calculation is

> Presented Text : [time to go shopping]
> Input Stream : [time to her shopping] (20 in total)
> C : [time to shopping] (17)
> IF : [](0)
> F : [](0)
> INF : [her] (3)

The resulting TER is 15.0%.

7.5 Results

Table 6 shows the results of the live text entry experiment. SilentSpeller's average session speed was 37 wpm. Average text entry accuracy (1 - TER) was 87%. Unlike the previous offline experiments, this accuracy metric considers failures of the user to type correctly, recognizer failures, and corrections. Participants mostly chose speed over accuracy, often leaving characters uncorrected, especially for P3, who chose not to edit at all. The average mini-QWERTY speed was 48wpm, well above the reported 36wpm average measured across 37,000 users in the literature [43], which might be expected for students recruited at technical universities. Average accuracy was 93%.

While SilentSpeller's speeds were slightly above the reported 36wpm mini-QWERTY average [43], for this group of users mini-QWERTY was faster. Certainly these technically savvy participants had much more experience with mini-QWERTY than SilentSpeller, which suggests that speeds and accuracies might improve with more practice. Participants quickly adapted to silently spelling words for text entry, though some participants did remark on improving with practice with the interface. P3 consistently improved his results each session. P2 discovered that his recognizer was good enough that he rarely waited to see the result of the output before continuing to the next word. This strategy resulted in a maximum 53 wpm speed while still maintaining 91% accuracy. When asked about his experience, P2 reported a sense of "flow" when the recognizer was working well which allowed him to keep a rhythm to the text input. This success suggests improving recognizer accuracy may cause the other participants to reach similar speeds.

At the end of the experiment, P1 and P2 attempted another informal 20 minute live text entry SilentSpeller session while walking. They achieved similar results to their seated performance, confirming the live system's tolerance to on-the-go usage, as is expected given the off-line experiment.

Note that the main live text entry experiment occurred about a week after the phrase training data was collected and weeks to months after the initial isolated words were collected. Again, the results suggest that the SmartPalate provides robust and reproducible results, even over significant time gaps between sessions.

When optimizing the system for recognizer speed, the average word recognition time was 200ms whereas spelling a word requires around 1 second. This result suggests that if we structured the system to execute the Viterbi recognition algorithm synchronously with the incoming data (as opposed to batching the data for each word), the user would perceive little to no delay in recognition.

7.6 Qualitative and Subjective Comments

P4 mentioned that SilentSpeller felt "magical" and was surprised at how well it worked. They believed it could be a "game changer" in sports or construction. While the dental retainer is considered a considerable drawback by most participants, P4 would consider using it while running or cycling. Besides that the device caused excess salivation, P4 considered it surprisingly comfortable during the long training sessions. P7, in contrast, mentioned having to keep their mouth open to avoid false triggering the system, which was tiring. P5 really liked the idea of discreet input but found the task of spelling to be too cognitively demanding. P5 also had the lowest recognition rates and said that caused him to "overthink" the interface. Several other participants mention the mental demand of spelling but like the ability to do hands-free text input. Perhaps one way to address this mental demand is to progressively include common spoken (not spelled) words in the base recognizer that are easily distinguishable to make a hybrid silent spelling/speech system. P2, who used the system while walking, remarked that he

		Speed (WPM)			Accuracy (%)		
	Method/Session	1	2	3	1	2	3
	mini-QWERTY (iPhone X)	35	36	38	95.7	95.9	96.7
P1	**SilentSpeller**	**42**	**41**	**43**	**95.5**	**93.4**	**93.8**
	mini-QWERTY (Pixel 4 XL)	35	39	43	92.8	91.0	89.7
P2	**SilentSpeller**	**46**	**53**	**52**	**91.2**	**89.8**	**91.2**
	mini-QWERTY (iPhone 11 pro max)	27	30	31	97.8	95.1	95.4
P3	**SilentSpeller**	**30**	**36**	**41**	**91.7**	**89.3**	**91.9**
	mini-QWERTY (Pixel 4)	43	45	49	90.5	91.7	93.6
P4	**SilentSpeller**	**30**	**38**	**34**	**83.8**	**87.8**	**86.4**
	mini-QWERTY (Pixel 4a)	80	80	92	93.7	91.4	91.8
P5	**SilentSpeller**	**38**	**33**	**37**	**78.7**	**82.1**	**74.6**
	mini-QWERTY (Oppo reno3 A)	50	44	45	85.8	86.5	93.4
P6	**SilentSpeller**	**31**	**28**	**26**	**82.5**	**92.3**	**90.2**
	mini-QWERTY (iPhone 11)	56	57	59	92.1	91.2	91.7
P7	**SilentSpeller**	**30**	**29**	**32**	**82.6**	**77.6**	**83.8**

Table 6: Live text entry results on SilentSpeller and mini-QWERTY keyboard. Words per minute (left) and accuracy (right) for each of the seven participants' three sessions. Accuracy is defined as 1 - Total Error Rate.

timed his silent speech input to the pace of his footsteps, which seemed to improve his speed and consistency while silent spelling.

8 DISCUSSION AND FUTURE WORK

In general, SilentSpeller achieves viable text entry rates for novices with the method, especially when compared to these users' expert virtual mini-QWERTY smartphone rates. In addition, SilentSpeller has the advantage that it could be used without the need to encumber the hands.

In retrospect, more training data might have been wise to reduce live text entry error rates. However, we wished to observe performance across users when a reasonable amount of training (about 1-2 hours) was collected to test the potential practicality of the method. Certainly, the system performed admirably for P1-P3 and P6, and P4 and P7 achieved text entry results that might be reasonable for informal SMS-like communication between colleagues. It would be interesting to re-run the experiment using all training data that was available for P1 and P2 and collecting the additional five or six hours of training from P3-P7 to achieve parity. By optimizing the recognizer across all data, perhaps most participants could achieve their mini-QWERTY speeds.

English skill and SilentSpeller's results do not seem to be correlated. P1, a non-native speaker who had not lived in an English-speaking country for more than 10 months, scored high on both accuracy and speed. However, the high scores may be better explained by the fact that P1 and P2 are skilled users who have continuously participated in the project and provided a large dataset of 2328 samples. P6, who has never been to an English-speaking country, achieved an accuracy of over 90% in the two sessions. On the other hand, P5 is a native speaker and an outstanding mini-QWERTY typer, but he had the lowest average accuracy in SilentSpeller.

An examination of the participant with the poorest accuracy (P5) reveals a concerning pattern. Letters pronounced with an "EE" sound (B, C, D, E, E, P, T, and V) are often confused. This result is understandable as the electrodes on the SmartPalate cannot sense the position of the lips, which are used to produce the sounds associated with these letters. This trend can be found in all the participants' data, suggesting that the triletter context modeling and dictionaries are needed to help differentiate words with these letters. The solution may be simple: add additional electrodes in front of the teeth. We are working with Complete Speech, the makers of Smart Palate, to create such a system.

The question remains as to why P5's result is so poor compared to the other participants. This result is especially curious as P2 and P5 are the only native English speakers in the experiment and have vastly different accuracies. P5 might differ from the other participants in that perhaps the SmartPalate did not fit as well, the mouth shape might differ in some way, or that the electrodes are miscalibrated in a subtle way that is not apparent upon inspection of the data using the visualizer. Recruiting more participants and comparing their results will help solve this puzzle.

Figure 13 demonstrates that early testing with a potential user before they give a full training set of data can predict future good or poor performance. P2 and P5 achieved average accuracies of 91% and 78% in live text entry experiments, respectively, after full training. However, early testing results with as little as 200 examples (about half an hour's worth of training data) would have allowed the researchers to predict whether or not continuing to collect data would have led to a good experience with SilentSpeller for the potential user.

Compared to EMG, ultrasound, and camera-based silent speech systems, SilentSpeller can be donned and doffed more easily (the equivalent of putting in a dental retainer) and can be used while on-the-go. With continued development, SilentSpeller hardware could be hidden from the view of casual spectators in an in-mouth retainer.

SilentSpeller may fit a niche for text entry for those who are on-the-go and need to communicate silently and hands-free. Alternatively, SilentSpeller might find use with those with manual

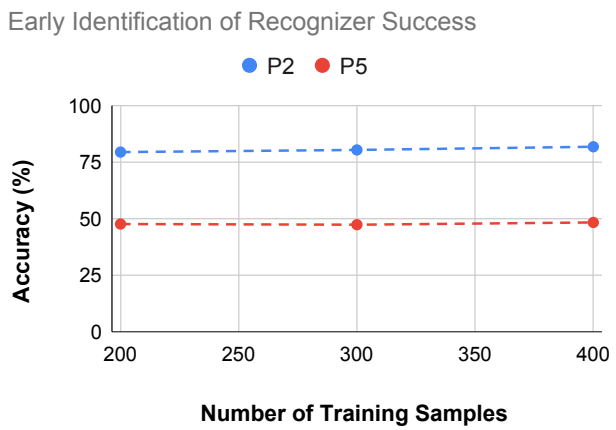

Figure 13: Leave one word out testing on small training sets (N=200, 300, and 400) can predict SilentSpeller's future success for a given user.

dexterity issues who need to be quiet while typing due to being in a public place. Most means of text entry for those with limited manual dexterity [39, 60, 68] report slower input rates and slower learning curves than observed here, suggesting a future direction of research directly comparing against current methods for participants with multiple sclerosis, Parkinson's, cerebral palsy, and muscular dystrophy. For people with both severe dysphonia and low dexterity, a more conservative version of SilentSpeller might be used where the user spells phrases for controlling home automation or choosing one of N phrases for communication.

While most of the literature focuses on silent speech, we decided to investigate silent spelling and were surprised at the viability of the idea. Using spelling for word-by-word text entry intuitively seemed too slow, but Wizard of Oz testing showed that reasonable text entry rates are possible. This observation suggests a re-examination of the silent speech literature. For example, EchoWhisper [16] reports 92% accuracy on 45 silently spoken words using ultrasonic backscatter from standard commodity smartphones. Perhaps by limiting the classes to 26 letters and using the SilentSpeller framework described here, a full text entry system could be created. Such a system might prove more valuable for the general public, allowing unobtrusive text entry on crowded public transportation, for example, without needing to learn the QWERTY keyboard (or perhaps keyboards for other languages).

8.1 Limitations

While this paper shows the potential of SilentSpeller, there are obvious limitations with respect to the number of participants, amount of training data, number of text entry sessions, hardware sensing, and the recognition pipeline. Here we review some options for future work.

8.1.1 Hardware improvements. The current SmartPalate retainer sends its sensing data through a USB cable, requiring the user to be tethered to a PC or smartphone, which limits the system's mobility. The current system can be made wearable for testing; Figure 1a shows such a system constructed using a Vufine head worn display, the Smart Palate, and the support hardware in a backpack. However, to make SilentSpeller more portable, as shown in Figure 14, we developed a compact Bluetooth Low Energy dongle that retrofits the SmartPalate retainer to enable it to send its data wirelessly. Nevertheless, the SmartPalate is still obtrusive in many social settings. Though the SmartPalate's sensing probes are completely in-mouth, they are connected to the capacitive sensing circuitry hanging outside the mouth through a flexible PCB ribbon. Our latest prototype (Figure 15) suggests it is possible to enclose the entire system, including sensing, processing, communicating and powering in medical grade silicone that resides completely in mouth as with Lee et al. [31]. While our current in-mouth prototype uses Bluetooth Low Energy for communication, backscattering might be used to significantly increase battery life. Sensing can be improved by providing analog instead of binary values. Lip-facing electrodes could be added to sense lip motions, and optical proximity sensors, similar to those in the parallel work by Stone and Birkholz [55] can be added on top of the mouth near the teeth to detect opening of the mouth. These extra sensors should improve recognition of letters that currently have similar palatograms (e.g., B/P and D/T/Z).

Each mouthpiece of the system is custom-made based on each user's dental impression to ensure a stable fit and consistent positioning of the tongue among different users. An alternative form-factor for the device is the mouth guard designed for treating bruxism (i.e., teeth grinding), where users close their mouth to hold the guard with their teeth. Though this variant might constrains a user's mouth movements, it fits the application of silent speech and subtle interfaces where minimal visible movement is preferred.

Electrical stimulation through the current electrodes could provide the user with feedback from the system [47, 48]. Such tactile stimulation could help users with feedback while using edit gestures. Alternatively, such an electrode array might allow for two-way communication between two users. Perhaps some words or gestures entered on one SmartPalate could be identified by another user through electrical stimulation playback on a second SmartPalate.

8.1.2 Recognizer improvements. Besides adding sensors to the SmartPalate to help recognize letters distinguished by lip movement, additional accuracy improvements might be obtained from using linear discriminant analysis (LDA) or independent component analysis (ICA) instead of PCA. Initial experiments on the data from P1 and P2 for the first experiment show potential improvements using a variation of segmentally-boosted hidden Markov models (SBHMMs) [70]. P1's error reduced 38% (2.70% to 1.67% character error and 7.17% to 4.85% word error), but P2's error increased 52% (3.43% to 5.22% and 9.11% to 13.25% word error). While initial work with deep learning Transformers [25, 61] proved disappointing (66% character error rate), additional data or data augmentation may eventually lead to improvements. As more participants are enrolled, we can experiment with user-independent and user-adaptive models. Other improvements include larger vocabularies, more sophisticated grammar models, and the ability to recognize out-of-vocabulary words.

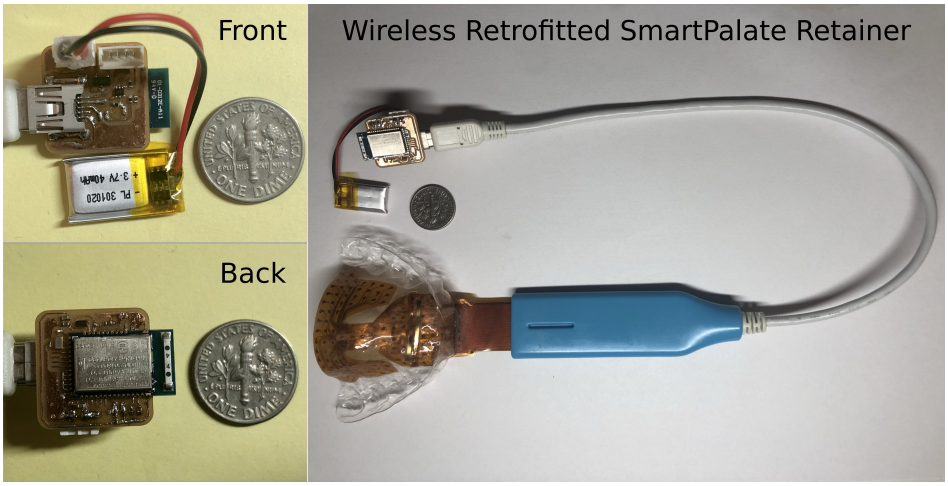

Figure 14: A compact Bluetooth Low Energy (nRF52832) dongle that enables the Smartpalate retainer to communicate wirelessly

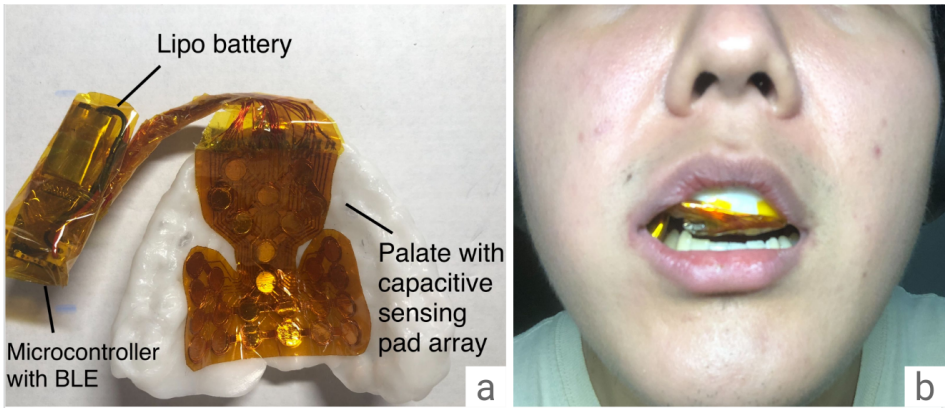

Figure 15: Wireless SilentSpeller prototype a) A portable wireless prototype for SilentSpeller that uses a Bluetooth low energy (nRF52832) micro-controller for communication and capacitive sensing b) A user wears the prototype by putting the micro-controller and battery in the cheek

Short, infrequent words were the most difficult to recognize with SilentSpeller, while longer words would often be recognized correctly even if the user misspelled them. This observation suggests we should optimize the number of examples of each word given as training by each user. In a pilot experiment, after identifying 12 troublesome short words and having our participants provide six more examples of each, those words became much more reliable.

Text entry speeds might be increased by improving the recognition of the ERASE-WORD and N-BEST commands. Increasing the speed of the recognizer would allow more fluid input by the user and perhaps lead to faster texting rates. If the interface were changed to be more akin to a transcribing speech recognizer, the user could spell multiple words or the entire phrase at a time and only correct when necessary. With such an interface change, the recognizer could use more language context when recognizing letters at the phrase level, which would lead to less errors.

8.1.3 User independent recognition. Since SilentSpeller already requires the creation of a custom fit retainer, one would expect the extra inconvenience of providing training data would not be too much of an additional barrier for interested potential users. However, starting with a user independent recognizer would allow immediate use and lead to faster improvements in accuracy for a user adaptive system. In an initial exploration, we performed leave one user out cross validation on the first 500 words (randomly) collected from the first five participants. The recognizer dictionary was limited to the 500 words in each test set. Note that some words in the dictionary would not have been trained in the training set due to the randomness of collection. Results averaged 55% character accuracy and 36% word accuracy.

8.1.4 Study Improvements. Punctuation, capitalization, and emojis are not included in this experiment as we are first attempting to explore whether this technique of text entry is feasible and be able to have a baseline comparison to other techniques. This area

can be the focus of significant future work (e.g., how should one spell characters with multiple pronunciations such as "&", "!", or even "0"?). However, many experiments in the text entry literature [7–9, 35, 66, 68, 69, 73] use a similar corpus to the one used here, arguing that there are compelling situations where the 26 letters and space are sufficient for communication and automation control.

Currently we are using a modification to the total error rate to address how SilentSpeller is used. Recently, Zhang and Wobbrock have developed new metrics to address situations with auto-correction and word prediction [73]. Future studies will adopt and adapt these metrics for SilentSpeller testing. In some senses, comparing SilentSpeller to smartphone QWERTY text entry is unfair without auto-correction using the same limited vocabulary and bigram. However, pilot testing adapting Zhang and Wobbrock's open source text entry testing software with these language modeling advantages shows a similarity to the results here. Further experimentation is needed.

9 CONCLUSION

We introduce SilentSpeller, an interface for text entry using unvoiced spelling of words. We evaluate SilentSpeller's recognition system on a dictionary of 1164 isolated words resulting in average 97% character accuracy. In another test, text entry speeds and accuracies were relatively unaffected by the user walking during input. Live text entry experiments with seven participants demonstrated texting rates competitive with smartphone virtual-QWERTY input rates but without necessarily encumbering the hands. These results suggest that SilentSpeller can be an efficient text entry system and may find niche applications for on-the-go, loud environment, hands-free text entry or silent text entry for people with movement impairments. Further work will explore specific application domains, tune recognition accuracy by adding sensors for the lips, and determine whether user-independent recognition and user adaptive recognition may be possible.

10 ACKNOWLEDGMENTS

Thanks to Dhruva Bansal for the initial recognition experiments using Transformers.

REFERENCES

[1] S. T. Ahi, H. Kambara, and Y. Koike. 2011. A Dictionary-Driven P300 Speller With a Modified Interface. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 19, 1 (2011), 6–14.

[2] Abdelkareem Bedri, Himanshu Sahni, Pavleen Thukral, Thad Starner, David Byrd, Peter Presti, Gabriel Reyes, Maysam Ghovanloo, and Zehua Guo. 2015. Toward silent-speech control of consumer wearables. *Computer* 48, 10 (2015), 54–62.

[3] Florent Bocquelet, Thomas Hueber, Laurent Girin, Christophe Savariaux, and Blaise Yvert. 2016. Real-Time Control of an Articulatory-Based Speech Synthesizer for Brain Computer Interfaces. *PLOS Computational Biology* 12, 11 (11 2016), 1–28. https://doi.org/10.1371/journal.pcbi.1005119

[4] Héctor A Caltenco, Björn Breidegard, and Lotte NS Andreasen Struijk. 2014. On the tip of the tongue: Learning typing and pointing with an intra-oral computer interface. *Disability and Rehabilitation: Assistive Technology* 9, 4 (2014), 307–317.

[5] Steven J Castellucci and I Scott MacKenzie. 2008. Graffiti vs. unistrokes: an empirical comparison. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 305–308.

[6] Xiaogang Chen, Yijun Wang, Masaki Nakanishi, Xiaorong Gao, Tzyy-Ping Jung, and Shangkai Gao. 2015. High-speed spelling with a noninvasive brain–computer interface. *Proceedings of the national academy of sciences* 112, 44 (2015), E6058–E6067.

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

[8] J. Clawson, K. Lyons, T. Starner, and E. Clarkson. 2005. The impacts of limited visual feedback on mobile text entry for the Twiddler and mini-QWERTY keyboards. In *Ninth IEEE International Symposium on Wearable Computers (ISWC'05)*. 170–177.

[9] James Clawson, Thad Starner, Daniel Kohlsdorf, David P. Quigley, and Scott Gilliland. 2014. Texting While Walking: An Evaluation of Mini-Qwerty Text Input While on-the-Go. In *Proceedings of the 16th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Toronto, ON, Canada) *(MobileHCI '14)*. Association for Computing Machinery, New York, NY, USA, 339–348. https://doi.org/10.1145/2628363.2628408

[10] Tamás Gábor Csapó, Tamás Grósz, Gábor Gosztolya, László Tóth, and Alexandra Markó. 2017. DNN-Based Ultrasound-to-Speech Conversion for a Silent Speech Interface. In *INTERSPEECH*.

[11] Bruce Denby, Thomas Schultz, Kiyoshi Honda, Thomas Hueber, Jim M Gilbert, and Jonathan S Brumberg. 2010. Silent speech interfaces. *Speech Communication* 52, 4 (2010), 270–287.

[12] 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 (2008), 419 – 425. https://doi.org/10.1016/j.medengphy.2007.05.003

[13] Torsten Felzer, I Scott MacKenzie, and Stephan Rinderknecht. 2014. Applying small-keyboard computer control to the real world. In *International Conference on Computers for Handicapped Persons*. Springer, 180–187.

[14] João Freitas, António Teixeira, Miguel Sales Dias, and Samuel Silva. 2017. *An Introduction to Silent Speech Interfaces*. Springer.

[15] Masaaki Fukumoto. 2018. SilentVoice: Unnoticeable Voice Input by Ingressive Speech. In *The 31st Annual ACM Symposium on User Interface Software and Technology*. ACM, 237–246.

[16] Yang Gao, Yincheng Jin, Jiyang Li, Seokmin Choi, and Zhanpeng Jin. 2020. EchoWhisper: Exploring an Acoustic-Based Silent Speech Interface for Smartphone Users. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 3, Article 80 (Sept. 2020), 27 pages. https://doi.org/10.1145/3411830

[17] Mayank Goel, Leah Findlater, and Jacob Wobbrock. 2012. WalkType: using accelerometer data to accomodate situational impairments in mobile touch screen text entry. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2687–2696.

[18] Mayank Goel, Jacob Wobbrock, and Shwetak Patel. 2012. GripSense: using built-in sensors to detect hand posture and pressure on commodity mobile phones. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 545–554.

[19] Jose A. Gonzalez, Lam A. Cheah, James M. Gilbert, Jie Bai, Stephen R. Ell, Phil D. Green, and Roger K. Moore. 2016. A Silent Speech System Based on Permanent Magnet Articulography and Direct Synthesis. *Comput. Speech Lang.* 39, C (Sept. 2016), 67–87. https://doi.org/10.1016/j.csl.2016.02.002

[20] Tamás Grósz, Gábor Gosztolya, László Tóth, Tamás Csapó, and Alexandra Markó. 2018. F0 Estimation for DNN-Based Ultrasound Silent Speech Interfaces. https://doi.org/10.1109/ICASSP.2018.8461732

[21] W. Hardcastle, W. Jones, C. Knight, A. Trudgeon, and G. Calder. 1989. New developments in electropalatography: A state-of-the-art report. *Clinical Linguistics & Phonetics* 3, 1 (1989), 1–38. https://doi.org/10.3109/02699208908985268 arXiv:https://doi.org/10.3109/02699208908985268

[22] 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 Commun.* 52, 4 (April 2010), 301–313. https://doi.org/10.1016/j.specom.2009.12.001

[23] T. Hueber, G. Aversano, G. Cholle, 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

[24] 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 (2010), 288 – 300. https://doi.org/10.1016/j.specom.2009.11.004 Silent Speech Interfaces.

[25] Yan Ji, Licheng Liu, Hongcui Wang, Zhilei Liu, Zhibin Niu, and Bruce Denby. 2018. Updating the Silent Speech Challenge benchmark with deep learning. *Speech Communication* 98 (2018), 42 – 50. https://doi.org/10.1016/j.specom.2018.02.002

[26] Biing-Hwang Juang and Lawrence R Rabiner. 2005. Automatic speech recognition–a brief history of the technology development. *Georgia Institute of Technology. Atlanta Rutgers University and the University of California. Santa Barbara* 1 (2005), 67.

[27] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *23rd International Conference on Intelligent User Interfaces*. ACM, 43–53.

[28] Arnav Kapur, Utkarsh Sarawgi, Eric Wadkins, Matthew Wu, Nora Hollenstein, and Pattie Maes. 2020. Non-Invasive Silent Speech Recognition in Multiple Sclerosis with Dysphonia. In *Machine Learning for Health Workshop*. 25–38.

[29] Naoki Kimura, Tan Gemicioglu, Jonathan Womack, Richard Li, Yuhui Zhao, Abdelkareem Bedri, Alex Olwal, Jun Rekimoto, and Thad Starner. 2021. *Mobile, Hands-Free, Silent Speech Texting Using SilentSpeller*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451552

[30] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300376

[31] Yongkuk Lee, Connor Howe, Saswat Mishra, Dong Sup Lee, Musa Mahmood, Matthew Piper, Youngbin Kim, Katie Tieu, Hun-Soo Byun, James P. Coffey, Mahdis Shayan, Youngjae Chun, Richard M. Costanzo, and Woon-Hong Yeo. 2018. Wireless, intraoral hybrid electronics for real-time quantification of sodium intake toward hypertension management. *Proceedings of the National Academy of Sciences* 115, 21 (2018), 5377–5382. https://doi.org/10.1073/pnas.1719573115 arXiv:https://www.pnas.org/content/115/21/5377.full.pdf

[32] Richard Li, Jason Wu, and Thad Starner. 2019. TongueBoard: An Oral Interface for Subtle Input. In *Proceedings of the 10th Augmented Human International Conference 2019*. 1–9.

[33] Kent Lyons, Thad Starner, Daniel Plaisted, James Fusia, Amanda Lyons, Aaron Drew, and EW Looney. 2004. Twiddler typing: one-handed chording text entry for mobile phones. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 671–678.

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

[35] I Scott MacKenzie and Kumiko Tanaka-Ishii. 2010. *Text entry systems: Mobility, accessibility, universality*. Elsevier.

[36] 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

[37] 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*. 357–360.

[38] Geoffrey S Meltzner, James T Heaton, Yunbin Deng, Gianluca De Luca, Serge H Roy, and Joshua C Kline. 2018. Development of sEMG sensors and algorithms for silent speech recognition. *Journal of neural engineering* (2018).

[39] Carlos H Morimoto and Arnon Amir. 2010. Context switching for fast key selection in text entry applications. In *Proceedings of the 2010 symposium on eye-tracking research & applications*. 271–274.

[40] Y Nakajima, Hideki Kashioka, Kiyohiro Shikano, and Nick Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 5, V – 708. https://doi.org/10.1109/ICASSP.2003.1200069

[41] Shuo Niu, Li Liu, and D Scott McCrickard. 2014. Tongue-able interfaces: Evaluating techniques for a camera based tongue gesture input system. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 277–278.

[42] Lotte NS Andreasen Struijk, Eugen R Lontis, Michael Gaihede, Hector A Caltenco, Morten Enemark Lund, Henrik Schioeler, and Bo Bentsen. 2017. Development and functional demonstration of a wireless intraoral inductive tongue computer interface for severely disabled persons. *Disability and Rehabilitation: Assistive Technology* 12, 6 (2017), 631–640.

[43] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How do people type on mobile devices? Observations from a study with 37,000 volunteers. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.

[44] Laxmi Pandey and Ahmed Sabbir Arif. 2021. LipType: A Silent Speech Recognizer Augmented with an Independent Repair Model. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[45] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779* (2019).

[46] J. Picone. 1990. Continuous speech recognition using hidden Markov models. *IEEE ASSP Magazine* 7, 3 (1990), 26–41.

[47] F. Robineau, F. Boy, J. Orliaguet, J. Demongeot, and Y. Payan. 2007. Guiding the Surgical Gesture Using an Electro-Tactile Stimulus Array on the Tongue: A Feasibility Study. *IEEE Transactions on Biomedical Engineering* 54, 4 (2007), 711–717. https://doi.org/10.1109/TBME.2006.889180

[48] Anne Roudaut, Andreas Rau, Christoph Sterz, Max Plauth, Pedro Lopes, and Patrick Baudisch. 2013. *Gesture Output: Eyes-Free Output Using a Force Feedback Touch Surface*. Association for Computing Machinery, New York, NY, USA, 2547–2556. https://doi.org/10.1145/2470654.2481352

[49] Sherry Ruan, Jacob O. Wobbrock, Kenny Liou, Andrew Ng, and James A. Landay. 2018. Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 159 (Jan. 2018), 23 pages. https://doi.org/10.1145/3161187

[50] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James A Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23.

[51] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. 2014. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*. ISCA, 338–342.

[52] Javier San Agustin, Henrik Skovsgaard, Emilie Mollenbach, Maria Barret, Martin Tall, Dan Witzner Hansen, and John Paulin Hansen. 2010. Evaluation of a low-cost open-source gaze tracker. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications*. 77–80.

[53] Tanja Schultz. 2010. ICCHP Keynote: Recognizing Silent and Weak Speech Based on Electromyography. 595–604. https://doi.org/10.1007/978-3-642-14097-6_96

[54] R. Soukoreff and I. MacKenzie. 2003. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. *Conference on Human Factors in Computing Systems - Proceedings*, 113–120. https://doi.org/10.1145/642611.642632

[55] Simon Stone and Peter Birkholz. 2020. Cross-Speaker Silent-Speech Command Word Recognition Using Electro-Optical Stomatography. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 7849–7853. https://doi.org/10.1109/ICASSP40776.2020.9053447

[56] 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* (Berlin, Germany) *(UIST '18)*. ACM, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599

[57] 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* (Berlin, Germany) *(UIST '18)*. ACM, New York, NY, USA, 581–593. https://doi.org/10.1145/3242587.3242599

[58] László Tóth, Gábor Gosztolya, Tamás Grósz, Alexandra Markó, and Tamás Gábor Csapó. 2018. Multi-Task Learning of Speech Recognition and Speech Synthesis Parameters for Ultrasound-based Silent Speech Interfaces.. In *INTERSPEECH*. 3172–3176.

[59] Jason Tu, Angeline Vidhula Jeyachandra, Deepthi Nagesh, Naresh Prabhu, and Thad Starner. 2021. Typing on Tap: Estimating a Finger-Worn One-Handed Chording Keyboard's Text Entry Rate. In *2021 International Symposium on Wearable Computers*. 156–158.

[60] Outi Tuisku, Päivi Majaranta, Poika Isokoski, and Kari-Jouko Räihä. 2008. Now Dasher! Dash away! Longitudinal study of fast text entry by eye gaze. In *Proceedings of the 2008 symposium on Eye tracking research & applications*. 19–26.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010.

[62] Jingxian Wang, Chengfeng Pan, Haojian Jin, Vaibhav Singh, Yash Jain, Jason I Hong, Carmel Majidi, and Swarun Kumar. 2019. RFID Tattoo: A Wireless Platform for Speech Recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 4 (2019), 1–24.

[63] Tracy Westeyn, Helene Brashear, Amin Atrash, and Thad Starner. 2003. Georgia Tech Gesture Toolkit: Supporting Experiments in Gesture Recognition. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (Vancouver, British Columbia, Canada) *(ICMI '03)*. Association for Computing Machinery, New York, NY, USA, 85–92. https://doi.org/10.1145/958432.958452

[64] Jacob O Wobbrock. 2006. EdgeWrite: A versatile design for text entry and control.

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

[66] Jacob O Wobbrock. 2019. Situationally-induced impairments and disabilities. In *Web Accessibility*. Springer, 59–92.

[67] Jacob O Wobbrock, Shaun K Kane, Krzysztof Z Gajos, Susumu Harada, and Jon Froehlich. 2011. Ability-based design: Concept, principles and examples. *ACM Transactions on Accessible Computing (TACCESS)* 3, 3 (2011), 1–27.

[68] Jacob O Wobbrock, Brad A Myers, and John A Kembel. 2003. EdgeWrite: a stylus-based text entry method designed for high accuracy and stability of motion. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*. 61–70.

[69] Jacob O Wobbrock, James Rubinstein, Michael W Sawyer, and Andrew T Duchowski. 2008. Longitudinal evaluation of discrete consecutive gaze gestures for text entry. In *Proceedings of the 2008 symposium on Eye tracking research & applications*. 11–18.

[70] Pei Yin, Thad Starner, Harley Hamilton, Irfan Essa, and James M Rehg. 2009. Learning the basic units in american sign language using discriminative segmental feature selection. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 4757–4760.

[71] Steve Young, G Evermann, M.J.F. Gales, Thomas Hain, D Kershaw, Xunying Liu, G Moore, James Odell, D Ollason, Daniel Povey, V Valtchev, and Philip Woodland. 2002. *The HTK book*.

[72] Shumin Zhai and Per Ola Kristensson. 2012. The word-gesture keyboard: reimagining keyboard interaction. *Commun. ACM* 55, 9 (2012), 91–101.

[73] Mingrui Ray Zhang and Jacob O. Wobbrock. 2019. Beyond the Input Stream: Making Text Entry Evaluations More Flexible with Transcription Sequences. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 831–842. https://doi.org/10.1145/3332165.3347922

[74] Qian Zhang, Dong Wang, Run Zhao, and Yinggang Yu. 2021. SoundLip: Enabling Word and Sentence-level Lip Interaction for Smart Devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–28.