Text Entry Tap Accuracy and Exploration of Tilt Controlled Layered Interaction on Smartwatches

Mark D Dunlop
University of Strathclyde
Glasgow, Scotland, UK
mark@textentry.org.uk

Marc Roper
University of Strathclyde
Glasgow, Scotland, UK
marc.roper@strath.ac.uk

Gennaro Imperatore
University of Strathclyde
Glasgow, Scotland, UK
gennaro.imperatore@strath.ac.uk

ABSTRACT
Design of text entry on small screen devices, e.g. smartwatches, faces two related challenges: trading off a reasonably sized keyboard area against space to display the entered text and the concern over "fat fingers". This paper investigates tap accuracy and revisits layered interfaces to explore a novel layered text entry method. A two part user study identifies preferred typing and reading tilt angles and then investigates variants of a tilting layered keyboard against a standard layout. We show good typing speed (29 wpm) and very high accuracy on the standard layout – contradicting fears of fat-fingers limiting watch text-entry. User feedback is positive towards tilting interaction and we identify ~14° tilt as a comfortable typing angle. However, layering resulted in slightly slower and more erroneous entry. The paper contributes new data on tilt angles and key offsets for smartwatch text entry and supporting evidence for the suitability of QWERTY on smartwatches.

Author Keywords
Text-entry; smartwatch interaction; user experimentation; layered interfaces; tap accuracy; fat finger.

ACM Classification Keywords
H.5.2 User Interfaces.

INTRODUCTION
Tapping accurately on small screens is a widely stated problem for touchscreen interaction. As a result smartwatch interfaces have largely avoided placing keyboards on the devices but have instead opted for canned replies, voice recognition or emoji drawing to support replies to received messages. These methods excessively constrain smartwatch users' ability to respond to messages received on the watch itself and limit other interactions where text input is necessary.

MobileHCI '17, September 04-07, 2017, Vienna, Austria

ACM ISBN 978-1-4503-5075-4/17/09.
http://dx.doi.org/10.1145/3098279.3098560

Recent research [15] has shown promise for text entry on smartwatches being able to achieve reasonable speeds for short phrases using a near standard QWERTY layout. However, this comes at the cost of a keyboard that uses over 85% of screen space (Figure 1). This prevents easy review of messages entered but is understandable as it is necessary to maximize the screen real estate for better separation of taps on very small targets of under 4 mm diameter.

In this paper we revisit semi-transparent layered interfaces as a potential solution that maximizes screen real estate for both the keyboard and for reviewing the typed message. After a background review of small screen text entry and layered interfaces, the paper presents two studies: An initial parameter setting study investigating reading and writing tilt angles; A comparative study investigating user performance on smartwatch text entry using a standard layout keyboard and two variants of the tilting keyboard.

Figure 1: Experiment Standard Text Entry Interface

APPROACHES TO WATCH TEXT ENTRY
Despite an array of research on alternative layouts (e.g. [5,6,13,31,35,39]) and alternative approaches (e.g. [24,36,44,46]), the traditional QWERTY layout has persisted on mobile phones and tablets despite being clearly sub-optimal for touchscreen text entry – following similar arguments to those for desktop concerning familiarity of the layout dominating over speed gains (e.g. [10]). Unlike large laptop/desktop keyboards where keypresses are assumed to be unambiguous, albeit often subject to spell-correction, entry on touch-screen keyboards is normally considered to be imprecise. User input is taken as an indication of his/her intension and interpreted by a combination of tap models and language models.

Tap models are required to interpret between continuous (x,y) screen coordinates and discrete keys. This is a particular concern on smartphones and smaller devices as people's fingers are large compared to the on-screen keys, typically 6-7 mm smartphone touchscreen key width compared to human finger tips of around 16-20 mm [9]. This is known as the “Fat Finger Effect” (e.g. [3,20,22]) after concerns raised particularly by older adults on using finger touch interaction [40]. How people tap varies between individual (e.g. [1]), between devices (e.g. [20]) and based on how people hold the device (e.g. [1]). Taking these factors into account can considerably improve tap accuracy and text entry (e.g. [21]). More advanced tap models have been developed that exploit, for example, the built-in accelerometers to adjust the tap model based on the walking pattern of the user [14].

Language models have long been a part of mobile text entry starting with unigram word “predictive text” approaches on 12-key physical pad phones [12,23,29] through bi-gram word models (e.g. [19]) to complex models based on combining multi-gram letter and multi-gram word models (e.g. [43]).

An alternative to one-touch-per-key text entry typical on phones is to build on the strength of language modelling to support a single gesture across the keyboard per word. Initially proposed with a custom keyboard [28] this has been widely adopted on smartphone text entry using the QWERTY layout [47]. Gesture approaches have also been developed that exploit accelerometers for touch-free entry (e.g. [24,36]) or to adjust button-based entry (e.g. [45]).

The fat-finger problem is particularly concerning when considering text entry on very small touchscreen devices such as smartwatches. Inspired by the success of predictive text on traditional phones, ambiguous keyboards have been researched to provide large but multi-letter keys on smartwatches either following an alphabetic layout [11,27] or an optimized layout [11]. Alternatively the standard QWERTY layout can be supported with additional interaction to make it more suitable for very small screens. The Zoomboard [34], for example, requires users to tap initially to zoom into a section of the keyboard before selecting an enlarged key. While ZShift [30] allows users to adjust their taps using a miniature zoom key callout of the area they are touching. These approaches, however, resulted in slow reported typing speeds: users are slow to learn a non-QWERTY layout and any form of interaction that involves mental response to displayed interaction tools is inherently slower than simple tapping. Simple tapping supported by a strong language model or gesture typing, e.g. Velocitap [43] and WatchWriter [15], have recently been shown to get speeds more in line with smartphone entry speed.

LAYERED INTERFACE DESIGN

Layering information so that partially-transparent layers are presented superimposed on the same display area is not common in interface design but can be effective. Originally developed for desktop interfaces such as toolbars [16] and menus [17] so that the interaction tools don’t obscure the underlying work area, they have been revisited for photographic manipulation tools with careful design of transparency not to interfere with underlying colour images [4]. Transparent controls have been used on mobile interfaces to overcome lack of screen estate [25] but have not been widely adopted.

PROPOSED LAYERED TILT CONTROLLED KEYBOARD

This paper reports our investigations to further investigate tap accuracy for typing and speed on a smartwatch and to explore tilt-based control. In our design tilting the watch fades between a full-screen keyboard (Figure 2 right) and a full-screen display of text of the message being typed (Figure 2 left). In intermediate states a transparency merged version is shown proportional to the tilt (Figure 1 middle). The tilting keyboard stretches the standard QWERTY layout to almost fill the watch screen leaving only space for a suggestion bar at the top of the screen (a 25% increase in the height of the keyboard) and exploits the widest part of the round face to increase the width of the middle row by 23%.

The rest of this paper reports two studies:

1. A parameter setting study to assess if there is a difference in the tilt angle of the watch for reading vs writing and parameterize the second study.
2. A comparative performance study of text entry on the standard layout (Figure 1) and the tilting layout (Figure 2).

Both studies were conducted at University of Strathclyde under institutional ethical approval with users given a small shopping voucher in appreciation of their time

Figure 2: Experiment Tilt Controlled Layered Text Entry Interface – full text (left) through to full keyboard (right)

STUDY 1: PARAMETER SETTING

Before finalizing the design of our keyboard we had to establish the tilt angles for typing and reading. In this parameter setting study participants were asked to (1) read three short passages of text on the watch, (2) enter six short phrases on the standard keyboard, and (3) demonstrate how they would angle the watch for reading and writing.

For reading tasks participants were given one short paragraph to read as practice while free to ask questions, they then read two longer paragraphs without interruptions. The paragraphs were the introductory sections of classic children’s stories of 94 [32], 150 [37] and 197 [2] words respectively. This involved scrolling using a vertical stroke gestures with a green next button () being double tapped to move on to the next phrase. Participants were told in advance that they would be asked some questions concerning the passages to encourage accurate reading. For phrase 2 and 3 the tilt angle from vertical (Figure 3) was recorded every second after the initial scroll gesture to eliminate initial user settling.

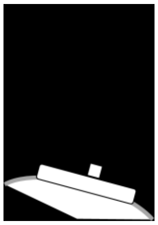

Figure 3: Tilt angles were measured from vertical parallel to the direction of the watch strap using internal accelerometer.

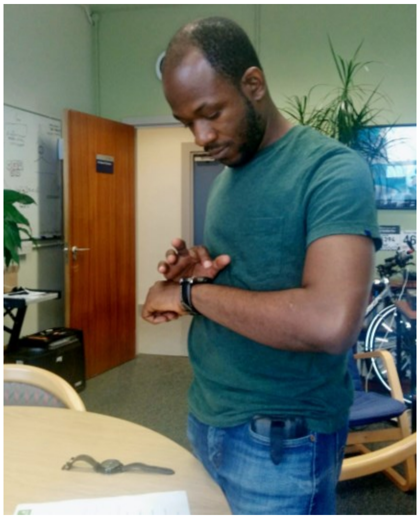

Figure 4: Participant conducting study

In writing tasks participants typed six short phrases from the standard Enron collection [41]. The first two were treated as practice with participants free to ask questions before silently completing the four test phrases (with double tap on ▶ to enter each phrase). While typing the angle was recorded after first keystroke.

At the end of the study users were individually verbally prompted as follows: "One of the reasons of this study is to see if people read and type at different angles on a smartwatch [experimenter's wrist rotated back and forth]. If you could separate reading and writing angle, please show me your reading angle [A] and your writing angle [B]." The displayed angle was logged by the experimenter pausing and tapping on the paired smartphone at A and B.

Participants and Equipment

Thirteen participants were recruited for this short study. All experiments were conducted while standing (figure 4) and wearing an LG Urbane smartwatch on their non-dominant hand with the index finger of their dominant hand used for typing. 13 participants were recruited through University mailing lists and were a range of undergraduate/postgraduate students and research assistants: 7 male, 6 female; median age 29 (range 18-41); none had previous experience of using a smartwatch but all had touchscreen phone experience. All users had self-declared normal, or corrected to normal with single vision lenses, eyesight and were self-declared native or fluent English speakers. The LG Urbane is a 1.3 inch diameter screen Android smartwatch with resolution of 320 x 320 pixels running Android Wear 1.0.

The keyboard used here (and the standard condition in study 2) approximately matched the screen usage and shape of Gordon et al's [15]. To increase alphabetic key sizes, space was implemented as a left-to-right swipe and backspace as right-to-left with the keys for these removed from the keyboard. For experimental purposes no capitalization or extended character panels were available, but these could be implemented using, say, vertical gestures or long presses. A fixed apostrophe and dash were available on the bottom row of the keyboard to support in-word punctuation. A tap on the single line text display revealed that text full screen for review, with a second tap returning to entry mode.

The keyboard used a 8-gram character language model combined with a tap to key centre measure to interpret user taps on the keyboard into characters. Given a tap coordinate x,y each letter was assigned a score, $T_{x,y}$ based on the Gaussian probability of tap being on the key. In the current implementation the mean tap was assumed to be centre of the key with standard deviation set at one-key width and a uniform circular distribution assumed. Analysis of study 2 results could improve this approximation in line with actual tap patterns, which tend to off-centre and elliptical [1]. Given a history string H a Witten-Bell [7] model of decay was used to score the next most likely letter, M_H, using a window of 7 previous characters and a base unigram model for the next character. The final score for a letter is given by $T_{x,y}^2\ M_H$ as this gave the best combination of language model accuracy and tap flexibility in pre-study tests. As the history itself is uncertain, a set of 6 most likely patterns was kept between taps and used to populate a new set of most highly scoring sequences. Space was modelled as an unambiguous character and, for efficiency, fixed the pre-space characters to the best, or selected, word. Backspace was modelled within word as rolling back the prediction and between words by recreating the tap pattern using key centres of the entered word.

Based on our combined tap and language model, the most likely suggestion was automatically inserted character-by-character as the user typed with three alternatives being offered on a suggestion bar. As per standard behaviour a space was inserted after suggestion bar choice, but within word punctuation had to be entered manually. The language model accounted only for miss-taps on the keyboard and did not include more advance spell correction for rarer omitted or duplicate letters (e.g. [8,43]). The first suggestion and text entry area are biased to show a complete word when possible, as opposed to prefixes of longer words, to reduce changes on space. As it has been shown that word predictions can lead to excess workload and can slow users [12,38], users were required to type whole words with use of the suggestion bar limited to correcting wrong predictions.

Our studies followed a standard text entry transcription task approach with users typing large-print phrases from paper onto the watch. Wizard-of-Oz style studies have been used for tap analysis so that users can tap naturally with only their position in the text being shown, and not actual characters entered (e.g. [1]). As we wanted to also assess speed of our different keyboards we felt this was not ideal and wanted users to have to type relatively accurately but not be over constrained by language modelling misinterpreting taps. We thus needed a high quality text entry method that would perform well in studies. Our tap and language model appeared to work fairly well but to improve performance given the processor and memory constraints of an experimental system we focused the training of the language model. We trained on the full Enron-50 memorable phrases augmented with three small lists of common words in English (containing 100, 200 and 200 words) and a list of 200 most common word-bigrams[2]. This tuned the language model closely to the study set but did not exclude erroneous typing or incorrect suggestions.

Results

Figure 5 shows the average tilt angle for reading and typing from parts one and two of study when users were doing these tasks without knowledge of the tilt sensor monitoring (left) and from part three where users stated both angles (right). The results show a significant difference in the observed average angle for reading (M_r=8.8°, SD_r=10.8°) and writing (M_w=15.0°, SD_w=6.3°) (t(12)=2.30, p=0.041). No significant difference was observed for stated angles (read $M_{r'}$=23.8°, $SD_{r'}$=15.5°; writing $M_{w'}$=15.6°, $SD_{w'}$=10.9°) (t(12)=1.70, p=0.116).

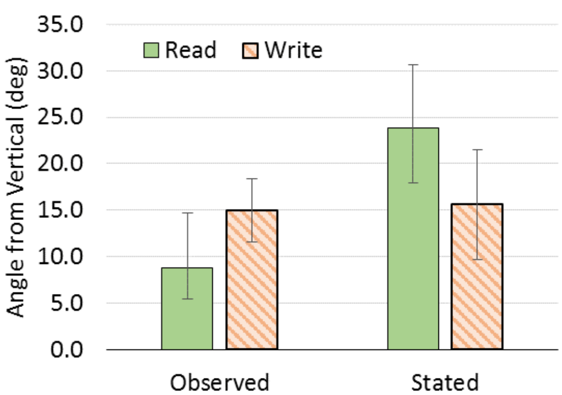

Figure 5: Observed and Stated Read and Write Angles (degrees from vertical, 95% confidence interval error bars)

Discussion

In both observed and stated measurements there was less variance on writing angle than reading angle. Furthermore the writing angle between observed and stated was very close while there was a marked difference between observed and stated reading angles. In observed tasks users overall preferred a flatter screen angle for reading but stated a stronger tilt when asked explicitly. This is unexpected and further studies need to be conducted to establish if this was an ordering issue as the initial observed shallow reading angle was before any text entry on the watch. However, there was also variation between users in both reading observed and stated angles.

In terms of parameters for part 2 of the study, this study strongly suggests a 15° angle for writing. For reading the most flexible would be to allow a reading position both sides of this to support both flatter and more tilted reading angle. In observations the mean read angle was 6.2° flatter than the mean write angle while when stated it was 8.3° more tilted suggesting a possible range of 15°±9° for a three position tilt keyboard but this range was too sensitive to fine movement so was extended to 15°±21° in study 2.

STUDY 2: KEYBOARD TEXT ENTRY

To assess typing accuracy, speed and keyboard usage in practice we conducted a standard laboratory based comparative study between our tilt-controlled layered keyboard and a more standard keyboard with minimal space for entered text.

Keyboard Parameter Setting

The layered keyboard has two layers: a full screen keyboard and a full screen text view showing the text entered. At one extreme the keyboard is shown nearly opaque with a watermark of the underlying entered text (Figure 2 right) while at the other extreme the text is shown with a watermarked keyboard (Figure 2 left). Between these two angles a linear function adjusted the opaqueness of each layer with simple exponential smoothing to reduce flicker. The colour saturation was also adjusted so that layers were given a green tint as they became more transparent.

Following study 1 the comfort angle for writing is 15°. As such a condition for study 2 is a tilting keyboard where the keyboard is opaque at 15° and users tilt forward or backward to see the entered text. However, it could be argued that users will start the study focusing on the keyboard and as they become more experienced will move towards a more transparent keyboard to focus on the text they are typing. In an ideal situation the final transparent keyboard angle should, thus, correspond with the preferred 15° writing position from study 1. Study 2 was designed to compare these two design options with each other and with the more standard layout used in study 1.

The study hypotheses were:

- H1. Text entry will be faster using the tilting keyboard;
- H2. Text entry will be more accurate using the titling keyboard;
- H3. The average angle of typing and variance of viewing angle will differ between the keyboards;
- H4. As the study progresses users will tend towards a more transparent keyboard (transparent conditions).

The three study conditions were:

1. Standard keyboard with manual swapping between near full screen keyboard and full screen text view (by user tapping on the text);

2. Keyboard-Focused Layered Keyboard: a layered keyboard with the central 15° tilt angle resulting in a *near opaque full screen keyboard* with watermark only of entered text (user tilts to see text);
3. Review-Focused Layered Keyboard: a layered keyboard with the central 15° tilt angle resulting in a *near opaque full screen text* with watermark only keyboard (user tilts to see keyboard).

As per study 1, the space key and backspace were replaced with a horizontal gestures.

	Standard	*Keyboard Focus*	*Read Focus*
Central 15° tilt angle			
Maximal offset from 15°			

Table 1: The three conditions in study 2

Participants and Equipment

26 users[1] were recruited through email and poster advertising at Anonymous University (20 male, 5 female, 1 declined; median age 22, range 18-39, 2 declined). Of the users 3 owned a smartwatch, 1 had limited prior use, and others had no previous use. All users were regular touch screen phone users with self-rated fluency in English and good, or good corrected using single-focus lenses, eyesight.

The same LG Urbane watches and same prediction engine as per study 1 were used in study 2 (with minor fixes to handling of backspaces in the keyboard). As per first study users stood throughout, wore the watch on their non-dominant wrist and typed with their dominant index finger (see Figure 4).

Study procedure

Within group study with users randomly allocated to 6 permutations of the three keyboard conditions. Users completed an initial demographic background form then typed 33 phrases from Enron set [41] on each keyboard in one practice block of 3 phrases then three blocks of 10 test phrases. All tasks done while standing. A NASA TLX [18] form was completed after each condition. Finally a short exit questionnaire on preferences and issues with was completed.

[1] Two further users withdrew from the study: one started feeling ill while the other exceeded the time limit of 1 hour. Their demographics and data was ignored.

Results

Text entry was analysed for speed (measured in words per minute, WPM) and accuracy (represented by number of backspaces used during typing and Character Error Rate, CER, the Levenshtein distance between target and typed phrases normalised to the length of the original phrase). Results were analysed using repeated measures ANOVA on 3 keyboards x 3 measures. The watches' average tilt angle and stability (RDMS) was analysed separately (3 keyboards x 2 measures). Mauchly's sphericity test and Šidák adjustment were used.

Speed

Timing was recorded from first keystroke to last keystroke of each phrase with words per minute defined using the standard five characters per word metric. Analysis shows a significant difference between the three keyboard conditions ($F(2,25)=5.29$, $p=.008$). Figure 6 shows the standard keyboard fastest with a mean entry rate of 29.2 wpm compared with 27.2 and 26.7 wpm for keyboard-focus and read-focus tilting keyboards. The difference between standard and keyboard-focus & read-focus was significant ($p=.046$ and $.045$ respectively) while the difference between the two tilting variants was not ($p=.89$).

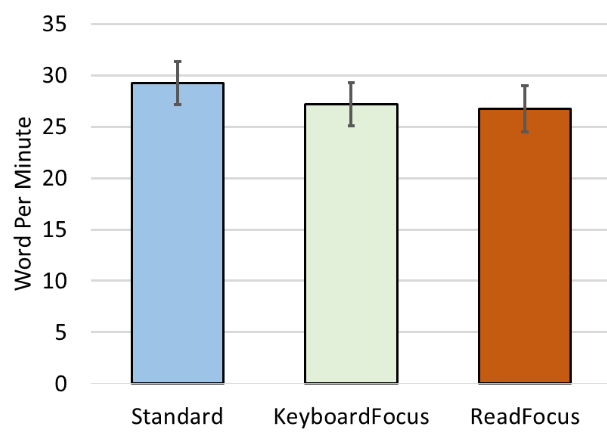

Figure 6: Speed of text entry on three keyboard versions (error bars are 95% confidence intervals)

Accuracy

Overall entry was very accurate with Levenshtein distances of 0.25 per phrase with little use of backspace (approx. 2 backspaces per phrase)(Figure 7). There was a significant difference between the three keyboard versions for backspace usage ($F(2,25)=4.34$, $p=.018$) but not for Character Error Rate, ($F(2,25)=0.098$, $p=.851$). Although usage was low, users did use backspace significantly less with the standard keyboard than the keyboard-focus condition ($M_{st}=1.79$, $M_{kf}=2.30$, $M_{rf}=2.39$ – pairwise Std-KF $p=.020$, Std-KF $p=.073$, KF-RF $p=.967$).

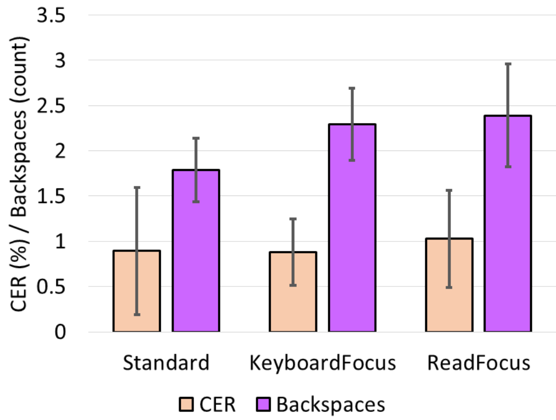

Figure 7: Accuracy of input (CER and backspaces per phrase)

Tilt angle and stability

We measured the tilt of the watch every time the user tapped a key and observed a distinct difference between the three keyboard variants. Unfortunately tilt log data was lost for early participants resulting in N=12 for tilt analysis.

The standard variant mean angle was 13.1° while the KeyboardFocus tilt condition was 10.0° and the ReadFocus 9.2° (see Figure 8). Based on study 1, we set the central angle at 15° at which the keyboard focus condition showed a near-opaque keyboard and the read focus condition near-opaque text. The angles observed of 10° for keyboard-focus and 9° for read-focus translate to a keyboard visibility of 77% and 26% respectively. The difference in overall angle was significant ($F(2,22)=4.015$, $p=.033$) with post-hoc Šidák tests showing a marginal difference between standard and read-focus, $p=.065$).

As we were interested in the learning effects we analysed tilt angle separately for final task set (final 10 phrases of 30 study phrases per keyboard). Here the mean angle was 14° (+1° from overall average) for the standard keyboard, 11° (+1°) for the keyboard-focus condition and 8° (-1°) for read-focus. A small change in all cases that was towards the "natural typing angle" identified in study 1 for both standard and keyboard-focus but away from in the read-focus condition. In both tilting conditions this lead to greater keyboard visibility of 80% (+3%) and 32% (+6%). As with

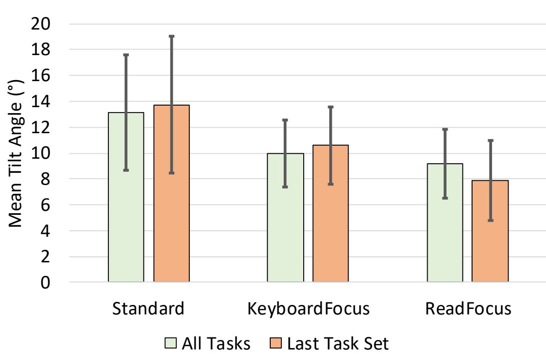

Figure 8: Mean tilt angle

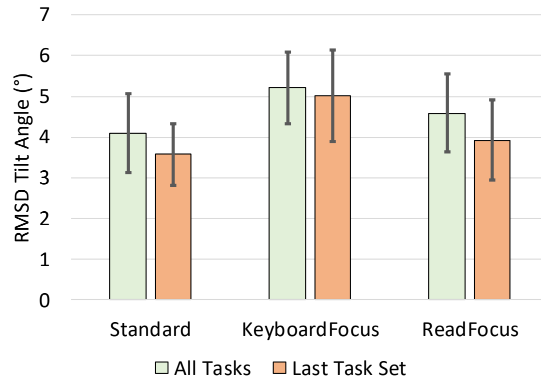

Figure 9: RMSD tilt variance from user average tap angle

overall mean angles, the difference in last-set angle was significant ($F(2,22)=5.51$, $p=.011$), post-hoc tests showed standard significantly different from read-focus, $p=.038$).

We were interested in how much users adjusted the watch angle while typing, for example to adjust the angle to make text more visible in between typing characters. In addition to logging the tilt angle whenever a keystroke was recorded the watch recorded the angle periodically during tasks. We calculated the Root Mean Square Deviation of these regular tilt recordings from each users' individual mean angle while typing. As shown in figure 9 the variance was lowest for the standard keyboard followed by the read-focus with the keyboard-focus having highest variance. This pattern was repeated with lower variance for the last 10 phrase set per condition. ANOVA analysis showed the difference for all phrases was not-significant while for the last ten phrases a significant difference does exist ($F(2,22)= 3.92$, $p=.035$) but no pairwise differences were significant.

Task Load Index

The NASA Task Load Index (TLX) analysis showed significant differences between the three keyboards in overall sum of scores ($M_{std}=39.2$, $M_{kf}=43.9$, $M_{rf}=49.2$; $F(2,50)=5.68$, $p=.006$) with post-hoc Bonferroni tests confirming workload significantly higher for read-focus than standard keyboards ($p=.007$). Figure 10 shows the individual TLX scales and identifies significant differences for mental, physical, effort and frustration. Figure 10 also highlights overall acceptable workloads with only one mean (physical read-focus) reaching the mid-point of the scales.

Qualitative feedback

In the end of the session users were asked to state three good and bad things about the tilting keyboard conditions. Table 2 lists representative samples of participant quotes (with similar quotes grouped and counts given in parenthesis). Overall this paints a picture of enthusiasm for tilt switching and keyboard accuracy but many problems with the cross-fading approach.

Positive Comments	*Negative Comments*
"You could easily check to see what was already written very easily." (8)	"A lot of flashing wouldn't allow me to focus on my writing." (9) "especially on [read-focus]" (2) and "foresee accidentally tilting … while walking" (1)
"All of the message is displayed." (7)	"The already entered text sometimes blocked your view of the keys." (7) or "was distracting" (3)
"Allowed you to easily switch between keyboard and typed view." (6)	"The point where the text can be clearly seen is hard to find and maintain." (7)
"Even when the focus was on the text, I was able to see the keyboard well." (5)	"It was not immediately obvious whether tapping on the screen while the keys were faded out would input text or perform another task, like move the cursor." (3)
"Accuracy of typing." (4)	"It was uncomfortable to move your wrist to the correct angle for the tilt to register."
"The [keyboard-focus] keyboard was better as it had stability when changing from normal to fading mode." (2)	"Tilting required me to tighten the strap."
"It allowed for more keyboard space on screen." (2)	"Had to tilt wrist unusually for it to fully operate."
"Layout was basic and simple to understand." (2)	"Autocorrect could be better."
"Word predictions were easier to read." (2)	"Tilt angle was in between both screens for my [normal] wrist angle."
"No clutter as use of swipes for space and delete."	"I didn't always realize when I had made mistakes."
"It was good to be able to check the sentence was correct."	"Preferred seeing everything on one screen."
"Use of colours."	"I think it was unnecessary. The [standard] keyboard was good enough."
"Better for longer sentences"	"Sometimes [tilting] was a bit tiring when compared with [standard]."
"Movement felt natural."	"Swipe for space could be a bit more sensitive."

Table 2: Sample Participant Quotes (counts for similar quotes in parenthesis)

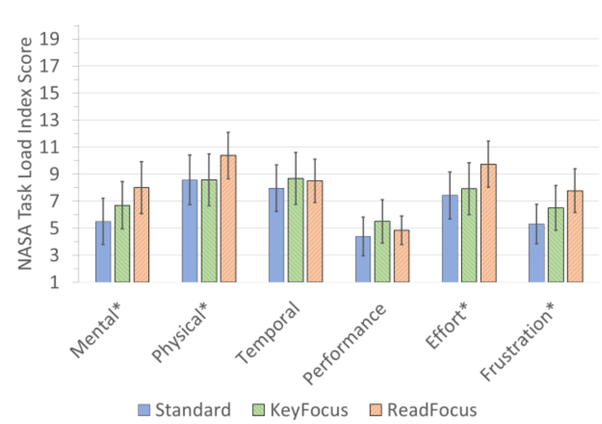

Figure 10: NASA TLX Results

TAP DISTRIBUTION

In line with previous studies of tap distribution in mobile text-entry (e.g. [1]) we were interested to assess how accurate user tapping was on the watch keyboard. Following the results above, we focused on the standard keyboard and excluded the rarest three characters (q,x,z) from main analysis as they did not occur frequently enough in our test data. When analysing individual taps the coordinate tapped has to be associated with the intended letter – this can be problematic as users do not attempt the correct letter all the time (they may miss a key, double tap or simply spell the word wrongly). We associated each tap with the nearest-letter on screen L_n, the expected letter in the phrase L_e and the language-model corrected letter that was inserted into the text L_i. We then filtered out cases where the tap was not correctly corrected by the language model to the expected letter in the target phrase (i.e. where $L_e \neq L_i$). As error rates were low this is a more open approach than a simple distance filter – in practice it excluded 12% of taps that had a mean distance of 87.7 pixels from the expected key.

Overall the average tap offset was 2.3 pixels horizontally and 3.7 vertically showing very close overall proximity to the centre of keys (0.24↔ 0.38↕ mm). Figure 11 shows the average offset per key with a pattern similar to previous analysis (e.g. [1]) for single finger tapping on a smartphone. Letters near the centre of the keyboard are tapped more accurately than the edges and there is a clear downwards trend on the top row. However, the average offsets from the key centres are very small considering that the screen measures 33 mm diameter (approx. 4 mm between key centres).

These average, however, may hide high variance either for each user or between users. Figures 12 shows the individual taps and does suggest fairly large tap variation around each key but still clear focus per letter.

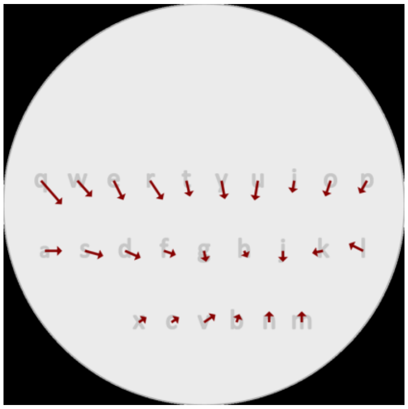

Figure 11: Average offset per key[2]

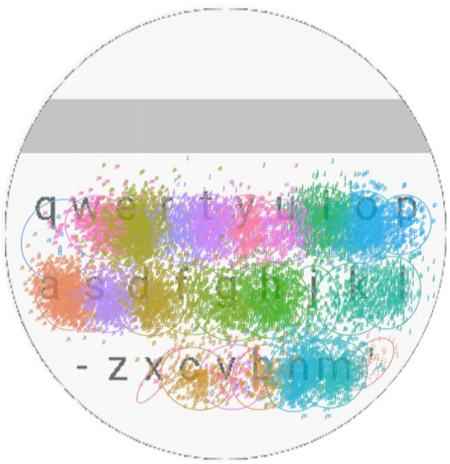

Figure 12: Cloud plot of tap distributions with 95% confidence ellipses[2]

Figures 13-16 show density plots for the keyboard's three horizontal rows and vertically. These show overlap but visible separation of plots between keys.

Finally we analysed the over variance of taps and compared the average variance per user with the average variance between users. Horizontally we see an average standard deviation per user of 12.3 pixels which is very close to the average deviation between users of 11.5 pixels. Similarly we see an average vertical user standard deviation of 11.0 pixels compared to a between user 10.2 pixels. Summarized as 95% confidence ellipse on figure 17, the results show that within user variance is very similar to between user variance on this small screen but that the variance overall is surprisingly small and largely in-line with expectations from large smartphone studies (Azenkot and Zhai [1] found within user standard deviations of 11.3 horizontally and 9.9 vertically on a smartphone with higher between subject variance – despite much smaller physical key separations).

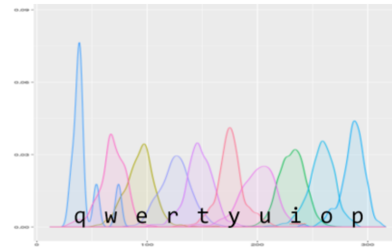

Figure 13: Top row horizontal density plot

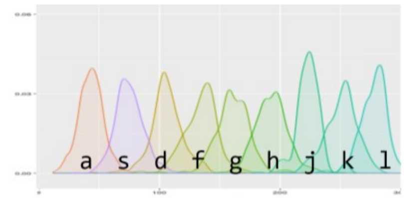

Figure 14: Middle row horizontal density plot

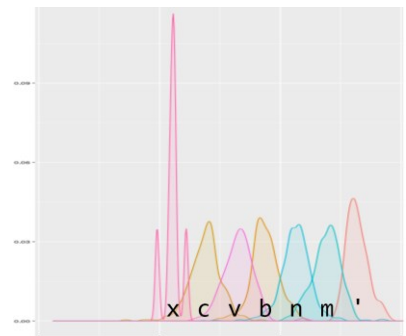

Figure 15: Bottom row horizontal density plot

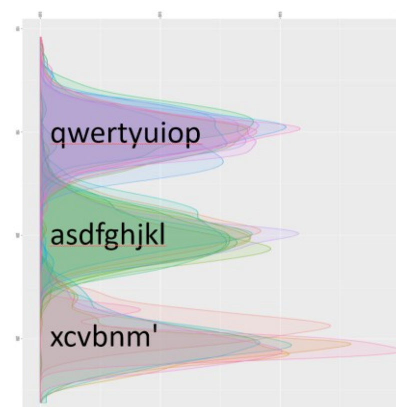

Figure 16: Vertical density plots

DISCUSSION

Reviewing our four hypotheses for study two:

H1 Text entry will be faster using the tilting keyboard

The results contradict this hypothesis – while the difference was not great, the standard keyboard was significantly faster.

[2] Word lists and log data, including all tap data, are available at http://watch.textentry.org.uk/

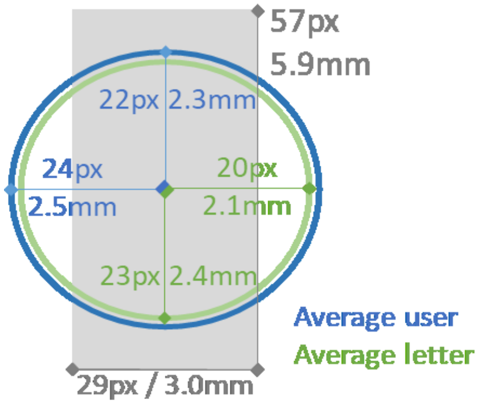

Figure 17: 95% confidence intervals for average tap distance horizontally and vertically (within vs between users) laid over average key dimension

H2 Text entry will be more accurate using the titling keyboard
In all conditions error rates and use of backspace were low. However, again the hypothesis is contradicted with the standard keyboard having lower error rates, despite the keys being closer together. Qualitative feedback, TLX performance ratings, and tap offset/variance analysis confirmed that typing was very accurate overall.

H3 The average angle of typing and variance of viewing angle will differ between the keyboards
The two parts of this hypothesis are confirmed: we see a different average angle of the watch at tap-time for the three conditions. The standard keyboard was closest to the preferred angle identified in study 1 with the two tilt-based keyboards being away from this central position by around 5°. The variation of angle during tasks was lowest for the standard keyboard (albeit with limited significance). There was little difference between the two tilt conditions but the keyboard-focus did appear to have higher variance of angle. In both tilting keyboards the mean angle corresponded to a position that had a partially transparent keyboard. It is interesting to note that even in the final block of phrases, the transparency between the two keyboards (80% and 32% for keyboard-focus and read-focus respectively) did not agree – showing that users weighted a comfortable typing angle higher than visibility optimization. Users also commented on a problem of maintaining uncomfortable angles.

H4 As the study progresses users will tend towards a more transparent keyboard.
This hypothesis is refuted – in the final task set users settled on higher opacity for the keyboard than in the phrase sets overall. Qualitative feedback suggested the transient transparent layering as confusing or distracting.

CONCLUSION AND FUTURE WORK

Tilt control was easily understood by participants in both studies and appreciated as a control method in study 2. However, the layered-tilting keyboard resulted in a slower and more erroneous typing, albeit by a small factor, with negative user comments focusing on the confusing nature of the merged transparent layers and the requirement to maintain an awkward angle. Our conclusion from this study is that users want to see both text and keyboard continuously so settle on a position where this is possible and will not tilt far beyond their comfortable typing angle to achieve this. While using the keyboard for 30 phrases users did not get comfortable enough with key locations to use a near transparent keyboard. As such, we must conclude that while the tilting interface was easy to understand and appreciated by users as a method for easily checking text and flicking between modes – the transient layering was not appreciated and did not help text entry where users preferred to see some of the text and the keyboard continuously. Adaptive monitoring of individual's tilt angles and more complex layering c.f [4] may overcome the visual issues. Tilt to control the display was effective, understood and "natural", with a fairly stable 13-15° tilt angle appearing most comfortable for typing, thus worthy of further research. Care is needed in design of tilt-based interfaces, however, to reduce the need to maintain an angle far from this central "comfort" angle. We suggest occasional flicking to view an alternative display, e.g. the whole text for review and/or carat placement in the case of text entry, may be a more suitable use of tilt-to control than fading proportionally between two superimposed views. Further research is planned into these issues in particular looking into how editing tasks could be supported. The studies here used traditional text-entry transcription tasks in a laboratory environment, we are also interested if different tasks (e.g. prompted composition [42] or image description [33]) or "in the wild" [26] study would affect usability of tilt and give greater challenges to the one short line approach of the standard watch Qwerty layout. We are also interested in looking further into the impact handedness may have on tap accuracy and what impact wearing the watch on the dominant has.

Our study confirmed accurate tapping on the small watch screen and acceptable input speeds (29 words per minute) with very low error rates and backspace usage. Tap analysis showed that users were very accurate on their individual taps and in-line with previous studies relative to key dimensions. Furthermore the use of a tuned language model did succeed in supporting users typing with good "accuracy" while still encouraging accurate tapping.

Speed, accuracy and tap analysis results are encouraging and provide support, with additional experimental data, for previous claims [15,43] that a standard QWERTY layout tied with a strong language model is suitable for text entry on smartwatches.

ACKNOWLEDGMENTS

All user based HCI studies owe a great debt to our participants – thank you! We are also grateful to our colleagues in the Mobiquitous Lab at Strathclyde for their feedback on drafts of this paper and to the anonymous reviewers for their suggestions for improvement.

REFERENCES

1. Shiri Azenkot and Shumin Zhai. 2012. Touch behavior with different postures on soft smartphone keyboards. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, 251–260.
2. JM Barrie. 1991. *Peter Pan, or, The Boy Who Wouldn't Grow Up (Originally 1911)*. Millennium Fulcrum Edition.
3. Patrick Baudisch and Gerry Chu. 2009. Back-of-device interaction allows creating very small touch devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1923–1932.
4. Patrick Baudisch and Carl Gutwin. 2004. Multiblending: displaying overlapping windows simultaneously without the drawbacks of alpha blending. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 367–374.
5. Xiaojun Bi, Barton A. Smith, and Shumin Zhai. 2010. Quasi-qwerty Soft Keyboard Optimization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 283–286. https://doi.org/10.1145/1753326.1753367
6. Xiaojun Bi and Shumin Zhai. 2016. IJQwerty: What Difference Does One Key Change Make? Gesture Typing Keyboard Optimization Bounded by One Key Position Change from Qwerty. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 49–58.
7. Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13, 4: 359–393.
8. James Clawson, Kent Lyons, Alex Rudnick, Robert A. Iannucci,Jr., and Thad Starner. 2008. Automatic Whiteout++: Correcting mini-QWERTY Typing Errors Using Keypress Timing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '08), 573–582.
9. Kiran Dandekar, Balasundar I Raju, and Mandayam A Srinivasan. 2003. 3-D finite-element models of human and monkey fingertips to investigate the mechanics of tactile sense. *Journal of biomechanical engineering* 125, 5: 682–691.
10. Paul A David. 1985. Clio and the Economics of QWERTY. *The American economic review* 75, 2: 332–337.
11. Mark D Dunlop. 2004. Watch-top text-entry: Can phone-style predictive text-entry work with only 5 buttons? In *Mobile Human-Computer Interaction-MobileHCI 2004*, 342–346.
12. Mark D Dunlop and Andrew Crossan. 2000. Predictive text entry methods for mobile phones. *Personal Technologies* 4, 2–3: 134–143.
13. Mark Dunlop and John Levine. 2012. Multidimensional Pareto Optimization of Touchscreen Keyboards for Speed, Familiarity and Improved Spell Checking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 2669–2678.
14. 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* (CHI '12), 2687–2696.
15. Mitchell Gordon, Tom Ouyang, and Shumin Zhai. 2016. WatchWriter: tap and gesture typing on a smartwatch miniature keyboard with statistical decoding. 3817–3821.
16. Beverly L Harrison, Gordon Kurtenbach, and Kim J Vicente. 1995. An experimental evaluation of transparent user interface tools and information content. In *Proceedings of the 8th annual ACM symposium on User interface and software technology*, 81–90.
17. Beverly L Harrison and Kim J Vicente. 1996. An experimental evaluation of transparent menu usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 391–398.
18. Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology*, Peter A. Hancock and Najmedin Meshkati (ed.). North-Holland, 139–183.
19. Jon Hasselgren, Erik Montnemery, Pierre Nugues, and Markus Svensson. 2003. HMS: A Predictive Text Entry Method Using Bigrams. In *Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods* (TextEntry '03), 43–50.
20. Niels Henze, Enrico Rukzio, and Susanne Boll. 2011. 100,000,000 taps: analysis and improvement of touch performance in the large. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 133–142.
21. Niels Henze, Enrico Rukzio, and Susanne Boll. 2012. Observational and Experimental Investigation of Typing Behaviour Using Virtual Keyboards for Mobile Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 2659–2668.
22. Christian Holz and Patrick Baudisch. 2010. The generalized perceived input point model and how to double touch accuracy by extracting fingerprints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 581–590.
23. Christina James and Michael Longé. 2000. Bringing text input beyond the desktop. In *CHI'00 Extended Abstracts on Human Factors in Computing Systems*, 49–50.
24. Eleanor Jones, Jason Alexander, Andreas Andreou, Pourang Irani, and Sriram Subramanian. 2010. GesText: Accelerometer-based Gestural Text-entry Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '10), 2173–2182.
25. Tomonari Kamba, Shawn A Elson, Terry Harpold, Tim Stamper, and Piyawadee Sukaviriya. 1996. Using small screen space more efficiently. In *Proceedings of the*

SIGCHI Conference on Human Factors in Computing Systems, 383–390.
26. Jesper Kjeldskov, Mikael B Skov, Benedikte S Als, and Rune T Høegh. 2004. Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In *Mobile human-computer interaction-MobileHCI 2004*, 61–73. https://doi.org/10.1007/978-3-540-28637-0_6
27. Andreas Komninos and Mark Dunlop. 2014. Text input on a smart watch. *Pervasive Computing, IEEE* 13, 4: 50–58.
28. Per-Ola Kristensson and Shumin Zhai. 2004. SHARK 2: a large vocabulary shorthand writing system for pen-based computers. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*, 43–52.
29. Cliff Kushler. 1998. AAC: Using a Reduced Keyboard. In *Proceedings of the CSUN 98 Conference*.
30. Luis A Leiva, Alireza Sahami, Alejandro Catalá, Niels Henze, and Albrecht Schmidt. 2015. Text Entry on Tiny QWERTY Soft Keyboards. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 669–678.
31. I Scott MacKenzie, Shawn X Zhang, and R William Soukoreff. 1999. Text entry using soft keyboards. *Behaviour & information technology* 18, 4: 235–244.
32. Edith Nesbit. 2004. *The Railway Children*. Shoes & Ships & Sealing Wax.
33. Emma Nicol, Andreas Komninos, and Mark D Dunlop. 2016. A participatory design and formal study investigation into mobile text entry for older adults. *International Journal of Mobile Human Computer Interaction* 8, 2: 20–46.
34. Stephen Oney, Chris Harrison, Amy Ogan, and Jason Wiese. 2013. ZoomBoard: A Diminutive Qwerty Soft Keyboard Using Iterative Zooming for Ultra-small Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 2799–2802.
35. Antti Oulasvirta, Anna Reichel, Wenbin Li, Yan Zhang, Myroslav Bachynskyi, Keith Vertanen, and Per Ola Kristensson. 2013. Improving Two-thumb Text Entry on Touchscreen Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '13), 2765–2774.
36. Kurt Partridge, Saurav Chatterjee, Vibha Sazawal, Gaetano Borriello, and Roy Want. 2002. TiltType: accelerometer-supported text entry for very small devices. In *Proceedings of the 15th annual ACM symposium on User interface software and technology*, 201–204.
37. Beatrix Potter. 2002. *The Tale of Peter Rabbit*. Pioneer Drama Service, Inc.
38. Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 83–88.
39. Jochen Rick. 2010. Performance Optimizations of Virtual Keyboards for Stroke-based Text Entry on a Touch-based Tabletop. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology* (UIST '10), 77–86. https://doi.org/10.1145/1866029.1866043
40. Katie A. Siek, Yvonne Rogers, and Kay H. Connelly. 2005. Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. In *Proceedings of the 2005 IFIP TC13 International Conference on Human-Computer Interaction* (INTERACT'05), 267–280. https://doi.org/10.1007/11555261_24
41. Keith Vertanen and Per Ola Kristensson. 2011. A Versatile Dataset for Text Entry Evaluations Based on Genuine Mobile Emails. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services* (MobileHCI '11), 295–298.
42. Keith Vertanen and Per Ola Kristensson. 2014. Complementing Text Entry Evaluations with a Composition Task. *ACM Trans. Comput.-Hum. Interact.* 21, 2: 8:1–8:33. https://doi.org/10.1145/2555691
43. Keith Vertanen, Haythem Memmi, Justin Emge, Shyam Reyal, and Per Ola Kristensson. 2015. VelociTap: Investigating Fast Mobile Text Entry Using Sentence-Based Decoding of Touchscreen Keyboard Input. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 659–668. https://doi.org/10.1145/2702123.2702135
44. David J Ward, Alan F Blackwell, and David JC MacKay. 2000. Dasher—a data entry interface using continuous gestures and language models. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*, 129–137.
45. Daniel Wigdor and Ravin Balakrishnan. 2004. A Comparison of Consecutive and Concurrent Input Text Entry Techniques for Mobile Phones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '04), 81–88.
46. 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.
47. Shumin Zhai and Per Ola Kristensson. 2012. The Word-gesture Keyboard: Reimagining Keyboard Interaction. *Commun. ACM* 55, 9: 91–101.