Mobile, Hands-free, Silent Speech Texting Using SilentSpeller

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

Tan Gemicioglu, Jon Womack,
Yuhui Zhao, Thad Starner
Georgia Institute of Technology
Atlanta, GA, USA

Abdelkareem Bedri
Carnegie Mellon University
Pittsburgh, PA, USA

Richard Li
Paul G. Allen School of Computer Science & Engineering
University of Washington
Seattle, WA, USA

Alex Olwal
Google Research
Mountain View, CA, USA

Jun Rekimoto
The University of Tokyo
Tokyo, Japan

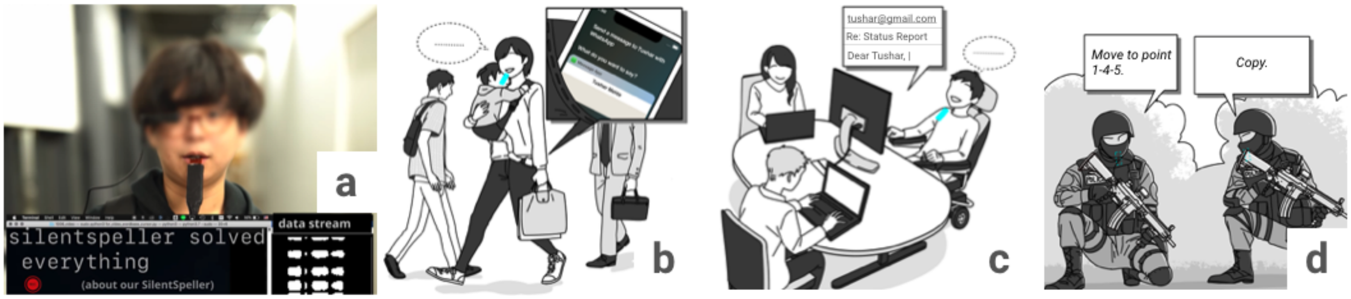

Figure 1: a) A SilentSpeller user wears the SmartPalate retainer whose 124 electrodes sense the position of the tongue at 100Hz. Applications include b) Hands-busy situations where speech is socially inappropriate c) users with low manual dexterity working in open office environments d) United Nations operations where silent communication is necessary

ABSTRACT

Voice control provides hands-free access to computing, but there are many situations where audible speech is not appropriate. Most unvoiced speech text entry systems can not be used while on-the-go due to movement artifacts. SilentSpeller enables mobile silent texting using a dental retainer with capacitive touch sensors to track tongue movement. Users type by spelling words without voicing. In offline isolated word testing on a 1164-word dictionary, SilentSpeller achieves an average 97% character accuracy. 97% offline accuracy is also achieved on phrases recorded while walking or seated. Live text entry achieves up to 53 words per minute and 90% accuracy, which is competitive with expert text entry on mini-QWERTY keyboards without encumbering the hands.

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, Jon Womack, Yuhui Zhao, Thad Starner, Abdelkareem Bedri, Richard Li, Alex Olwal, and Jun Rekimoto. 2021. Mobile, Hands-free, Silent Speech Texting Using SilentSpeller. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI '21 Extended Abstracts), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3411763.3451552

1 MOTIVATION AND CONTRIBUTIONS

Many conditions, such as stroke, multiple sclerosis (MS), Parkinson's disease, essential tremor, amyotrophic lateral sclerosis (ALS), cerebral palsy, and arthritis can limit a computer user's manual dexterity and necessitate alternative text entry methods. One solution is silent speech, which recognizes text entry via non-voiced speech. However, the hardware for current silent speech interfaces is often challenging, given technology that needs to be mounted on the body (e.g., ultrasound probes, electrodes or microphones attached to the neck and face) [3, 4, 8, 9]. In addition, most silent speech interfaces have high error rates when the user is on-the-go [4] and are limited to a relatively small vocabulary [3]. We present SilentSpeller, a device in the form of a dental retainer that tracks the tongue at 100 Hz using 124 capacitive touch sensors on the roof of the mouth (Figures 1 and 2). Instead of mouthing words, users spell words without voicing. Typing feedback may be displayed on a head worn display (HWD), smartwatch, mobile phone, or desktop

CHI '21 Extended Abstracts, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8095-9/21/05.
https://doi.org/10.1145/3411763.3451552

or each word can be spoken to the user through an earbud as it is recognized. While SilentSpeller is currently wired, which creates challenges with appearance, a wireless version could be made that fits completely in the mouth [5].

In addition to medical disabilities, situational impairments, such as mobility, interfere with text entry. For example, a mobile user may not have hands available to hold a smart phone or the visuo-manual attention to attend the screen for gesture typing (Figure 1b) [12]. While head-worn displays (HWD) allow a hands-free way to view a mobile screen, many users are reluctant to use speech input in public due to privacy concerns or social opprobrium. One-handed controllers for fast, silent, and eyes-free mobile text entry, such as the Twiddler keyboard, require significant training to operate at rates above hunt-and-peck desktop speeds (30 words per minute) [7]. Some tasks require communication in high noise or silent environments that preclude the use of voiced speech recognition; interviews with special operations leaders indicate a need to communicate silently among members of the team (Figure 1d), and soldiers have described a need for subtle and silent communication while on presence patrols. Any text messaging system should be hands-free, robust to body movements, and, preferably, not require much training to achieve fast texting rates.

With SilentSpeller, we offer the following contributions:

- **Optimization experiments** to determine the amount of data needed to train a SilentSpeller recognizer per user. Two participants each spelled 2328 words (1164 unique words twice). SilentSpeller achieves an average 97% character accuracy (92% word accuracy) and reaches maximum accuracy within 1500 words of training.
- **Walking versus seated experiments** that establish that SilentSpeller tolerates user movement during input with little degradation of performance.
- **An interactive text entry system** that combines the spelling recognizers with gestures for editing.
- **Text entry experiments** using the standard MacKenzie-Soukoreff phrase set where SilentSpeller users “type” up to 53 words per minute (43 average) with 90% accuracy (88% average).

2 SILENT SPELLING TEXT ENTRY

In Silent Spelling, users silently mouth each letter in the context of spelling a word. An interesting advantage over mouthing words is that there is no ambiguity due to homophones. We use Complete Speech's SmartPalate, which is a dental retainer-type device with 124 binary capacitive sensors that lines the user's palate and captures tongue movements (Figure 2). SmartPalate was originally developed for speech therapy to correct pronunciation. Data is sampled at 100Hz and sent to a personal computer or smart phone via a wired USB hub for analysis. Since the SmartPalate fits firmly in the top of the mouth, we expect SilentSpeller to be tolerant to body movements [6]. SmartPalate requires each user obtain a dental impression so that the electrode array can be custom fit to each user's mouth (Figure 2a). Covid-19 restrictions limited the number of participants who could be fitted at this time.

2.1 Recognizer Pipeline

Principal Component Analysis is performed on training data sets (which are kept independent from test data). Based on empirical testing, we chose the top 16 components for recognition. As each silently spelled word is collected, each data frame of 124 binary electrode values is projected to the top 16 principal components. The resulting 100Hz 16-dimensional signal is then decoded using hidden Markov models. Preliminary testing suggests HMMs outperforms neural net-based methods for this data set. The models are first trained on letters and then on triletters, akin to phone and triphones in conventional speech recognition systems. Tied-state triletters are used to reduce error for triletters with limited occurrence in the training data set. We choose a 12-state left-to-right HMM topology with no skip transitions based on early experiments.

2.2 Corpus and Participants

We use the Mackenzie-Soukoreff phrase set, which consists of 500 phrases, 1164 unique words, and 7048 letters [11]. Each phrase is about five 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. 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. All experiments were conducted in participants' respective homes, using Apple MacBook Pro laptops.

To collect samples of silent spelling, we developed a push-to-talk style recording application. 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 re-record the word. Participants are allowed to take a break when desired. After every word is recorded, an estimate of speed (wpm) is displayed. The 2328 word data sets required approximately five hours of input for each of the two participants.

2.3 Offline Isolated Word Testing (1164-word Dictionary)

Using the 2328-word data sets from P1 and P2, we optimized the parameters of the model. In general, we average results over 10-fold cross-validation (i.e., independent training and test sets, random 10% for testing each fold) for testing. We swept over two through 18 states and discovered that 12 states provided good overall accuracy and still worked on the fastest articulated letters. Recognizers trained from the two 2328 word data sets performed exceedingly well, achieving 97% character accuracy and 93% word accuracy on P1 and 97% and 91%, respectively, on P2.

3 TOLERANCE TO ON-THE-GO INPUT

SilentSpeller, by its nature, can not be as fast at text entry as a conventional silent speech system; however, it may enable text entry while in motion, providing an advantage over EMG, camera, and ultrasound systems. The electrode array fits snugly in the mouth, and the tongue is relatively isolated from the mechanical shock of walking; otherwise, voiced speech while walking would not be possible. Given these attributes, we expect SilentSpeller to be as accurate at recognizing silently spelled words when the user is walking as when seated.

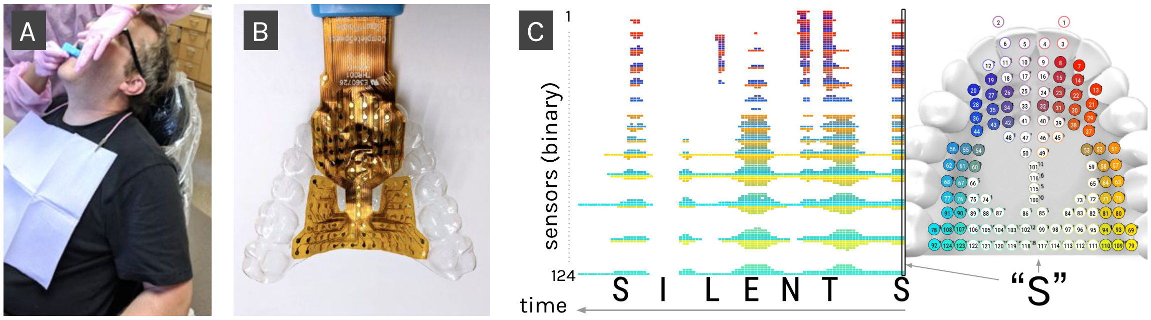

Figure 2: a) Dental impression needed for custom-fit SmartPalate. b) Resulting SmartPalate. c) Palatogram and electrode map. Note that individual letters are not recognized in real-time but are added to the image for illustration purposes. Time flows from the right so that letters are spelled in correct order and the right side of the palatogram displays the current state of the electrodes.

3.1 Experiment Settings

P1 and P2 provided 107 phrases each for both a walking and seated condition (a total of 428 phrases). The 107 phrases are from the MacKenzie-Soukoreff phrase set and consist of 556 words; 321 of them are unique. Common words occur repeatedly in the phrases. For example, the most frequently used word was "a" which appeared 24 times. 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 while capturing the 107 phrases. The seated condition was captured at a desk.

Since our goal is to compare walking versus seated text input we chose to use the most advantageous training that is reasonable for this study. For each of the two participants, the recognizer is trained on their 2328 isolated dictionary words plus the 107 phrases from the condition not being tested. In other words, the recognizer for the seated condition was trained with the 2328 words plus the 556 words from the 107 phrases collected during the walking condition. Similarly, the recognizer for the walking condition was trained with the 2328 words plus the 556 words from the 107 phrases collected during the seated condition. No part of any test set is used in training. During recognition, the system is limited to a dictionary constructed from the 321 unique words from the 107 phrases. A bigram is constructed using only the 107 phrases and Laplace smoothing (so that any word combination is possible).

3.2 Results and Discussion

Table 1 presents the results of the study. There is almost no difference in the accuracy between the seated and walking conditions, demonstrating the robustness of SilentSpeller to body motion.

participant-condition	character (word) accuracy
1-seated	97% (95%)
1-walking	97% (95%)
2-seated	94% (90%)
2-walking	93% (91%)

Table 1: Comparing walking to seated text input.

4 LIVE TEXT ENTRY USER STUDY

In informal experiments emulating SilentSpeller by simply spelling the words in the MacKenzie-Soukoreff phrase set as fast as possible, we found text entry surpassed 50 wpm, which is equivalent to the expert text entry rates on physical [2] and virtual [10] mini-QWERTY keyboards. While accuracy varies between the participants in the experiments above, the results show that SilentSpeller holds promise as a means of text entry. Adding an interface so that the user can select between the top n-best candidates returned by the recognizer should improve the speed and usefulness of the interface.

4.1 Participants and corpus

The same two participants performed the live text entry user study. The recognizer was trained with the 556 words from the 107 phrases collected while seated for each participant plus 500 words were chosen at random from the 2328 isolated words (this method was chosen for compatibility with an on-going study with more participants). For testing, the participants attempted to input the 107 phrases again, as quickly and as accurately as possible.

We employ a bigram stochastic grammar trained on the 107 phrases with Laplace smoothing. Upon inference, HTK returns a 20-best list of candidates with their likelihoods. Applying the bigram to this list determines the top candidates.

4.2 Text Entry using SilentSpeller

Following previous work [1, 7, 11], we implemented an interface application to test the speed and accuracy of text entry using SilentSpeller on the MacKenzie-Soukoreff phrase set. We provided a video to instruct participants in how to use the interface. The application presents phrases to the participant who then transcribes them over the course of 20 minutes. Using the SilentSpeller app is similar to using a gesture keyboard [12], included on most smartphones. Three interactions are provided: INPUT (silent spelling),

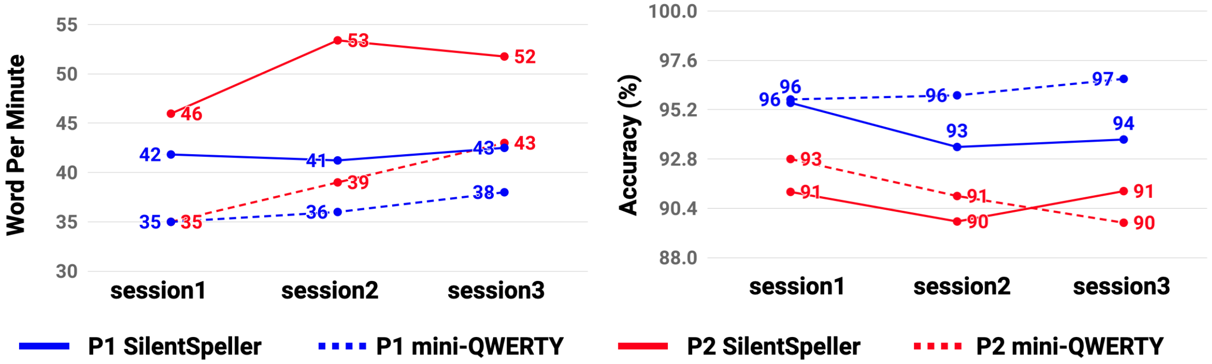

Figure 3: Live text entry results. Words per minute (left) and accuracy (right) for each of the participants' three sessions. Solid lines shows the results using SilentSpeller. Dotted lines shows the results using usual mini-QWERTY keyboard.

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). For each phrase, the user presses a push-to-record button and inputs an individual word by silently spelling with the SmartPalate. Upon button release, the captured data frames are sent to the recognizer. About a second later, the interface shows the user a list of the five best word predictions in order of probability. If the first best candidate is correct, it is accepted as soon as the next input is started. If the correct answer is in the list of five, the user selects the best candidate with the TAP gesture. If there is no correct answer among the five, the candidates are deleted with the STICK gesture and the system returns to the input state. Once the user has completed a phrase, the user presses the right shift to advance to the next phrase.

4.3 Results

Figure 3 shows the results of the live text entry experiment. Participants did the same experiments using the mini-QWERTY keyboard on their smartphones for comparison. First, it was shown that for both participants, text entry by SilentSpeller was faster than by mini-QWERTY. The average maximum session speed over each participant's three sessions was 48 wpm. Average text entry accuracy (1 - TER) for those respective sessions was 92%. Unlike the previous offline experiments, this accuracy metric considered user failures in typing the correct letter, recognizer failures, and corrections. As expected, participants quickly adapted to silently spelling words for text entry. 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 90% 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 more investment in improving the recognition rates may cause the other participants to reach similar speeds. At the end of the experiment, P1 and P2 attempted another 20 minute live text entry session while walking and saw similar results to their seated performance, as expected.

5 CONCLUSION

We present 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 demonstrate texting rates competitive with mobile phone virtual QWERTY typing, but without encumbering the hands. These results suggest SilentSpeller can be an efficient text entry system and may find niche applications for on-the-go, silent, hands-free text entry or silent text entry for people with movement impairments. Further work will explore specific application domains, additional sensors for the lips to tune recognition accuracy, and user independent and user adaptive recognition.

REFERENCES

[1] Edward Clarkson, James Clawson, Kent Lyons, and Thad Starner. 2005. An Empirical Study of Typing Rates on Mini-QWERTY Keyboards *(CHI EA '05)*. ACM, New York, NY, USA, 4.

[2] 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 *(MobileHCI '14)*. 339–348.

[3] 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.

[4] Arnav Kapur, Shreyas Kapur, and Pattie Maes. 2018. AlterEgo: A Personalized Wearable Silent Speech Interface. In *IUI2018*. ACM, 43–53.

[5] 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.

[6] Richard Li, Jason Wu, and Thad Starner. 2019. TongueBoard: An Oral Interface for Subtle Input. In *AH2019*. 1–9.

[7] 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.

[8] 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.

[9] Jun Rekimoto Naoki Kimura, Michinari Kono. 2019. An Ultrasound Imaging-Based Silent Speech Interaction Using Deep Neural Networks. CHI2019.

[10] 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.

[11] R. Soukoreff and I. MacKenzie. 2003. Metrics for text entry research: An evaluation of MSD and KSPC, and a new unified error metric. *CHI2003*, 113–120.

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