Multi-Modal Text Entry and Selection on a Mobile Device

David Dearman
University of Toronto

Amy Karlson
Microsoft Research

Brian Meyers
Microsoft Research

Ben Bederson
University of Maryland

ABSTRACT

Rich text tasks are increasingly common on mobile devices, requiring the user to interleave typing and selection to produce the text and formatting she desires. However, mobile devices are a rich input space where input does not need to be limited to a keyboard and touch. In this paper, we present two complementary studies evaluating four different input modalities to perform selection in support of text entry on a mobile device. The modalities are: screen touch (*Touch*), device tilt (*Tilt*), voice recognition (*Speech*), and foot tap (*Foot*). The results show that Tilt is the fastest method for making a selection, but that Touch allows for the highest overall text throughput. The Tilt and Foot methods—although fast—resulted in users performing and subsequently correcting a high number of text entry errors, whereas the number of errors for Touch is significantly lower. Users experienced significant difficulty when using Tilt and Foot in coordinating the format selections in parallel with the text entry. This difficulty resulted in more errors and therefore lower text throughput. Touching the screen to perform a selection is slower than tilting the device or tapping the foot, but the action of moving the fingers off the keyboard to make a selection ensured high precision when interleaving selection and text entry. Additionally, mobile devices offer a breadth of promising rich input methods that need to be careful studied in situ when deciding if each is appropriate to support a given task; it is not sufficient to study the modalities independent of a natural task.

KEYWORDS: Mobile device, multi-modal, text entry, text formatting, target selection, foot, tilt, touch, speech

INDEX TERMS: H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

1 INTRODUCTION

Text entry is a fundamental and common activity that users perform on their mobile devices. Although the mobile phone is most commonly used to send and receive simple unadorned text messages [7], users are looking for new ways to improve the expressivity their devices can afford. Rich text entry tasks such as writing a detailed and structured email, posting an update to a Blog, and editing a Word document are becoming increasingly more common. These rich text task are typically supported by selectable interface features that allow for faster (*e.g.,* word completion), accurate (*e.g.,* word correction), descriptive (*e.g.,* font, format, colour) and structured (*e.g.,* bullets, indentation) text entry. Selecting these features is currently accomplished by touching an on-screen widget, or navigating among a list of options using the directional pad.

Figure 1. A user entering text on a mobile device and selecting the appropriate character level formatting.

Touching an on-screen widget or manipulating the directional pad are natural methods of selection, but they require users to interleave selection and typing, slowing the user's rate of text input. Given that text entry is already considered a bottleneck for expression on mobile devices, we wondered how well alternative input channels (*e.g.,* accelerometers and speech recognition) might be used to efficiently complement text entry. These alternatives have the potential to be used in parallel with typing, allowing the user to dedicate her fingers to the primary typing task and reduce the impact that editing functions have on throughput.

The focus of this research is to better understand the efficiency potential that alternate input channels hold for increasing the expressivity and throughput of mobile text entry. Specifically, we are interested in comparing screen touching to alternate input channels as a means to support selection during text entry. The input channels we have chosen to evaluate are device tilt (*Tilt*), voice recognition (*Speech*), and foot tap (*Foot*). While there are a large number of alternate input channels that we could explore, we compared Touch with Tilt, Speech, and Foot because they cover a range of input technologies supported by devices today (noting that foot sensing is becoming more common place with the Nike + iPod Sports Kit [9]).

To better understand the relative tradeoffs that screen touch, device tilt, voice input and foot tapping offer users for performing multimodal edit-based selections during text entry, we conducted a controlled laboratory evaluation. Our goal was not to evaluate the technologies themselves, but rather their unique impacts on the flow of text input and formatting. Overall, we found:

- Touch resulted in the highest text throughput. Thus, our core hypothesis that parallel input channels would be faster was false. Coordinating selection and typing was difficult using tilt, foot and voice.
- A significant trade-off exists between selection speed and accuracy. Selection was quickest with device tilt and foot tapping, but screen touch and voice resulted in fewer errors.

dearman@dgp.toronto.edu
{karlson, brianme}@microsoft.com
bederson@cs.umd.edu

- There are interesting human performance issues with respect to the orientation of a target within an input type. For example, tapping the ball of the foot is more accurate than using the heel.
- The time to select a target is slower than the time to resume typing the text.

2 Related Work

The dominant modes for interacting with a mobile device currently require a user to touch the screen or use the directional pad. However, a mobile phone can support alternate input modalities such as accelerometers [3, 4, 10, 11, 15, 19, 20], speech recognition [14, 16, 17, 22], cameras [23] and chording [24] have been explored, some of which are already common in today's smart mobile devices.

Touching the screen with a finger or stylus, manipulating the directional pad and typing on a QWERTY or 12-button keypad are typical modes for interacting with mobile devices. The majority of phones provide one or more of these modes to interact with the information space. ChordTap by Wigdor *et al.,* based on the principles of a chorded keyboard, extends the default keypad by adding three buttons on the back of a mobile phone [24]. The ChordTap keys allow users to differentiate between the multiple characters for each button on a 12 button keypad.

Speech has been used as a means to provide direct input [22] and facilitate text correction [1, 5]. Jiang *et al.* fused word candidates generated by typing Chinese characters while speaking in parallel to generate a single reduced word set [5]. Similarly, Ao *et al.* corrected recognition errors in Chinese handwriting by having people verbally repeat the sentence [1].

Tilt and orientation of a mobile device have been used extensively to navigate through lists and menus [3, 4, 10, 11] and disambiguate between characters when typing [11, 15, 19, 20, 24]. Oakley *et al.* used tilt to navigate through one dimensional lists and menus [10, 11], invoking commands by rotating the device into one of three target regions along a 90 degree rotational space (vertical to horizontal). Unigesture, a tilt-to-write system by Sazawal *et al.,* enabled single handed text entry [20]. Rather than typing on a keypad, three to four characters are organized into seven regions that are selected using the orientation of the device. Partridge *et al*. refined the unigesture technique in TiltType [15], allowing users to disambiguate between the characters within a region by pressing one of four buttons. TiltText, a technique designed specifically for a mobile phone, utilized 30 and 60 degree rotations along the vertical and horizontal plane to disambiguate between the available characters on a standard 12-button mobile phone keypad [24]. Vision TiltText [23] mimicked the functionality of Wigdor *et al.*'s TiltText, but used the mobile phones built-in camera to differentiate between characters by detecting the user's hand movement, rather than the devices tilt.

Foot based input for a mobile device is generally a discounted mode of interaction. However, many professional examples confirm that feet can be elegantly engaged in tasks (*e.g.,* musicians, audio scribes). Pearson and Weiser explored alternate topologies of foot movement and the design space to support interacting with desktop computing [16]. They later implemented one such design called the planar slide mole, assessing its performance against a mouse for target selection [17]. Although the mouse was generally faster and less prone to errors, the foot was extremely quick in gesturing. Pakkanen *et al*. conducted a similar study utilizing a trackball to perform target selection, comparing the foot to the hand [14]. Their results highlight that although the foot is not as dexterous as the hand, it is appropriate for tasks that do not require precision or quick homing and selection. The design of our study is in keeping with these recommendations. The foot is not required to perform homing, only selection.

Many alternate input modalities exist for mobile devices to support direct text input. These methods have never been directly compared in order to assess their effectiveness as a parallel input channel to enrich text entry. Understanding the performance of these alternate input modes will provide a better understanding of how multimodal selection and text entry can be integrated [12] and how to best support user's integration patterns [13].

3 Study

We performed two studies in order to explore human capabilities while performing selections during text entry. Four input modes were compared: standard screen touching (*Touch*), device tilting (*Tilt*), voice recognition (*Speech*), and foot tapping (*Foot*). Tilt, Speech and Foot all allow the user to keep her fingers on the keyboard, and so have the potential to be used independently from the act of letter selection. However, these channels also have characteristics that may impair the text entry task. For example, Speech is considered a "natural" form of input that does not require much in the way of physical effort, but verbalizing commands while typing words may impose additional cognitive demands. Tilt may be familiar to many modern mobile phone or game users, but coordinating Tilt with text entry may be difficult or uncomfortable in practice. Finally, while many professional examples confirm that feet can be elegantly engaged in tasks, Foot selection for the uninitiated user may simply be too awkward and too "distant" from the mental/manual focus of the task to be used effectively.

To evaluate user performance with our chosen input methods, we devised two experiments. Experiment one involved a stimulus-response *Target Selection* task to establish the speed and accuracy with which users can make target selections using each of our four chosen input modes. In experiment two, we used a *Text Formatting* task to evaluate how quickly and accurately users can apply text formatting using each input method while completing a text entry task. The purpose of studying the two tasks independently was to isolate systematic differences in users' abilities to perform selections using the different input methods (*Target Selection*) from other influences affecting the flow and throughput of text entry (*Text Formatting*). Initially we had envisioned using a more common word correction or prediction feature for this task. However, those tasks would not have allowed us to control for *when* the user applied the correction or prediction feature. The usefulness of correction and prediction features is dependent on the input text and the user's perception that selecting a word is faster than correcting or typing the word. Text Formatting was a realistic text entry task that allowed us to maintain control over when and where a selection is made.

The Touch, Tilt, Speech and Foot input methods vary greatly in the granularity of expression they support. For example, voice supports a large unconstrained input space, limited only by the human capacity to label and verbalize a selection. In contrast, researchers have formally characterized the limits of hand tilt to a much smaller input space [18]. Because our focus was on understanding the relative tradeoffs between the input methods during text entry and not comparing the expressive limits of each method, we chose a selection space of four options to achieve parity across the input methods. Limiting the selection space also allowed for straightforward visual mappings between the input gestures and on-screen selection targets.

3.1 Target Selection and Text Formatting Tasks

The *Target Selection* experiment involved a stimulus-response task designed to evaluate the speed and accuracy with which participants could identify and select on-screen targets in four different positions using the four input methods (Figure 2: top row). The goal of this task was to understand users' motor abilities across input methods. The target placement and alignment differed for each input method, but for each method the placement corresponded to the physical action necessary to perform a selection (Figure 2). Each trial began with a blank screen, requiring the participant to press the 'F' and 'J' keys simultaneously to display the target object, shown in red. The start posture was designed to ensure that the device was held in a consistent manner across trials and participants. Selection time was calculated from the time the 'F' and 'J' keys were pressed until a selection was made.

The *Text Formatting* experiment involved a modified text entry task that required participants to reproduce short text phrases that included visual formatting characteristics (Figure 2: bottom row). The goal of this task was to evaluate the speed and accuracy with which users could interleave the selection and de-selection of formatting states while concurrently entering text, and how the four input methods impact the primary text entry task. This is in contrast to the *Target Selection* task, which simply evaluated the user's ability to execute four distinct commands using each of the four input types.

Participants were required to enter the characters of a text phrase and apply formatting to the text at various positions in the phrase. The tasks allowed for partially formatted words, meaning that format mode activations were required both at the beginning and middle of a word, and format mode deactivations were required at both the middle and end of a word. In practice, the format commands were modal—only one format could be active at a time. For example, selecting 'Orange' would activate the orange text mode. All subsequent characters entered would be shown in orange until the user selected 'Orange' again; returning the text mode to "regular" entry mode.

The placement and alignment of the selectable format buttons was identical to that of the *Target Selection* task. The interface supported formatting sequences of three or more characters. Words of three to five characters could be formatted in whole, while words of six or more characters could be formatted in whole or in part. A partially formatted word would always start and end with an unformatted character. Formatting characters according to this schema ensured that a minimum of three characters would be entered between toggling formats. The phrases were chosen randomly from MacKenzie's English phrase dictionary [7]. No phrases were repeated and the order was consistent between participants. After each phrase was entered and properly formatted, the participant pressed the 'Enter' key to advance to the next phrase.

Figure 3. Experimental setup for the Foot input condition.

To conduct this study we developed an application test-bed that provides input to a HTC Touch Pro 2 (shown in Figure 1) using screen Touch, Tilt of the device, Speech, and Foot tapping. The test-bed consisted of two components: a desktop computer running Windows Vista and the HTC Touch Pro 2 running Windows Mobile 6.1. The desktop and mobile device was connected wirelessly by a dedicated Linksys Wireless router using 802.11g.

3.2 Apparatus

Foot and Speech input was accomplished through the desktop computer by wirelessly communicating commands to the mobile device. Input for Foot was performed using two X-keys 3 switch foot pedals connected to the desktop computer via USB. One foot pedal was placed sideways under each foot such that the ball and heel of the foot depressed the respective left and right switch (Figure 3). In the default state, the switches were depressed by the pressure of the user's foot resting on them. A selection was registered when a switch was released, not pressed. This implementation allows for four possible inputs by lifting the ball or heel of the left and right foot. For example, lifting the heel of

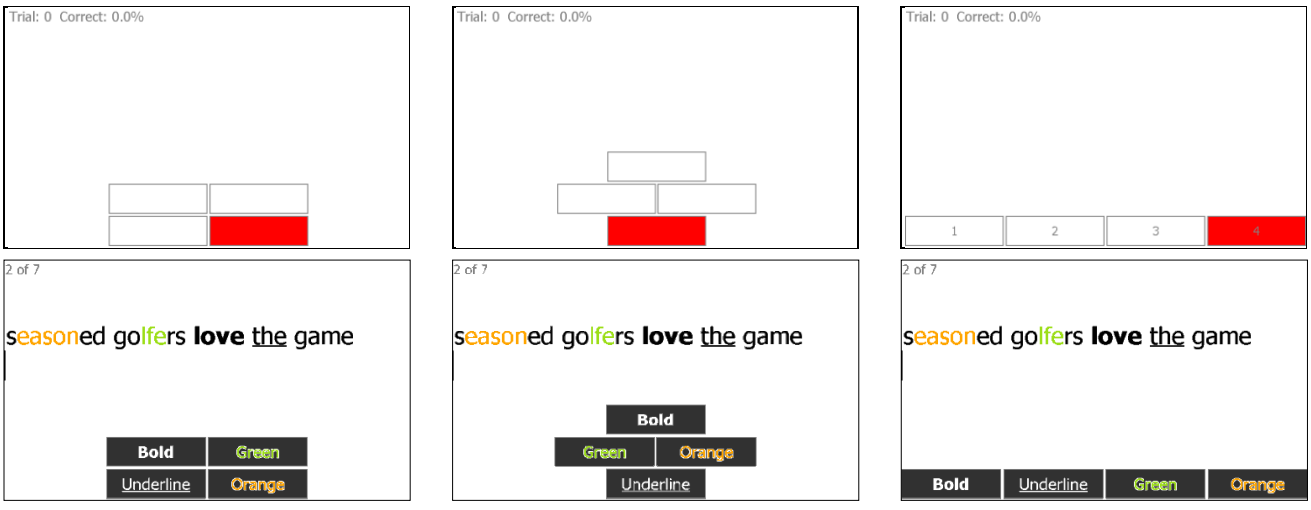

Figure 2. The software interface for the Target Selection (top) and Text Formatting (bottom) tasks. Presented are the Foot (left), Tilt (middle) and Touch/Voice (right) interfaces. The target to select for the Target Selection task is red. The phrase used in the Text Formatting task includes all four formats: 'eason' is orange; 'lfe' is green; 'love' is bold; and 'the' is underlined.

the right foot would select the red target (Figure 2: top-left) or 'Bold' format mode (Figure 2: bottom-left).

Input with the Speech condition was performed by saying the target's label. The speech recognition component of our test bed was implemented using a Wizard of Oz simulation. We chose not use computer-based speech recognition because the systems we tested for the desktop and mobile device incurred a noticeable lag between the time when a label was spoken and interpreted. We wanted the latency between saying a label and it being selected to be as small as possible to allow for a fair comparison. To accomplish this, we relied on a human wizard to listen to the participants' verbal selection and press one of four corresponding keys on a keyboard connected to the desktop computer. The selection was then wirelessly communicated to the mobile device. For example, saying "four" would result in the selection of the red target (Figure 2: top-right). Similarly, saying "Bold" would select the 'Bold' target and enter 'Bold' input mode (Figure 2: bottom-right).

Tilt and Touch inputs were implemented directly on the mobile device and did not require the desktop computer. Input using the Tilt of the mobile device was implemented by sampling the integrated six degree of freedom accelerometer at 25 Hz. We interpreted four Tilt gestures: tilting the device forward, backward, left and right. Gestures exceeding 30 degree changes from a continually updated "neutral" position were recognized as a command along the direction of movement. We chose to implement a relative rather than absolute origin because it allowed users to choose a comfortable position and angle at which to hold the device. For example, tilting the device backward would select the red target (Figure 2: top-middle), or the 'Underline' formatting state (Figure 2: bottom-middle).

Input using Touch was performed by physically pressing the appropriate target displayed on the device's resistive screen. For example, pressing the red target would select it (Figure 2: top-right). Similarly, pressing the 'Bold' target would activate the 'Bold' formatting state (Figure 2: bottom-right).

3.3 Participants

Twenty-four participants recruited from the general population took part in the study – 11 females and 13 males. The age of participants varied between 18 and 38, with a median age of 26. We recruited participants that currently owned a mobile device (e.g., Blackberry, HTC Touch, iPhone or iPod Touch) that they currently use to enter text on a daily basis and have done so consistently for at least the past four months. All participants owned a device with either a physical or touch screen based QWERTY keyboard. All but one participant were right handed. Participants were compensated (removed for review) for their time.

3.4 Procedure

The *Target Selection* experiment was administered first. The participant was first introduced to all four input methods and trained how to use each. The number of training trials varied across participants, but always continued until both the participant and experimenter felt comfortable with the participant's performance. The participant then completed the selection task for each of the four input methods, completing all trials for a given input method before continuing to the next method. Participants were instructed to make the selections as quickly and accurately as possible.

For the *Text Formatting* experiment, the participant was first asked to read a document describing the formatting task and then enter 10 to 12 training phrases for each input method. If the participant felt uncomfortable after the training tasks for any input mode, she was allowed to continue training until both the participant and experimenter felt comfortable with her performance. After training on all input methods, the testing phase began, during which the participant completed four sets of test phrases grouped by input type. Participants were instructed to enter and format the text as quickly and accurately as possible, and to correct all mistakes. However, perfect input was not enforced. After completing trials for each input type, a modified NASA TLX survey was administered. After the final input method was completed, a concluding survey was administered asking participants to rank the inputs in order of preference and to provide subjective feedback with respect to the least and most preferred input.

3.5 Design

The order Tilt, Touch, Speech and Foot were presented was fully counterbalanced across the twenty-four participants for the Target Selection and Text Formatting experiments. The *Target Selection* experiment was a 4×4 design. It comprised the following factors and levels:

- *Input Type* {Touch, Tilt, Speech, Foot}
- *Target Position* {1, 2, 3, 4}

Each *Input Type* was evaluated over six blocks of trials (1 training; 5 testing) with 20 test trials per block - five trials for each of the four *Target Positions*. Each participant performed 4×5×4×5 = 400 test trials or 9,600 among all 24 participants. The presentation order of the target positions was randomly assigned within each block, but consistent across participants. The first block for each *Input Type* was training and excluded from the analysis.

The *Text Formatting* experiment was a 4×3×4 design. It comprised the following factors and levels:

- *Input Type* {Touch, Tilt, Speech, Foot}
- *Format Position* {Start, Middle, End}
- *Target Position* {1, 2, 3, 4}

Each *Input Type* was evaluated over five blocks of trials (one training and four testing) with between 8 and 12 phrases per block. Each block required 48 format selections – four trials for each *Format Position* × *Target Position*. Participants performed 4×4×3×4×4 = 768 format selection and entered 3,111 characters or 18,432 format selections and 74,664 characters among all 24 participants. The length of the words and the number of words per phrase dictated the overall number of phrases required to meet the 48 format selections per block. The presentation order of the format position and the type of format to be applied was randomly assigned within blocks, but presented consistently across participants. The first block for each *Input Type* was training and excluded from the analysis.

4 Results

The *Target Selection* and *Text Formatting* experiments were conducted independently, and thus analyzed separately. Selection time trials that exceeded three standard deviations from the mean were removed as outliers. To account for the variability in human selection, the median selection time for each participant was used in the analysis. Timing data was analyzed with repeated measures ANOVAs using Wilk's Lambda. Event-count measures such as error were analyzed with nonparametric Friedman tests and post-hoc pairwise comparisons were conducted with the Wilcoxon test. All post-hoc comparisons were conducted using Holm's sequential Bonferroni correction.

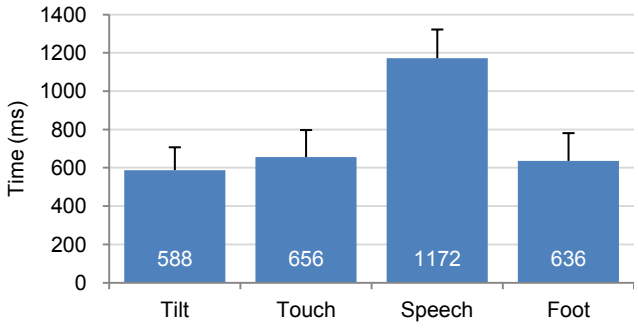

Figure 4. The average selection time for the Target Selection experiment grouped by the Input Types. The error bars indicate the standard deviation.

In addition to evaluating the participants' performance for each *Input Type*, we evaluated how each input impacted the task of text entry by analyzing the participant's character stream with Wobbrock and Myers TextTest StreamAnalyzer [25].

We do not directly compare the *Target Positions* across the four *Input Types* because the positions vary and are therefore not equivalent. Rather, we focus our analysis on the selection time and error between *Target Positions* within *Input Types*.

4.1 Target Selection Results

Seventy-six outliers (0.8%)—three standard deviations from the mean—were excluded from the analysis. Analysis of the blocks yielded no evidence of a learning effect; therefore all test trials are included in the time and error analyses.

4.1.1 Target Selection Time

Repeated measures ANOVA on the median selection times yielded a significant effect of *Input Type*, $F_{3, 21}$=879.98, p<0.001. Post-hoc pairwise comparisons show that the overall selection time for Speech (1172 ms) was slower than the other *Input Types* (Figure 4), all at p<0.001. Although Touch (656 ms) and Foot (635 ms) have similar mean selection times, Tilt (588 ms) was found to be faster than the other *Input Types*, all at p<0.001.

Repeated measures ANOVAs for *Target Positions* within *Input Types* found a significant effect on selection time for Tilt ($F_{3, 21}$=8.64, p<0.001), Touch ($F_{3, 21}$=13.80, p<0.001) and Speech ($F_{3, 21}$=11.10, p<0.001), but not for Foot.

Tilt. Tilting the device forward (561 ms) was faster than tilting in the other directions: forward was 8.2% faster than left (594 ms; p<0.05), 5.9% faster than right (591 ms; p<0.001), and 5.3% faster than backward (607 ms; p<0.01).

Touch. The second target (680 ms; ordered left to right) was 4.0% slower than the first (654 ms; p<0.001), and 7.1% slower than the fourth (635ms; p<0.001). While somewhat surprising, the slower selection time for the second target is likely attributed to the majority of our participants being right-hand dominant. This agrees with our observation that many participants opted to use their right thumb to reach across the screen to the second target rather than using their proximally closer left thumb.

Speech. The fourth target (1199ms; ordered left to right) was 3.5% slower than the first (1158ms; p<0.001), and 3.9% slower than the second (1154ms; p<0.001). While the slower selection time of the fourth target might be attributed to a tendency for participants to scan left to right, this explanation seems somewhat suspect since we would have expected target recognition to occur preattentively and be immune to position bias. Furthermore, it is useful to remember that the Speech condition involved two independent human response components (participant and wizard). Together with the fact that overall differences in selection times across positions was very small (≤ 41 ms) it is unlikely that the position effect has practical meaning with respect to user performance for Speech.

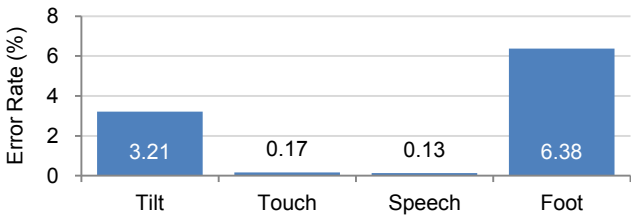

Figure 5. The selection error rate for the Target Selection experiment grouped by the Input Types.

4.1.2 Target Selection Errors

The overall selection error rate was 2.47%. A Friedman test showed a significant main effect of *Input Type*, $\chi^2_{(3, N=24)}$ = 55.29, p<0.001. The error rates for Touch (0.17%) and Speech (0.13%) are negligible (<1%) and significantly lower than Tilt (3.21%) and Foot (6.38%), all with p<0.001. In addition, the error rate for Foot is greater than Tilt, p<0.005.

Pairwise analysis of the *Target Positions* within the *Input Types* yielded significant differences in the error rate for Tilt ($\chi^2_{(3, N=24)}$ = 7.88, p<0.05) and Foot ($\chi^2_{(3, N=24)}$ = 8.52, p<0.05). Post-hoc pairwise comparisons were conducted with the Wilcoxon test.

Tilt. Forward tilt (n=29; 1.21%) resulted in a higher error rate than backwards (n=10; 0.04%), p<0.005. No differences were found for left (n=18; 0.75%) and right (n=20; 0.83%).

Foot. The error rate for the left-heel (n=52; 2.17%) was greater than the right-ball (n=27; 1.13%), p<0.005. No differences were found for the left-ball (n=32; 1.33%) or right-heel (n=42; 1.75%). We also compared the combined error rate of the ball and heel of the left and right foot. Overall, the heel (n=94; 3.92%) has an error rate greater than the ball of the foot (n=59; 2.46%), p<0.05. Since most of our participants were right-footed, it makes sense that participants would be most agile with the ball of their dominant foot (right) and least agile with the heel of their non-dominant foot (left).

4.2 Text Formatting Results

Two-hundred and ten (1.1%) selections were identified as outliers (greater than three standard deviations from the mean) and removed from the analysis. The timing data for P1 using Touch (192 entries) was not included in the analysis because of a software logging error. In the analysis we differentiate:

- *Selection Time* – the time difference between typing a character and selecting a subsequent formatting.
- *Resumption Time* – the time difference between selecting a format and typing a subsequent character.

Pairwise comparison shows that selection time is slower, p<0.001, than resumption time (Figures 6 and 7).

4.2.1 Format Selection Time

Repeated measures ANOVA of the median selection time yielded a significant effect of *Input Type*, $F_{3, 20}$ = 95.23, p<0.001, and *Format Position*, $F_{2, 21}$ = 15.0, p<0.001. Similar to the *Target Selection* results, post-hoc pairwise comparisons show that the selection time for Speech (1146 ms) was slower than the other *Input Types* (Figure 6), all at p<0.001. Although Touch (855 ms), Tilt (797 ms) and Foot (834 ms) have similar mean selection times, Tilt was found to be faster than Touch, p<0.001. Analysis

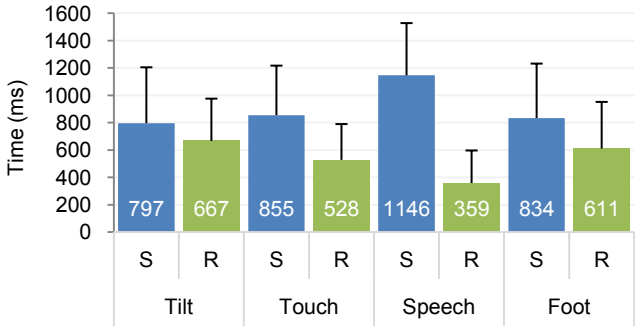

Figure 6. The average selection time (S) and resumption time (R) for the Text Formatting experiment grouped by the Input Types. The error bars indicate the standard deviation.

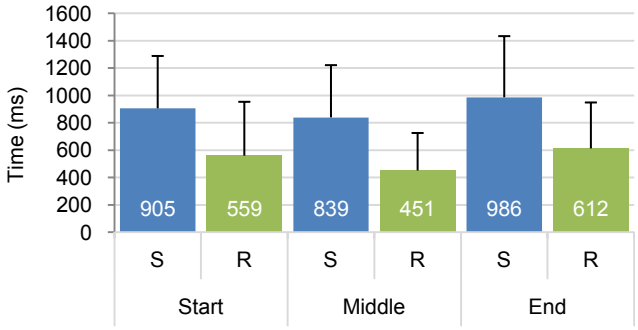

Figure 7. The average selection time (S) and resumption time (R) for the Text Formatting experiment grouped by the Format Positions. The error bars indicate the standard deviation.

of the *Format Position* revealed that toggling a format selection at the End of a word (839 ms) is faster than the Start (905ms; $p<0.01$) and Middle of a word (985 ms; $p<0.001$).

Repeated measures ANOVAs for *Target Position* within *Input Types* found a significant main effect on selection time for Touch ($F_{3, 19}=11.04$, $p<0.001$), Speech ($F_{3, 19}=11.62$, $p<0.001$), and Foot ($F_{3, 18}= 7.30$, $p<0.005$), but not Tilt.

Touch. The second target (920 ms) is 13.6% slower than the third (810ms) and 12.1% slower than the fourth (821 ms), all at $p<0.001$, which matches our findings from the Target Selection study.

Speech. The second target (1192 ms) is slower than all the other target positions: 5.7% slower than the first (1128 ms; $p<0.005$); 7.6% slower than the third (1108 ms; $p<0.001$); and 3.3% slower than the fourth (1154 ms; $p<0.05$).

Foot. The left-heel (903 ms) is 14.0% slower than the right-ball (792 ms; $p<0.005$) and 11.3% slower than the right-heel (811 ms; $p<0.001$), but not the left-ball (834 ms).

4.2.2 Format Resumption Time

Repeated measures ANOVAs on the median resumption times yielded a significant effect of *Input Type*, $F_{3, 20}=22.27$, $p<0.001$, and *Format Position*, $F_{2, 21}=80.10$, $p<0.001$. In contrast to the selection time results where Speech was the slowest *Input Type*, Speech (359 ms) was the fastest *Input Type* for resumption, all at $p<0.001$. Touch (528 ms), Tilt (667 ms) and Foot (611 ms) have comparable mean resumption times, but Touch was faster than Tilt, $p<0.001$. Analysis of the *Format Position* revealed that toggling a format at the End of a word (451 ms) is faster than the Start (559 ms) and Middle of a word (611 ms), all at $p<0.001$. In addition, toggling a format at the Start of a word is faster than the Middle, $p<0.001$.

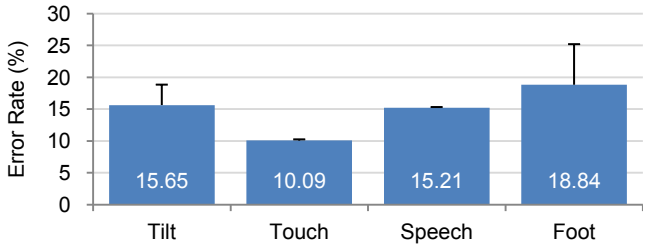

Figure 8. The selection error rate for the Text Formatting experiment grouped by the Input Types.

Repeated measures ANOVAs for *Target Position* within *Input Types* found a significant main effect of resumption time for Touch ($F_{3, 19}=7.89$, $p<0.001$.), but not Tilt, Speech or Foot.

Touch. The first target (502 ms) was faster than the other target positions: 5.2% faster than the second (529 ms; $p<0.05$); 8.0% faster than the third (543 ms; $p<0.001$); and 7.0% faster than the fourth (538 ms; $p<0.005$).

4.2.3 Format Selection Errors

The overall error rate was 14.95%. A Friedman test showed a significant main effect of *Input Type*, $\chi^2_{(3, N=23)}=27.05$, $p<0.001$, but not the *Format Positions*. Post-hoc pairwise comparisons showed that Touch incurs fewer errors than the other *Input Types*, all at $p<0.001$.

Analyses of *Target Position* within *Input Types* yielded significant differences in error rate for Speech ($\chi^2_{(3, N=23)}=10.08$, $p<0.005$) and Foot ($\chi^2_{(3, N=24)} = 8.52$, $p<0.05$), but not Touch and Tilt. Post-hoc pairwise comparisons were conducted with the Wilcoxon test.

Speech. The error rate for the second target ($n=154$; 3.34%) was lower than the third target ($n=206$; 4.47%), $p<0.005$, but not the first ($n=171$; 3.71%) or fourth ($n=170$; 3.69%).

Foot. The error rate for the left-ball ($n=262$; 5.69%) was greater than the right-ball ($n=207$; 4.49%), $p<0.001$, and right-heel ($n=185$; 4.01%), $p<0.001$, but not the left-heel ($n=214$; 4.64%).

4.2.4 Text Throughput (Characters per Second)

Across all participants, the average character per second (CPS) text throughput was 1.36 (see Table 1). The CPS reported here (1.36) is much lower than the mini-QWERTY CPS (2.65) reported in prior research [2] because of the additional formatting requirements. There is a significant main effect of *Input Type* on CPS, $F_{3, 21}=19.06$, $p<0.001$. Post-hoc pairwise comparison of the CPS shows that *Touch* resulted in a 9.8% greater throughput than Tilt, $p<0.001$ and 10.7% greater throughput than Foot, $p<0.001$.

4.2.5 Format Errors and Text Corrections

Formatting the phrases correctly required participants to select the four *Target Positions* a total of 4,608 times for each *Input Type*.

Table 1. For each Input Type, the difference between the number of formats required and the actual number of formats used (FEC), backspace count (BS), characters per second (CPS), uncorrected (UER) and corrected error rate (CER). The CPS, UER and CER are calculated using Wobbrock and Myers StreamAnalyzer [25].

	FEC (N)	*BS (N)*	*CPS (N/s)*	*UER (N/s)*	*CER (N/s)*
Tilt	839	1062	1.32	0.0055	0.0522
Touch	219	1048	1.45	0.0033	0.0506
Speech	184	1619	1.37	0.0037	0.0770
Foot	1320	1451	1.31	0.0019	0.0702

Table 1 shows the total number of formatting errors (FEC), the total number of backspaces (BS), and the uncorrected (UER) and corrected (CER) error rates. Note that UER and CER only capture character level errors, and not format-level errors. Each measure gives insight into how *Input Type* impacted the overall text throughput. We observed a significant main effect of *Input Type* for FEC, $\chi^2_{(3,\ N=24)}=57.85$, $p<0.001$, backspaces, $\chi^2_{(3,\ N=24)}=20.53$, $p<0.001$ and CER, $\chi^2_{(3,\ N=24)}=25.65$, $p<0.001$, but not UER.

Post-hoc pairwise comparisons show that the Foot has a greater FEC than Tilt, $p<0.05$, Touch, $p<0.001$, and Speech, $p<0.001$; and that Tilt has a greater FEC than Touch and Speech, both $p<0.001$. In contrast, the number of backspaces used with Tilt and Touch is lower than Speech (both $p<0.001$) and Foot ($p<0.05$ and $p<0.01$ respectively). Similarly, the CER for Tilt and Touch is lower than Foot and Speech, both $p<0.001$. The higher error rate of Foot and Tilt was likely due to in-place formatting mistakes. The higher correction rate of Speech and Foot suggest that users made more errors placing the formats relative to the text. This is what we would expect if users had trouble coordinating the more asynchronous inputs of Speech and Foot entry with text input.

4.3 Qualitative Responses

The results of the modified NASA TLX surveys and the ranked preferences of the *Input Types* are shown in Figure 9. Participants rated each *Input Type* using the NASA TLX measures on a 7-point Likert scale (1='Very Low'; 7='Very High'). The subjective measures were: mental demand (*Mental*), physical demand (*Physical*), task success (*Success*), speed of use (*Speed*), ease of use (*Ease*) and ease of learnability (*Learning*).

The participants responses revealed a significant difference in their ranked preference for the *Input Types,* $\chi^2_{(3,\ N=24)} = 14.9$, $p<0.005$. Post-hoc pairwise comparison showed a preference for Touch over Foot, $p<0.001$. The majority of participants ranked Touch (11) and Speech (9) as their preferred input, but a small number of participants showed a strong preference for Tilt (3) and Foot (1).

Friedman tests of the NASA TLX measures showed a main effect of *Input Type* for: mental demand, $\chi^2_{(3,\ N=24)}=22.31$, $p<0.001$; physical demand, $\chi^2_{(3,\ N=24)}=30.16$, $p<0.001$; task success, $\chi^2_{(3,\ N=24)}=17.31$, $p<0.001$; speed of use, $\chi^2_{(3,\ N=24)}=13.48$, $p<0.005$; ease of use, $\chi^2_{(3,\ N=24)}=20.00$, $p<0.001$; and ease of learnability, $\chi^2_{(3,\ N=24)}=19.26$, $p<0.001$. Touch and Speech were perceived as more successful than Tilt (both $p<0.005$) and Foot ($p<0.005$ and $p<0.05$ respectively), and as easier to use than Tilt ($p<0.005$ and $p<0.01$) and Foot ($p<0.001$ and $p<0.005$). Speech was easier to learn than Foot, $p<0.005$, and Foot was more mentally and physically demanding to use than Tilt ($p<0.005$ and $p<0.05$), Touch (both $p<0.001$), and Speech ($p<0.005$ and $p<0.001$). Although Tilt was no more mentally demanding than the other *Input Types*, it was more physically demanding to use than Touch, $p<0.01$ and Speech, $p<0.001$. Overall, participants correctly felt that they were faster with Touch than Foot, $p<0.005$.

5 DISCUSSION

Examination of selection speed, independent of the text input, revealed that selecting targets was quickest using Tilt, rather than Touch, Speech and Foot. However, with respect to the text formatting task, participants experienced the greatest text throughput (characters per second entry) using Touch and Speech – the two slowest *Input Types* for performing selections. While these results are seemingly at odds with one another, the impacted text throughput for Tilt and Foot is partially due to their higher error rate. Tilt and Foot allowed for the fastest selections, but resulted in the greatest number of errors. In contrast, Touch and Speech

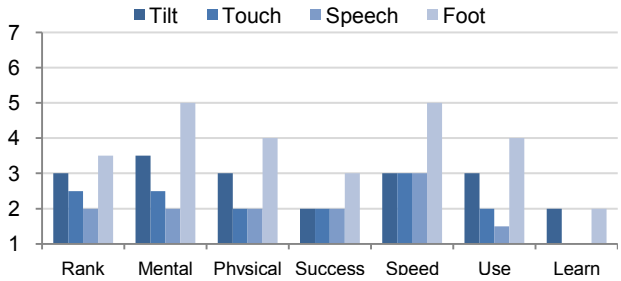

Figure 9. The participants responses (median values) to their ranked preference and the modified NASA TLX survey for each input type. Lower values are better.

resulted in the slowest selection times, but the lowest number of errors. The time required to correct the erroneous selections and any improperly formatted characters accumulated, impacting the overall text throughput for Tilt and Foot. Although we trained participants until both the participant and experimenter felt comfortable with the participant's performance, some participants commented that Tilt and Foot were "*awkward to coordinate effectively.*" The common sentiment among participants was that Touch and Speech were "*the most natural*", making them easier to use and less mentally/physically demanding.

It was our core hypothesis for the Text Formatting experiment that Tilt, Speech and Foot would result in the highest test throughput. We believed that the ability to make format selections and enter text in parallel would improve overall text throughput. However, contrary to our initial expectations, Touch resulted in the highest text throughput. The sequential ordering of selection and text entry allowed Touch to strike a balance between speed and a low rate of error. For Speech and Foot, it was difficult to coordinate selection in parallel with typing, resulting in a higher per character error rate than Touch: "*I tend to format a split second before I start typing … I'd type a letter or two before the formatting would take place … I would find myself typing the word that was supposed to be green … before saying green.*" While Tilt and Foot "*seemed easy to use because you could keep your fingers on the keyboard"* and Touch "*felt like it was slowing me down having to move my thumbs up [to select targets]*", it was in fact more difficult to make an accurate selection with Tilt and Foot, resulting in a higher format error rate than Touch. Moving the thumbs off the keyboard to make a selection, although perceived as slowing down the text entry, ensured participants did not have to coordinate parallel selection and typing. Depending on the granularity required for a text application, the concurrency problem with Tilt, Speech and Foot can be corrected through simple modifications to the interface. Rather than requiring a format selection to be activated at a precise position within the word (*e.g.,* beginning), the format could be activated by the user at any point during word entry and applied to the entire word. If more precision is needed, the format could be applied "back in time" to the characters that were typed when, for example, the utterance was started, rather than after it was completed and recognized.

The majority of participants preferred Touch and Speech, but a small minority preferred Tilt and Foot. We believe that the perceptions of the Foot modality may have been negatively impacted by our use of foot pedals to register the selections. One participant who did not rank Foot as their preferred input commented that, "*I think I performed best with foot, but I'm on a cell phone, why would I use my feet? If it was integrated into my shoe, feet would be number one.*" We explained to participants in the concluding interview that a more realistic implementation would

integrate within actual footwear such as the Nike + iPod Sport Kit [9]. Sensing integrated within a user's own footwear may perform better and be perceived as more useful.

Alternate target placements for each input type highlighted interesting design considerations with respect to the target orientations and input. Handedness and footedness is important when choosing target placement. The left to right ordering for touch resulted in the second button being slower than the third and fourth. It was most frequently the case that participants would use their right thumb to select all buttons except the first one: "*I kept trying to figure out which thumb to use [to select the second button]. It's not like my left thumb was tired, but I just kept crossing all the way over with my right thumb*" Selection using the heel resulted in more errors than the ball of the foot, "*it's easier to trigger with my toes than my heels.*"

6 CONCLUSION

Mobile devices support many rich text entry tasks that require a user to interleave typing and selection to produce the text and formatting she desires. For example, text completion, correction and formatting functions improve the speed of text entry and enrich the presentation of text, but currently require a user perform a selection (via screen touch or the directional pad) while typing. In this paper we performed two complementary studies to explore the performance and limitations of using standard screen touch (*Touch*), device tilt (*Tilt*), voice recognition (*Speech*), and foot tap (*Foot*) to perform selections in support of text entry. The results show that Tilt is fastest for selecting a target, but that selection with Touch while entering text allows for the greatest character per second text throughput. Coordinating selection with Tilt, Speech and Foot in parallel with typing proved to be difficult, resulting in a higher per character error rate than Touch. Although perceived as slow—and slower than Tilt for selecting targets—moving the thumbs off the keyboard to make a selection ensured a natural interleaving of selection and typing; producing a low rate of error, resulting in the need for fewer corrections producing the highest text throughput.

Additionally, we highlighted human performance issues with respect to the orientation of the targets within an input type for handedness and footedness. The left to right ordering for Touch resulted in the second button being slower than the third and fourth because participants would use their right thumb to reach across and select all except the first button. Similarly, tapping the ball of the foot is more accurate than using the heel.

7 REFERENCES

[1] Ao, X., Wang, X, Tian, F., Dai, G. and Wang, H. Crossmodal error correction of continuous handwriting recognition by speech. In *Proc. IUI 2007*, ACM Press (2007), 243-250.

[2] Clarkson, E., Clawson, J., Lyons, K. and Starner, T. An empirical study of typing rates on mini-QWERTY keyboards. *Ext. Abstracts CHI 2005*, ACM Press (2005), 1288-1291.

[3] Harrison, B.L., Fishkin, K.P., Gujar, A., Mochon, C. and Want, R. Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces. In *Proc. CHI 1998*, ACM Press (1998), 17-24.

[4] Hinckley, K., Pierce, J., Sinclair, M. and Horvitz, E. Sensing techniques for mobile interaction. *In Proc. USIT 2000*, ACM Press (2000), 91-100.

[5] Jiang, Y., Wang, X., Tian, F., Ao, X., Dai, G. and Wang, H. Multimodal Chinese text entry with speech and keypad on mobile devices. In *Proc. IUI 2008*, ACM Press (2008), 341-344.

[6] Kume, Y. Foot interface: fantastic phantom slipper. *Ext. Abstracts SIGGRAPH* 1998, ACM Press (1998), 114.

[7] MacKenzie, I.S. and Soukoreff, R.W. Phrase sets for evaluating text entry techniques. *Ext. Abstracts CHI 2003*, ACM Press (2003), 754-755.

[8] Nielsen. Flying Fingers: Text-messaging overtakes monthly phone calls, http://en-us.nielsen.com/main/insights/consumer_insight/issue_12/flying_fingers

[9] Nike + iPod Sport Kit, http://www.apple.com/ipod/nike/

[10] Oakley, I. and O'Modhrain, S. Tilt to Scroll: Evaluating a motion based vibrotactile mobile interface. In *Proc. WHC 2005*, IEEE Computer Society (2005), 40-49.

[11] Oakley, I. and Park, J. A motion-based marking menu system. *Ext. Abstracts CHI 2007*, ACM Press (2007), 2597-2602.

[12] Oviatt, S. Ten myths of multimodal interaction. *Commun. ACM 42*, 11 (1999), 74-81.

[13] Oviatt, S., Lunsford, R. and Coulston, R. Individual differences in multimodal integration patterns: what are they and why do they exist? In *Proc. CHI 2005*, ACM Press (2005), 241-249.

[14] Pakkanen, T. and Raisamo, R. Appropriateness of foot interaction for non-accurate spatial tasks. *Ext. Abstracts CHI 2004*, ACM Press (2004), 1123-1126.

[15] Partridge, K., Chatterjee, S., Sazawal, V., Borriello, G. and Want, R. TiltType: accelerometer-supported text entry for very small devices. In *Proc. UIST 2002*, ACM Press (2002), 201-204.

[16] Pearson, G. and Weiser, M. Of moles and men: the design of foot controls for workstations. In *Proc. CHI 1986*, ACM Press (1986), 333-339.

[17] Pearson, G. and Weiser, M. Exploratory evaluation of a planar foot-operated cursor-positioning device. In *Proc. CHI 1988*, ACM Press (1988), 13-18.

[18] Rahman, M., Gustafson, S., Irani, P. and Subramanian, S. Tilt techniques: investigating the dexterity of wrist-based input. In *Proc. CHI 2009*, ACM Press (2009), 1943-1952.

[19] Rekimoto, J. Tilting operations for small screen interfaces. In *Proc. UIST 1996*, ACM Press (1996), 167-168.

[20] Sazawal, V., Want, R. and Borriello, G. The unigesture approach. In *Proc. Mobile HCI 2002*, Springer-Verlag (200), 256-270.

[21] Schoning, J., Daiber, F., Kruger, A. and Rohs, M. Using hands and feet to navigate and manipulate spatial data. *Ext. Abstracts CHI 2009*, ACM Press (2009), 4663-4668.

[22] Voice command for Windows Mobile, http://www.microsoft.com/windowsmobile/en-us/downloads/microsoft/about-voice-command.mspx

[23] Wang, J., Zhai, S. and Canny, J. Camera phone based motion sensing: interaction techniques, applications and performance study. In *Proc. UIST 2006*, ACM Press (2006), 101-110.

[24] Wigdor, D. and Balakrishnan, R. A comparison of consecutive and concurrent input text entry techniques for mobile phones. In *Proc.* CHI 2004, ACM Press (2004), 81-88.

[25] Wobbrock, J.O. and Myers, B.A. Analyzing the input stream for character-level errors in unconstrained text entry evaluations. *ACM Trans. Comput.-Hum. Interact. 13*, 4 (Dec. 2006), 458-489.