Vouch-T: Multimodal Text Input for Mobile Devices Using Voice and Touch

Minyoung Lee and Gerard J. Kim(✉)

Digital Experience Laboratory, Korea University, Seoul, Korea
{mignon0o0b, gjkim}@korea.ac.kr

Abstract. Entering text on a small mobile device is not easy due to the relatively small screen space and the fat finger problem. We consider a multimodal text input method, called Vouch-T (Voice + tOUCH - Text) combining the touch and voice input in complementary way. With Vouch-T, the user makes an approximate touch among densely distributed alphabetic keys, accompanied with the voice input to effectively disambiguate the target among possible candidates. We have assessed the potential of Vouch-T in terms of the usability and performance compared to the conventional touch-only based method. We considered two types of text input layout, namely, the QWERTY and 3 × 4 telephone keypad on two sizes of mobile devices, the smart phone and smart watch. The comparative simulation experiment validated that the multimodal approach of Vouch-T improves the input performance and usability for the smaller-sized smart watch, but only marginally for the larger smart phone.

Keywords: Keyboard · Text input · Multimodal input · Voice · Smart watch · Smart phone · Usability · Task performance

1 Introduction

Small multi-purpose mobile devices, such as the smart phone and smart watch, often require text input, albeit mostly simple and comprising of just several words or phrases such as in short text messages, entering phone numbers and contact and making daily schedules. Despite such a need, text entry on small mobile devices remain to be not so easy, e.g. on smart phones. The situation is obviously worse with the even smaller smart watches. While in many cases, smart watches are used in a supplementary way to the smart phone where the text entry is made instead. But it is also true that smart watches are gaining popularity as an independent unit, sometimes over-taking the usual roles (e.g. including those as related to text entry) of the smart phones.

Therefore, while it is not likely smart watches will be replacing smart phones, there are still definite merits to improving the text input capabilities of smart watches so that it can be increasingly used independently from a mother device like the smart phone. In fact, various proposals have been made and techniques developed for facilitating simple text/command input for small screened smart watches, using touch gestures, hand/finger written letter recognition, multi-layer keyboard, distributed icon layout, voice recognition, and etc. [4, 7, 8, 11, 14, 15, 26, 27].

M. Kurosu (Ed.): HCI 2017, Part II, LNCS 10272, pp. 208–224, 2017.
DOI: 10.1007/978-3-319-58077-7_17

In the light of this situation, we consider a multimodal input method as an attractive and classic solution to improving interaction performance and user satisfaction, particularly even more effective under difficult operating conditions [17]. We propose to combine two modalities, namely, touch and voice. Touch input has high familiarity and good ergonomic accessibility, but limited by the fat finger problem (or equivalently, the screen size) [11, 22] and sensitive to user motion [1]. Voice input is mostly immune to user motion, but its reliability may suffer by environment noise. However, such respective characteristics can complement each other very well in coping with the difficult operating condition of smart watch. The pioneering work of "Put-that-there" [3] had illustrated how two lesser reliable modality input methods can complement each other to boost the overall recognition performance.

In our case, the user can attempt to touch among densely distributed small icons or keys and while difficult to pinpoint at a target, the finger position will at least be in the vicinity of the wanted key or icon. The accompanying voice input can be used effectively (despite some environment noise) to disambiguate the target among just a small number of possible candidates, if not recognize the target right out (see Fig. 1). Such a process has the added benefit of relieving the burden of the user having to

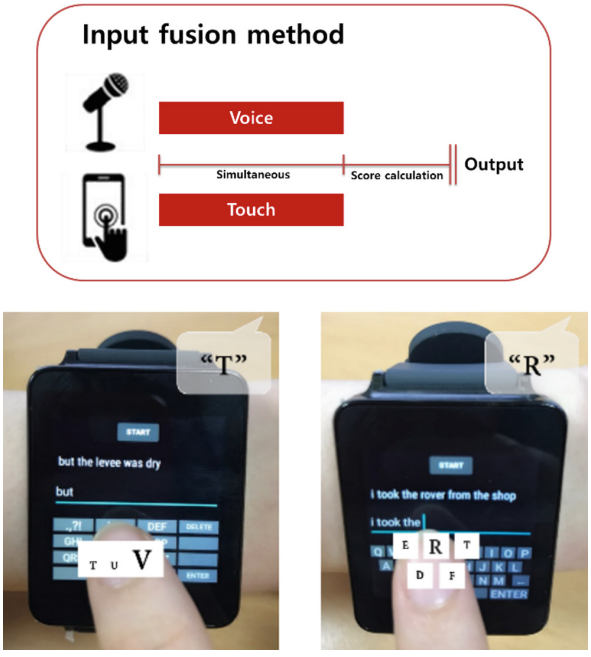

Fig. 1. The interaction model of Vouch-T for text entry on small mobile devices: (1) 3 × 4 keypad: the user touches a particular cell containing the target letter, and a second menu containing three possible letters appears, chosen by a near simultaneous voice input (left), and (2) QWERTY: the user touches a particular target letter on the layout, which is disambiguated by a near simultaneous voice input (right).

consciously pinpoint to the already small icons/keys. We call this approach the "Vouch-T" (Voice + tOUCH for Text input)[1].

We assess the potential of Vouch-T in terms of the usability and performance compared to the conventional touch-only based method through a simulation experiment. In the study, two types of text input layout and device size are considered respectively, namely, the QWERTY and 3 × 4 telephone keypad layout on the smart phone and even smaller smart watch.

2 Related Work

Even before smart watches were commercially available and became popular, several input methods for small media devices have been studied [1, 2, 19]. Albinsson et al. [1] introduced the Zoom-pointing in which the user first enlarges the interaction space (nearby the target) enabling easier selection of small objects. Roudaut et al. [19] developed similar methods called the TapTap and Magstick. Baudisch et al. [2] proposed to use the back of the device (although not applicable for smart watches) in order to avoid the occlusion by the fingers in interacting with small screened devices.

Although similar in terms of being a key selection task, text input on smart watches present a more difficult situation than object/icon selection because there are many possible keys and the input has to be made continuously [21]. Consequently, methods more specific to text entry have been devised [4–8, 10, 12, 14, 17]. For example, Oney et al. [17] have applied the Zoom pointing to QWERTY keyboard input for smart watches. The SplitBoard [12] and Virtual Sliding [4] divide the keyboard layout into several parts (so that each of them are presented in a large enough size for fast selection of keys) which are accessible by flicking through them. Dunlop et al. [8] suggested for a layout with just six large buttons, three each on the top and bottom rows of the watch screen (middle row used for showing the text entry). Each large key (of the six) maped to 3–6 letters (with the most likely one entered as the first choice) covering all alphabets. The layout could be switched to that for numbers or special characters by flicks. Combined with the auto-completion of words, this method significantly reduced the number of touches compared to the Splitboard or Virtual Sliding. Chen et al. [5] developed the SwipeBoard that used two swipes in a continuous fashion, which first enlarged the part (grouped in three letters) of the QWERTY keyboard, then picked the wanted letter from the three. The DragKeys [6] used a circular menu, spatially mapped in the 8 principal directions. The first swipe selected the group that contained the target letter among the eight, which brought up the subgroup menu with the 1–5 letters laid out in 4 principal direction, but the most likely one in the middle position. If the wanted letter was already in the middle position, that letter was entered by default, otherwise one more directional swipe was needed to choose the wanted letter.

For those methods based on QWERTY layout, a minimum of two flicks/swipes is inescapable (and possibly more for special characters or capitalization), while those

[1] In a previous work, we have suggested for the same basic idea, combining touch and voice, for digit input on smart watches [27].

with non-QWERTY layout generally require a significant amount of training to get oneself skillful and acquire effectiveness. Moreover, approaches that rely on touch input and thus use multi-step input naturally exhibit a trade-off between the size of the key (and ease of selection) and the number of steps required to complete the input.

Voice input is increasingly being used for small hand-held devices with the improved recognition accuracy and easily accessible cloud based high end services. Voice input performs quite stably for entering single or short composite commands/words, but still limited in terms of their accuracy being sensitive to environmental noise and speaker characteristics [16]. In order to reduce the effects of the environmental noise, it is often the case that the user has to speak inconveniently close to the device (and visually confirm the command is recognized).

Instead, using voice to help type in the individual letter, as proposed here, could be a viable alternative (at least for entering short amount of texts). By relying on letters only, the voice recognition engine could be much lighter (to be even put locally on a smart watch itself) and the accuracy improved (e.g. with the voice recognition itself having less target candidates for a match (i.e. just the alphabets) already filtered by the touch input. In fact, such a multimodal input method is an attractive and effective solution to improving interaction performance, leveraging on different individual characteristics of the fused methods. For instance, the pioneering work of "Put-that-there" [3] had illustrated how two lesser reliable modality input methods can complement each other to boost the overall recognition performance. While multimodal interfaces are increasingly being applied to mobile and hand-held devices (e.g. touch + speech [23], speech + gesture [13, 23, 25], and speech + pen [9]), not many exist for making input to smart phones or watches. Voice input method, by itself or in combination with others, has rarely been applied to alphanumeric entry (by individual letters).

3 Vouch-T

As noted, Vouch-T combines touch and voice for individual letter input on small mobile devices. The touch input is selected because it is the most natural, popular and familiar to most users and has good ergonomic accessibility. The reliability of touch input is much affected by the size of the key, finger size and user motion. The voice input is selected for its ease of use and independence of its reliability to user motion. It is still generally less reliable than touch input due to environment noise, distance to the microphone, accent, intonation and other speaker characteristics. We project that these modality characteristics can complement each other to improve the usability and recognition performance.

The interaction model for entering a letter is simple (Fig. 1): (1) make an approximate touch on a familiar but small keyboard displayed on the smart device (generating possible candidates around the vicinity of the touch position), and (2) accompany a voice input to single out the target letter (and vice versa). The interaction model is expected to be fast (by the simultaneous or one stage sequential one-stage input [20]), easy (by just an approximate touch and using familiar vocabulary with light cognitive load), and accurate (by combining multiple evidences for recognition [18]).

In our study, Vouch-T is tested using two keyboard layouts, namely, the 3 × 4 keypad and the traditional QWERTY. More detailed explanation follows.

3.1 3 × 4 Keypad Layout

The telephone keypad is an ad hoc layout originally used for the old rotary and touch tone phones as a quick and dirty way to enter text. It may not be as popular as the QWERTY keyboard, but still somewhat familiar. There are 12 keys laid out in a 3 × 4 grid, and each cell is mapped to three or four alphabets, ordered from the top to bottom and left to right (see Fig. 1 – lower left). This layout has an advantage over the QWERTY for being compact, thereby making each button larger than those of the QWERTY (given the same screen space area). To enter a letter, two touches would be needed, first to select the group, then to select among the three letters in the group.

With Vouch-T, after the initial input, instead of another touch, the voice input is used to disambiguate among the three possible targets. The computational load for the voice input would be thus low, and relative recognition performance high, because there are only three search targets. Note that the voice input alone may not be sufficient for selecting among "all" letters, e.g. because of the usually relatively far microphone location and the environment noise in a mobile situation. Since, the three letters (assigned to a button) are already known, the voice input can be made simultaneously with the touch.

3.2 QWERTY Layout

The QWERTY layout is arguably the most popular alphanumeric key input layout, adopted for most typewriters and keyboards (including the virtual and touch screen) [30]. Depending on variations, all alphabets (plus special characters) are individually mapped to a single key (about 33 keys). Therefore, while it is the most familiar, it is also error-prone due to the reduced size given a small mobile device screen. A single touch/press is enough to enter a letter.

With Vouch, for each letter, the probability of the simultaneously accompanied voice input match and a score based on the distance from the finger touch input are used to compute a score to determine the finally selected letter. The final score is a weighted sum of the two factors, with the weights (w_1 and w_2) set empirically. As for the distance based score, a small circle is drawn around a given letter whose radius is set by the nearest neighboring letter (see Fig. 2). If a touch is made outside the circle, a full score of 1 is given for that letter. Otherwise, the normalized distance score is actual distance to the target letter divided by the radius of the circle. That is,

$$\text{Score}(\text{letter}) = \left\{ w_1 * \left(1 - P_{\text{voice\_recognition}}(\text{letter})\right) + w_2 * D(\text{letter}) \right\}^{-1}, \quad (1)$$

where $P_{\text{voice\_recognition}}$(letter) is the voice recognition score/probability for given letter and D(letter), the normalized distance score defined as:

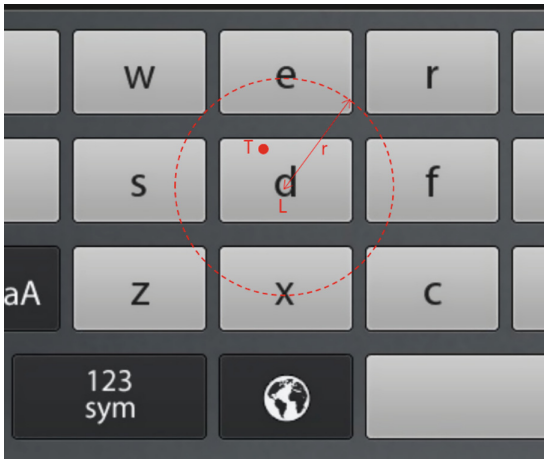

Fig. 2. Computing the distance based score for a given letter, "d". The position of the letter "d" is denoted L, the actual touch location, T, and the radius of circle, r. Outside the circle the normalized distance is clamped to 1, otherwise, (T−L)/r.

$$\mathrm{D(letter)} = 1, \mathrm{if}\ |\mathrm{T}-\mathrm{L}| > \mathrm{r}, \mathrm{otherwise} = |\mathrm{T}-\mathrm{L}|/\mathrm{r}, \tag{2}$$

where T is the touch position, L, the letter position, and r is the distance to the closest neighboring letter from L.

4 User Study

4.1 Experimental Design

We experimentally compared Vouch-T to the conventional unimodal touch-only method in terms of the text entry performance and usability. The main purpose of the experiment was to highlight the projected advantage of the multimodal approach for the small layout/key size. Comparison of Vouch-T to other notable "specialized" small screen text entry methods [4, 5, 8, 12, 17], as summarized in the Related Work section, is left as future work.

The experiment was designed as a 3 factor (2 layout types × 2 interface types × 2 screen sizes), but divided into two parts (Experiments 1-1 and 1-2): 2 layout types × 2 interface types on the smart phone and the same on the smart watch (thus two 2 factor, 2 × 2 within-subject repeated measure). Table 1 summarizes the experimental design and the total of eight test conditions.

The user was given a set of short sentences to enter under the respective test condition, and the following dependent variables were measured: (1) task completion time/rate (e.g. words per minute), (2) number of individual key inputs, (3) error rate (e.g. number of backspace/delete keys used), and (4) responses to the usability survey questions (answered in the 7 level Likert scale asking the ease of use, naturalness, ease

Table 1. Experimental design and the eight test conditions

Independent variable	Interaction type	Layout type	Symbol
Experiment 1-1: screen size type = smart phone	Touch-only	QWERTY	P-TQ
	Touch-only	3 × 4 Keypad	P-T34
	Vouch-T	QWERTY	P-VQ
	Vouch-T	3 × 4 Keypad	P-V34
Experiment 1-2: screen size type = smart watch	Touch-only	QWERTY	W-TQ
	Touch-only	3 × 4 Keypad	W-T34
	Vouch-T	QWERTY	W-VQ
	Vouch-T	3 × 4 Keypad	W-V34

of learning, fatigue, general satisfaction/future usage and preference, we omit the actual questionnaires for lack of space). We hypothesized that Vouch-T would be significantly superior to the touch-only interface at least for the case of smart watches for both QWERTY and 3 × 4.

4.2 Experimental Task and Set-up

The user was given a set of short sentences to enter under the respective test condition. For each condition, the subject was given 20 sentences all sampled from the MacKenzie and Soukoreff's test set [28], such that all alphabets would have to be used at least once. All subjects were given the same set of sentences. A different set of 20 sentences were used for each test condition.

As for the experimental set-up, the LG G2 was used as the smart phone and the LG G-Watch for the smart watch (both running the Android operation system) on which the respective key layout was displayed in the lower half and text entry made/shown in the upper. Figure 3 shows the screen display with the respective key layout on the smart phone and smart watch.

Ideally, the experiment should have been carried out with actual voice recognition used and running on the smart phone or smart watch. However, in the case of the smart phone, no voice recognition software API (on the Android) that provides recognition certainty values could be found. Using the cloud based service would present the same problem (plus latency for the fast paced stream of individual letter input vs. word level recognition). Moreover, smart watches (including the one we tested) are not computationally powerful enough to run voice recognition to begin with. Therefore, our experiment simulated the use of the voice by preconfiguring the letter recognition probability values off-line. Our implementation used the Daum Newtone speech recognition API (for the Android) was used with which only the top 10 ranked recognized candidates were made available [29]. We estimated the probabilities of each letter recognition, Prob_{est} (letter), in this way.

$$Prob_{est}(letter) = \frac{\sum_i w_i x_i}{\sum_{allletters} Prob_{est}(l)}, w_i = 1/2^i. \quad (3)$$

where, w_i is the weight given to the rank i, and X_i is the number of occurrences of the letter with rank i. The weight was arbitrarily set to $1/2^i$ (e.g. half decay function) for the

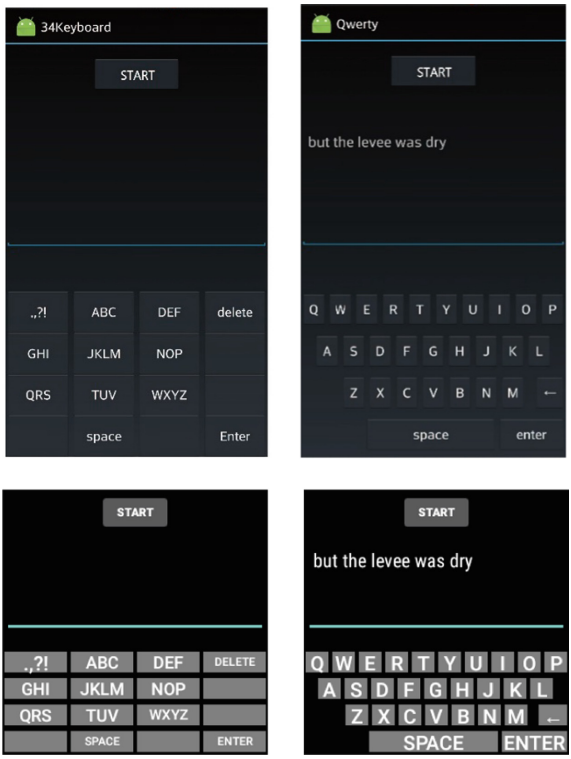

Fig. 3. The 3 × 4 keypad and QWERTY key layout on the smart phone (top) and smart watch (below). The key layout was located in the lower half part of the screen and the input made and shown in the upper part. The start button was used for the testing purpose.

ith ranked data. Thus, for example, if in an attempt to recognize the letter/word, "a" (pronounced "eigh") 10 times, the "a" could be ranked the first, 5 times, ranked the second, 3 times, ranked the third, 0 times, and the fourth, 2 times (see Table 2 as an example), resulting into:

$$\text{Prob}_{\text{est}}(\text{“a”}) = 0.5 * 5 + 0.25 * 3 + 0.125 * 0 + 0.0625 * 2 = 3.375 \text{ (unnormalized)}$$

Since voice recognition performance is quite speaker-dependent, the performance table as shown in Table 2 was constructed and probabilities computed for each subject. Even so, such an estimation is not correct. Yet, it was deemed close enough for modelling the gross behaviour of the letter recognition performance and assessing the effect of adding the voice input to the touch. The user was still instructed to speak out the desired letter (assumed to be correct) even though the recognition was simulated according to the preconfigured probabilities internally.

In actual practice, combining the touch and voice brought about the synergistic effect of disambiguation. For instance, in the case of 3 × 4 key layout, when entering for the letter "D" (thus pressing the "DEF" button and speaking out "dee"), the voice

Table 2. Recognition results (10 runs) for the letter "a" (pronounced "eigh") and rankings of the possible matches (top 10)

Run no.	Rankig (w_i)				
	1 (1/2)	2 (1/4)	3 (1/8)	…	10 (1/1024)
1	a	j	h	…	n
2	a	h	j	…	m
…				…	
10	a	k	j	…	l

recognition produces the match candidates in the approximate order of {"b," "p"," "d," "t," "v," "e," "k," "g," "a," "j"}. Despite the letter "d" being ranked only the third, "b" and "p" are not even considered and removed from the match list.

4.3 Experimental Procedure

Ten paid subjects (5 men and 5 women between the ages of 20 and 36, mean = 27.9/SD = 5.36) participated in the experiment. After collecting their basic background information, the subjects were briefed about the purpose of the experiment and given instructions for the experimental tasks. A short training (3–7 min or until the subject felt sufficiently trained) was given to allow the subjects to become familiarized with the experimental process and the touch-only or Vouch-T based text entry methods on the respective devices. All subjects were right handed, used to wearing watches on their left wrist, and never had the prior experience of using the smart watches.

The subjects held the smart phone or wore the smart watch in their left hand/wrist comfortably. The user sat and interact using their upper body naturally with one's arms and fingers (see Fig. 4). There were total of 8 different treatments (or conditions), namely, 2 interfaces × 2 key sizes × 2 devices (total 8 blocks). Each treatment was presented in a balanced fashion using the Latin square methodology. The user carried out the given text entry task (e.g. "I am a boy you are a girl"). In each treatment block,

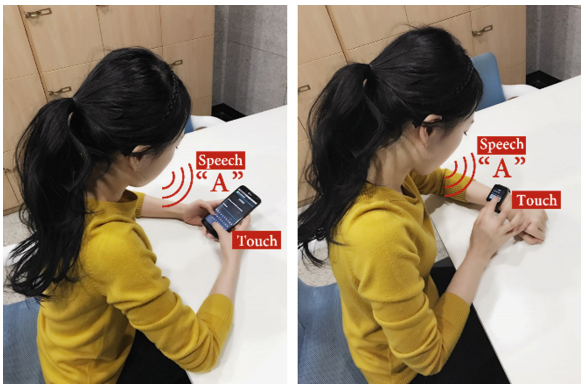

Fig. 4. A subject carrying out the experimental task (entering 20 sentences) using the smart phone (left) and the smart watch (right).

the user entered a series of 20 sentences. The task started by the user pressing the "start" button, then 20 sentences would appear for which the user was to enter. The single block was finished by touching the "enter" button after entering all the sentences. Both the task completion time and the number of delete key usage were recorded. The user was asked to make the entries correctly as fast as possible. If errors were made, the user had to use the backspace/delete key to make corrections. After the session, the subjects filled out a usability questionnaire. The user rested between each treatment and the total experimental session took about 1 h.

5 Results

5.1 Smart Phone

On the smart phone, subjects clearly performed better in terms of the task completion time using the QWERTY (P-TQ and P-VQ) vs. 3 × 4 (P-T34 and P-V34, Tukey/p-value = 0.05). Subjects reported of the strong preference of the QWERTY and the unfamiliarity and fatigue with the 3 × 4. While using the voice was helpful for improving the typing performance for the 3 × 4 (P-V34 < P-T34), it was not so for the QWERTY (no statistically significant difference between P-TQ and P-VQ). On the smart phone, the keys were sufficiently large and touch-only operation too familiar, making the effect of the voice disambiguation only marginal. See Fig. 5.

Figure 6 (top) shows the total number of keys entered for the four test conditions. Only the case of P-T34 showed an excessive (almost two times) key input compared to the others with a statistically significant difference by ANOVA (vs. P-V34, p-value = 0.000). On the other hand, Fig. 6 (bottom) shows the number of delete/backspace keys used for making corrections, and both touch-only cases, P-TQ and P-T34 exhibited much higher rates (but no statistically significant difference). It supports what was observed with the task completion time results in that the while the added voice input helped the text input using the 3 × 4 key layout, it was the reverse for the QWERTY (more proportion of delete keys over the total key input with P-VQ than with P-TQ).

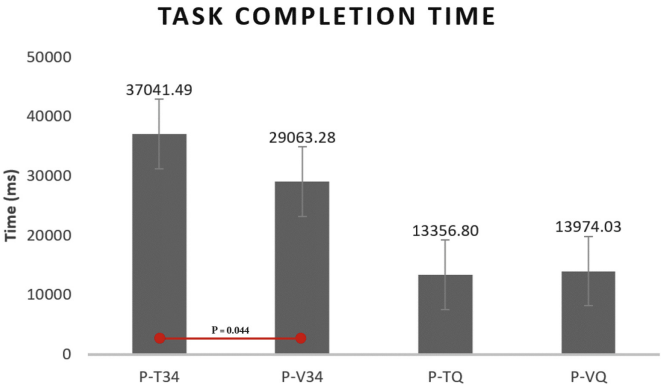

Fig. 5. Task completion time (time taken for entering 20 sentences) on the smart phone.

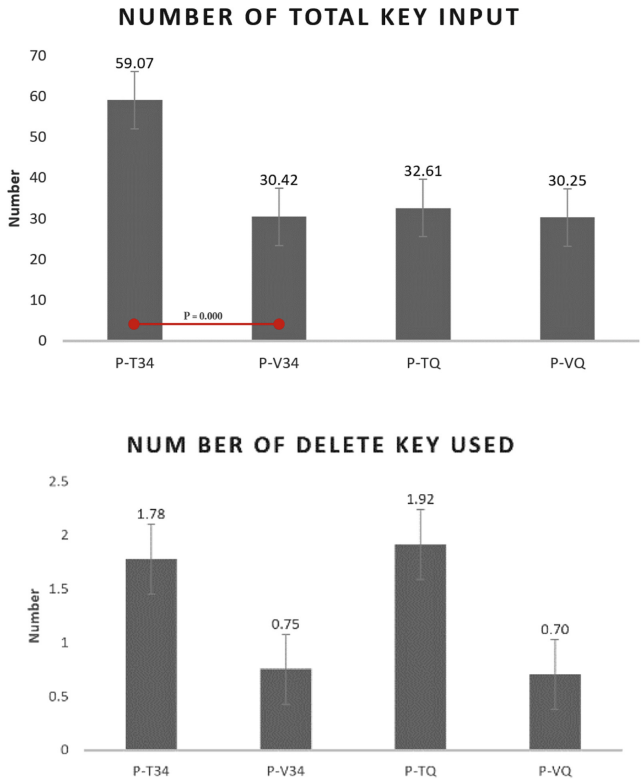

Fig. 6. Total number of key inputs made (top) and total number of delete keys used (bottom) on the smart phone.

5.2 Smart Watch

On the smart watch, Fig. 7 shows that the added voice input helped both the QWERTY (W-VQ < W-TQ, p-value = 0.033) and 3×4 key layout (W-V34 < W-T34, p-value = 0.000) in terms of the task completion time. Between the W-VQ and W-V34, there was not statistically significant difference. Thus, with the smaller sized watch, the advantage of the QWERTY with respect to the familiarity was reduced.

In terms of the total and types of keys entered, most inputs (and thus corrections) were made with the 3×4 key layout (compared to all other conditions), thus causing the longer overall task completion time. W-VQ also produced statistically less total and delete key input compared to W-TQ, differently from the case of the smart phone (See Fig. 8). The use of voice had a more pronounced effect of improved task performance on the smart watch.

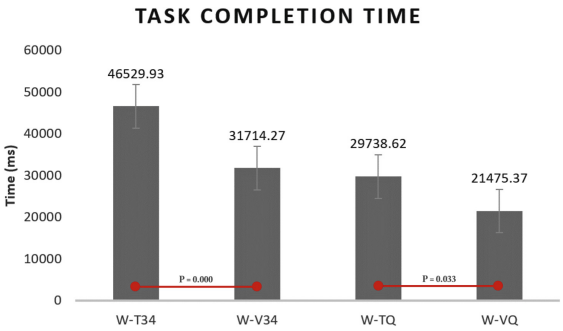

Fig. 7. Task completion time (time taken for entering 20 sentences) on the smart watch.

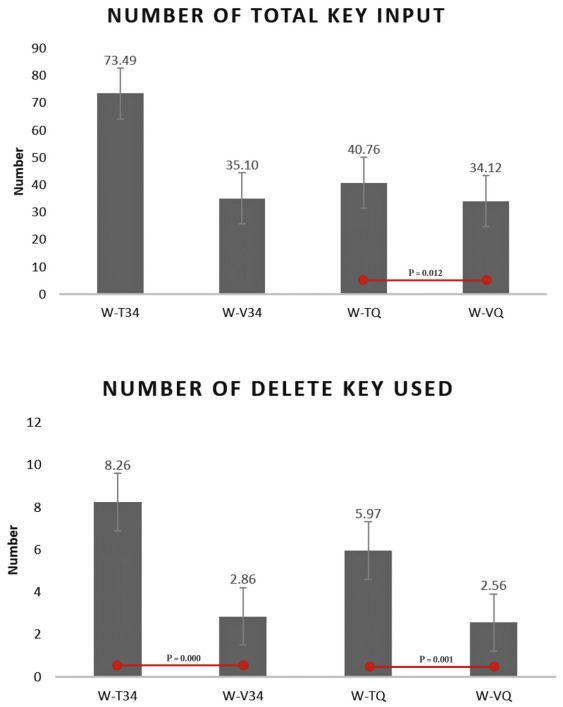

Fig. 8. Total number of key inputs made (top) and total number of delete key inputs made (bottom) on the smart watch.

5.3 Smart Phone Vs. Smart Watch

Between the smart phone and smart watch, as can be expected, the task complete times were generally longer on the smaller sized smart watch (P-VQ vs. W-VQ, P-TQ vs. W-TQ (nearly two times, p-value = 0.000), P-V34 vs. W-V34, and P-T34 vs. W-T34,). However, the difference was not statistically significant when the voice was used

(P-VQ vs. W-VQ and P-V34 vs. W-V34). This indicates that the task performance is less influenced by the screen/key size if the voice is added as a secondary input channel, and the same time, the familiarity of the QWERTY is less effective with the smaller sized device as well. In reverse, on the smart watch, the added voice turned out to be effective in making W-QV more performing than W-TV (p-value = 0.033), and not so on the larger smart phone. The statistical data with regards to the total and types of keys entered were consistent as well (e.g. more keys and deletes used with the smart watch). See Figs. 9, 10 and 11.

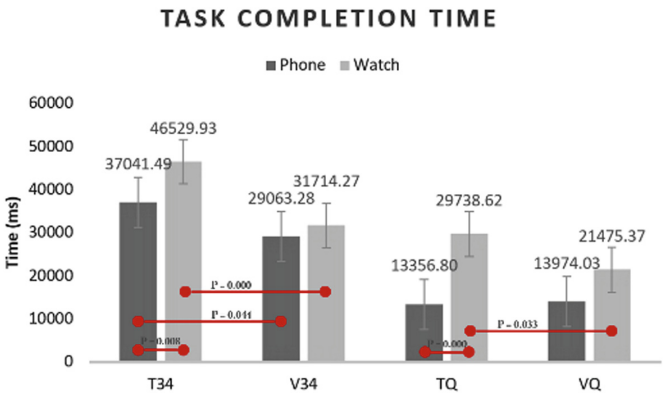

Fig. 9. Task complete times between the smart phone and smart watch.

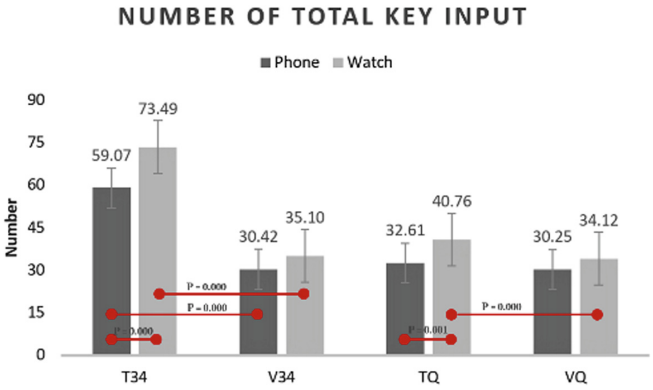

Fig. 10. Total number of keys used between the smart phone and smart watch.

5.4 Usability

The usability survey generally confirmed our general expectation of the familiarity and preference of the QWERTY over the 3 × 4 layout, and higher subjective usability (over most categories) when the larger phone was used to enter the text. Users also

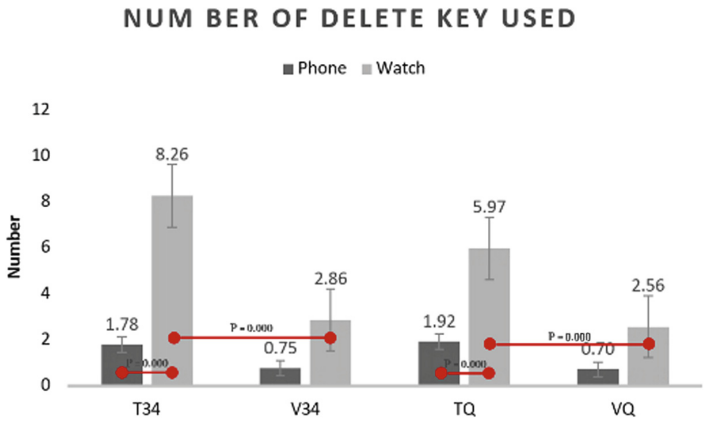

Fig. 11. Total number of delete keys used between the smart phone and smart watch.

generally felt that the added voice modality helped improve the usability for both the QWERTY and 3 × 4 layout and also for both the smart phone and smart watch.

5.5 Summary

The quantitative experiment results can be summarized as follows.

- The input performance was better on the larger smart phone (than on the small smart watch) as easily expected.
- The added voice input modality was helpful on the smaller smart watch for both QWERTY and 3 × 4 key layout, but, on the smart phone, not so on the smart phone, only for the 3 × 4.
- On the smart phone, the best performance was obtained with the QWERTY (P-TQ) where the Vouch-T did not help very much.
- On the smart watch, the best performance was obtained with QWERTY-V (W-VQ) Vouch-T did help in a significant manner.
- QWERTY was preferred over the 3 × 4 on both devices due to its familiarity.

6 Conclusion

In this paper, we presented a prototype implementation of Vouch-T which mixes up the use of touch and voice in a multi-level alphanumeric input for small hand-held or wearable devices. The combination of voice and touch creates a synergy in that it effectively allows for the use of larger sized keys on the screen and near simultaneous input across the input levels, making it a faster method.

We have assessed the potential of Vouch-T in terms of the usability and performance compared to the conventional touch-only based method. We considered two types of text input layout, namely, the QWERTY and 3 × 4 telephone keypad on two

sizes of mobile devices, the smart phone and smart watch. The comparative simulation experiment validated that the multimodal approach of Vouch-T improves the input performance and usability for the smaller-sized smart watch, but only marginally for the larger smart phone. Even though our experiment was based on a simulated voice recognition, we believe it was sufficient to show the effectiveness of the multimodal approach especially for under difficult operating conditions such as with small smart watches.

As future work, Vouch-T needs to be compared to methods other than just the touch-only case, such as using word-level voice input and other small screen input methods as outlined in our related work (e.g. Zoom-board [17]). We also hope to assess the advantages of the multimodal approach in other difficult situations such as while moving and or under low visibility situations.

Acknowledgement. This work was supported by the Technology Innovation Program (or Industrial Strategic Technology Development Program (10053638, International technology standardization of 3D data information, and industrialization) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea) and also by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2011-0030079).

References

1. Albinsson, P.A., Zhai, S.: High precision touch screen interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2003), pp. 105–112. ACM (2003)
2. Baudisch, P., Chu, G.: Back-of-device interaction allows creating very small touch devices. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009), pp. 1923–1932. ACM (2009)
3. Bolt, R.A.: "Put-that-there": Voice and Gesture at the Graphics Interface. ACM, New York (1980)
4. Cha, J.M., Choi, E., Lim, J.: Virtual sliding QWERTY: a new text entry method for smartwatches using Tap-N-Drag. Appl. Ergon. **51**, 263–272 (2015)
5. Chen, X.A., Grossman, T., Fitzmaurice, G.: Swipeboard: a text entry technique for ultra-small interfaces that supports novice to expert transitions. In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST 2014), pp. 615–620. ACM (2014)
6. Cho, H., Kim, M., Seo, K.: A text entry technique for wrist-worn watches with tiny touchscreens. In: Proceedings of the Adjunct Publication of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST 2014), pp. 79–80. ACM (2014)
7. Darbara, R., Senb, P. K., Dasha, P., Samantaa, D.: Using hall effect sensors for 3D space text entry on smartwatches. In: Proceedings of the 7th International Conference on Intelligent Human Computer Interaction (IHCI 2015) (2015)
8. Dunlop, M.D., Komninos, A., Durga, N.: Towards high quality text entry on smartwatches. In: CHI 2014 Extended Abstracts on Human Factors in Computing Systems, pp. 2365–2370. ACM (2014)

9. Dusan, S., Gadbois, G. J., Flanagan, J.L.: Multimodal interaction on PDA's integrating speech and pen inputs. In: INTERSPEECH (2003)
10. Funk, M., Sahami, A., Henze, N., Schmidt, A.: Using a touch-sensitive wristband for text entry on smart watches. In: CHI 2014 Extended Abstracts on Human Factors in Computing Systems, pp. 2305–2310. ACM (2014)
11. Holzinger, A.: Finger instead of mouse: touch screens as a means of enhancing universal access. In: Carbonell, N., Stephanidis, C. (eds.) UI4ALL 2002. LNCS, vol. 2615, pp. 387–397. Springer, Heidelberg (2003). doi:10.1007/3-540-36572-9_30
12. Hong, J., Heo, S., Isokoski, P., Lee, G.: SplitBoard: a simple split soft keyboard for wristwatch-sized touch screens. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2015), pp. 1233–1236. ACM (2015)
13. Kurihara, K., Goto, M., Ogata, J., Igarashi, T.: Speech pen: predictive handwriting based on ambient multimodal recognition. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2006), pp. 851–860. ACM (2006)
14. Kwon, S., Choi, E., Chung, M.K.: Effect of control-to-display gain and movement direction of information spaces on the usability of navigation on small touch-screen interfaces using Tap-N-Drag. J. Ind. Ergon. **41**(3), 322–330 (2011)
15. Leiva, L.A., Sahami, A., Catalá, A., Henze, N., Schmidt, A.: text entry on tiny QWERTY soft keyboards. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2015), pp. 669–678. ACM (2015)
16. Nebeling, M., Guo, A., Murray, K., Tostengard, A., Giannopoulos, A., Mihajlov, M., Bigham, J.P.: WearWrite: orchestrating the crowd to complete complex tasks from wearables. In: Proceedings of the 28th Annual ACM Symposium on User Interface Software and Technology (UIST 2015), pp. 39–40. ACM (2015)
17. Oney, S., Harrison, C., Ogan, A., Wiese, J.: 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 2013), pp. 2799–2802. ACM (2013)
18. Oviatt, S.: Mutual disambiguation of recognition errors in a multimodel architecture. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1999). pp. 576–583. ACM (1999)
19. Paliwal, K., Basu, A.A.: Speech enhancement method based on Kalman filtering. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Process (ICASSP), pp. 177–180 (1987)
20. Roudaut, A., Huot, S., Lecolinet, E.: TapTap and MagStick: improving one-handed target acquisition on small touch-screens. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 146–153. ACM (2008)
21. Schüssel, F., Honold, F., Schmidt, M., Bubalo, N., Huckauf, A., Weber, M.: Multimodal interaction history and its use in error detection and recovery. In: Proceedings of the 16th International Conference on Multimodal Interaction, pp. 164–171. ACM (2014)
22. Siek, K.A., Rogers, Y., Connelly, K.H.: Fat finger worries: how older and younger users physically interact with PDAs. In: Costabile, M.F., Paternò, F. (eds.) INTERACT 2005. LNCS, vol. 3585, pp. 267–280. Springer, Heidelberg (2005). doi:10.1007/11555261_24
23. Tsourakis, N.: Using hand gestures to control mobile spoken dialogue systems. Univ. Access Inf. Soc. **13**(3), 257–275 (2014)
24. Turunen, M., Hurtig, T., Hakulinen, J., Virtanen, A., Koskinen, S.: Mobile speech-based and multimodal public transport information services. In: Proceedings of MobileHCI 2006 Workshop on Speech in Mobile and Pervasive Environments (2006)
25. CMU Sphinx. http://cmusphinx.sourceforge.net/
26. Watch, L.G.: http://www.lg.com/us/smart-watches/lg-W100-g-watch/

27. Lee, J., Kim, G.J.: Vouch: multimodal input for smart watches under difficult operating conditions using touch and voice. J. Multimodal Interfaces (2016, submitted)
28. MacKenzie, I.S., Soukoreff, R.W.: Phrase sets for evaluating text entry techniques. Ext. Abstracts CHI 2003, pp. 754–755. ACM Press (2003)
29. Newtone, D.: https://developers.daum.net/services/apis/newtone
30. QWERTY. https://en.wikipedia.org/wiki/QWERTY