A	Apple
B	Bacon
C	Camera
D	Doctor
E	Ear
F	Fox
G	Game
H	Hand
I	Ice
J	Jack
K	King
L	Lucky
M	Money
N	Nest
O	Orange
P	Park
Q	Queen
R	Rabbit
S	Soccer
T	Tour
U	Uncle
V	Violin
W	Wine
X	X-ray
Y	Young
Z	Zoo
SPACE	Space

Table 1: Our custom phonetic alphabet

Error Correction of Speech Recognition by Custom Phonetic Alphabet Input for Ultra-Small Devices

Kazuki Fujiwara
The University of Tokyo
Tokyo, Japan
isfujiwara@is.s.u-tokyo.ac.jp

CHI'16 Extended Abstracts, May 07-12, 2016, San Jose, CA, USA
ACM 978-1-4503-4082-3/16/05.
http://dx.doi.org/10.1145/2851581.2890380

Abstract
Automatic speech recognition (ASR) is one of the most effective ways to input text, in particular, for ultra-small devices such as smartwatches. Although the accuracy of ASR has been improving these days, it still often makes recognition errors. If you want to correct words that have been recognized incorrectly, you need to use a software keyboard or read out the words again. However, it is difficult and annoying to input text correctly using a software keyboard on a small display. Besides, even if you read out the same phrase again, there is no guarantee that your speech will be recognized correctly. To address this problem, we designed a custom phonetic alphabet optimal for ASR. It enables the user to input words more accurately than spelling them out directly or using the NATO phonetic alphabet, which is known as the standardized phonetic alphabet used for human-human speech interaction under noise. Furthermore, we conducted user studies to verify our method's efficiency in correcting speech recognition errors on a small display.

Author Keywords
speech recognition; phonetic alphabet; smartwatch

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: Input devices and strategies; Interaction styles

A	Alpha
B	Bravo
C	Charlie
D	Delta
E	Echo
F	Foxtrot
G	Golf
H	Hotel
I	India
J	Juliet
K	Kilo
L	Kima
M	Mike
N	November
O	Oscar
P	Papa
Q	Quebec
R	Romeo
S	Sierra
T	Tango
U	Uniform
V	Victor
W	Whiskey
X	X-ray
Y	Yankee
Z	Zulu

Table 2: NATO phonetic alphabet

Introduction

Smart devices are becoming more and more miniaturized these days. The displays of smartwatches are too small to input text with software keyboards like those on smartphones. There are some previous works related to ultra-small software keyboards based on the QWERTY keyboard (e.g. ZoomBoard [5], SplitBoard [3], Swipeboard [2]), but it is still not easy to input text. To begin with, it is uncertain whether touch-based text input is suitable for smartwatches [1].

On the other hand, automatic speech recognition (ASR) is one of the most effective ways to input text without using software keyboards. However, it is still difficult to input text perfectly considering ambient noise and the dialect of the speakers, even though the accuracy of ASR has improved [4]. Furthermore, there are many words that are difficult to recognize; the existence of proper nouns complicates matters still further. It is also said that the computing power of smartwatches is limited for complex speech recognition [6]. Thus, we need another approach to deal with this problem.

In this paper, we propose an ASR-based approach. The users input text using ordinary speech recognition, and then correct incorrectly recognized words by using the phonetic alphabet, where each spoken word is associated with a corresponding alphabet or symbol. The result of the speech recognition is displayed, and then the users select parts they want to correct by tapping and swiping on the incorrect parts if there are wrong parts. Selecting parts in units of word makes it easier to designate range on a small display. Finally, the users input spellings by using the phonetic alphabet.

Furthermore, we designed a custom phonetic alphabet for ASR. The NATO phonetic alphabet is known as a standardized phonetic alphabet for human-to-human speech interaction under noise, but it is not designed for ASR and everyday use. We adopted words that are shorter in average than those in NATO, and which are used in daily life and recognized easily (e.g. “apple”, “bacon”, “camera”). These words were chosen from words that everyone knows. Our study confirmed that our phonetic alphabet is more easily recognized than NATO’s.

The main contributions of this study are:

- Using the phonetic alphabet for ASR error correction
- Designing a custom phonetic alphabet suitable for ASR

Related Works

Swipeboard

Swipeboard [2] is a text entry technique for ultra-small devices such as smartwatches. Nine regions subdivided from QWERTY keyboards are shown on a display at first, and each region has three or four characters. The user does two actions to input one character. Each action is either a swipe or a tap. The first action specifies one of the regions, and the second action determines a character to input. With less than two hours’ training, Swipeboard users achieved 19.58 words per minute, making it 15% faster than ZoomBoard [5], its previous work.

The Phonetic Alphabet for Non-native Speakers

It was found that the pronunciation habits and characteristics of the Chinese are different from those of Europeans and Americans [7]. For people whose mother tongue is English, it is easy to get the corresponding letter from the phonetic alphabet. However, it is very difficult for most Chinese people to use it. To deal with this issue, the English

Figure 1: The result of the speech recognition is displayed on the top, and the part that the user selected is highlighted in green.

Figure 2: The input by our custom phonetic alphabet is displayed in the middle of the display. In this case, the user should say “Fox, Ice, Lucky, Ear.”

Phonetic Alphabet applicable to Chinese people was designed in 2013 [7]. As a result of the study, it was revealed that there is a huge difference between Chinese and English pronunciation, so it is difficult to combine the phonetic alphabet for native speakers and the phonetic alphabet for non-native speakers.

Interaction Workflow

Figure 1 shows the appearance of the display. Normal speech recognition starts when the ’Record’ button is pressed. If there are some misrecognized words, the user can select the words by tapping and swiping on the words. In this study, the user can choose the area only word by word because ASR makes errors in words rather than in characters. Then, the selected area becomes highlighted. When the user pushes the ’Record’ button in this state, error correction by our custom phonetic alphabet will start (see Figure 2). The result of the input will be displayed in the center of the display after the corresponding phonetic alphabets are read out. If the result corresponds to the user’s intended meaning, the selected parts will be replaced by pushing the ’OK’ button; otherwise, the user can input characters from the beginning by pushing the ’Record’ button again.

As a proof-of-concept, we used a prototype of the application for smartwatches in JavaScript and native touch events on iPhone 6. The size of the display is 30mm $\times$ 24mm.

The Phonetic Alphabet

Our custom phonetic alphabet consists of 27 words (26 words corresponding to each alphabet + "space") as shown in Table 1. Since some words used in the NATO phonetic alphabet are unfamiliar to non-native English speakers, we chose words used in daily life and that are familiar to them. The words are shorter in average than the words used in the NATO phonetic alphabet. We also chose words that are easily distinguished by ASR. These words are currently selected subjectively by the authors based on the recognition performance of our own speeches.

In this study, we prepared 27 words first for the phonetic alphabet. Then, after having these words pronounced by non-native speakers, we replaced the incorrect words with other candidate words. We performed this operation iteratively, until we had finally adopted all 27 words. Our phonetic alphabet has higher recognition-accuracy than the NATO phonetic alphabet since we chose words that are easily recognized not by people but by ASR.

Preliminary Study

We conducted a preliminary study to test the accuracy of ASR.

Procedure

We recruited five participants (two females and three males; the average age is 26.8 years, and all participants are Japanese). Each participant was asked to read out seven phrases (see Table 3) that are shown on a display in a sequence and we collected the input data. The first six phrases were chosen from phrases used in daily conversation, and the last one was chosen from sentences that are often recognized incorrectly by ASR. In this study, we used Dragon Mobile SDK[1] for ASR.

Result

The accuracy rate of ASR was 62.8% of the total, as shown in Table 3. No participant could input all sentences perfectly, so we conclude that it is difficult to input sentences using ASR without any error correction. Phrases that are

[1] Dragon Mobile SDK - Nuance Developers: https://developer.nuance.com/

Sentence	User1	User2	User3	User4	User5	Accuracy Rate
Thank you very much for your help.	✓	✓	✓	✓	✓	100%
I'm afraid I can't attend the class.	x	✓	✓	x	x	40%
I'll give you a call later.	x	✓	✓	✓	✓	80%
Please send me the file.	✓	✓	✓	x	x	60%
How's the weather there.	✓	✓	✓	✓	x	80%
We had snow this morning.	x	x	✓	✓	x	40%
The order of these words is not important.	x	✓	x	x	✓	40%

Table 3: The results of the recognition-accuracy test in the preliminary study: We counted the number of inputs without any mistakes. The average of accuracy rate was 62.8%.

pronounced less loudly are likely to be missed or recognized incorrectly (e.g. "I'll" can be recognized as "I", and "the class" as "a class" or "class"). Besides, there was some misrecognition unique to Japanese, such as the difference between "l" and "r" (e.g. "file" is likely to be mistaken for "fire"). It was also revealed that longer sentences are likely to be recognized incorrectly.

Evaluation Study

We evaluated the effectiveness of our method through a comparative study.

Procedure

We compared four text-input methods, as follows, to verify our method's efficiency in correcting errors on a small display.

- Our method (our custom phonetic alphabet)
- NATO phonetic alphabet
- Direct alphabet input (e.g. saying "C, A, R" for "car")
- QWERTY keyboard

In this study, a sentence with the incorrect words highlighted appears on a display of the smartwatch, and the correct input is shown outside display in advance. We measured the task completion time. We started timing when the 'start' button was pushed, and stopped when the 'finish' button was pushed. Both buttons are located outside display of the smartwatch. We stopped timing if participants needed more than 40 seconds to fix the incorrect words. We added the 'space' command to the NATO phonetic alphabet and Direct alphabet input by saying "space".

We recruited five participants (three females and two males; the average age is 20.6 years, and all the participants are Japanese and different from the participants of the preliminary study). The participants were asked to correct five sentences, as shown below, for each of the input methods ($5 \times 4 = 20$ sentences in total). All the incorrect sentences used in this study were actually observed in ASR.

- Task 1: Please send me the fire ("fire" $\rightarrow$ "file")
- Task 2: I'm afraid I can't attend across ("across" $\rightarrow$ "the class")
- Task 3: We have snow this morning ("have" $\rightarrow$ "had")

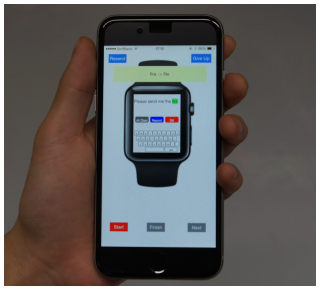

Figure 3: Study screen: In this study, we used a prototype of the application for smartwatches on iPhone 6.

- Task 4: The order of these wars is not important (“wars” → “words”)
- Task 5: You’ll and house your vocabulary by reading the news (“and house” → “enhance”)

Note that there is no command for removing text character by character like the ’delete’ key on software keyboards. We did not adopt the “delete” command, as we did not want it to be recognized as another phonetic alphabet. In our method and the NATO phonetic alphabet, the participants had access to the table of words during the study. We used OpenEars[2] for ASR in our custom phonetic alphabet, the NATO phonetic alphabet and Direct alphabet input.

Result

Table 4 shows the percentages of the cases where participants successfully fixed sentences correctly within 40 seconds for each method. Some tasks for which ASR was used were not finished in time, while all tasks for which a software keyboard was used were completed in time. Especially in the direct alphabet input method, almost all tasks ended in failure. Our phonetic alphabet had a higher recognition rate than the NATO phonetic alphabet.

Method	Task 1	Task 2	Task 3	Task 4	Task 5	Avg.
Our method	100%	60%	80%	80%	60%	76%
NATO phonetic alphabet	100%	40%	20%	60%	20%	48%
Direct alphabet input	20%	0%	0%	0%	40%	12%
QWERTY keyboard	100%	100%	100%	100%	100%	100%

Table 4: The percentages of the cases where participants could fix sentences correctly within 40 seconds for each method. Our ASR method achieved higher accuracy than other methods using ASR.

[2]OpenEars - free speech recognition and speech synthesis for the iPhone: http://www.politepix.com/openears/

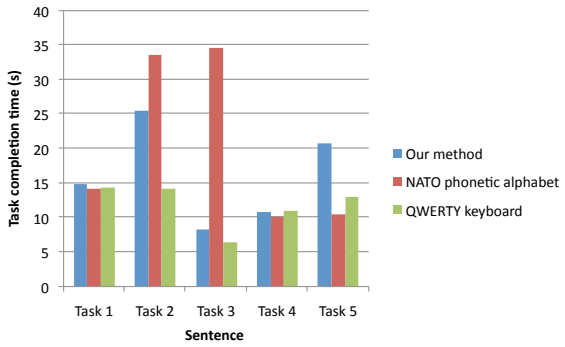

Figure 4: The average time required to complete tasks for each method: Our method could input text faster than NATO phonetic alphabet.

In our method, we observed that shorter words are more likely to cause misrecognition. For example, in the case of the word "hand", the “h” sound was often missed or recognized from ambient noise. Words that have similar sounds (e.g. "apple" and "uncle") were also often misrecognized.

Figure 4 shows the average time required to complete tasks for each method. We excluded the direct alphabet input in this graph because almost none of the tasks were completed in time. Note that the average times in this graph also excluded cases where it takes more than 40 seconds to correct errors. Almost all participants could correct errors by using our phonetic alphabet as fast as by using software keyboards. There is a case in which our method took much longer than the NATO phonetic alphabet (Task 5), but accuracy was significantly low (20%) for the NATO in this case. These results show that using our phonetic alphabet is more efficient than the NATO phonetic alphabet for error correction on an ultra-small display.

Discussion and Future Work

The result shows that our method is slightly slower than software keyboards in this particular hardware setup. However, as screen sizes get smaller, software keyboards become harder to use. We expect the benefits of our method to be more profound in such cases. In addition, there is room for improvement in recognition accuracy and speed by making some modifications.

Our current alphabet is an adhoc selection by the authors. An immediate future work would be to device a systematic procedure to design a custom phonetic alphabet for a particular user group and recognition engine. A possible approach would be to systematically measure recognition accuracy of many candidate words and find an optimal set that minimizes the total number of recognition errors.

Another immediate future work would be to implement the "delete" method in ASR. We did not adopt the "delete" command, as we did not want it to be recognized as a word in this study. The user will be able to input text faster by choosing proper phonetic alphabet for the "delete" command.

Semi-real-time speech recognition will also improve the usability of ASR. As of now, the users receive the result of recognition after short interval, and then decide whether they would like to continue to input text or read out the same phrase again. Receiving the results while reading out sentences will enable the users to notice errors earlier and relieve their stress.

Conclusion

ASR is an effective way to input text on ultra-small devices such as smartwatches, but we need a method for correcting errors because speech recognition often makes mistakes. We consider that, rather than speaking the same word again for correcting errors, using the phonetic alphabet is preferable. By using our custom phonetic alphabet, we can input text faster and more accurately than by using the existing phonetic alphabet that is widely used in radiotelephone communications, such as the NATO phonetic alphabet. Our evaluation study showed the effectiveness and potential of ASR, and it is expected that its effectiveness will be improved by making further modifications.

References

[1] Barbara S Chaparro, Jibo He, Colton Turner, and Kirsten Turner. 2015. Is Touch-Based Text Input Practical for a Smartwatch? In *HCI International 2015-Posters' Extended Abstracts*. Springer, 3–8.

[2] Xiang 'Anthony' Chen, Tovi Grossman, and George Fitzmaurice. 2014. Swipeboard: A text entry technique for ultra-small interfaces that supports novice to expert transitions. In *Proc. UIST '14*. ACM, 615–620.

[3] Jonggi Hong, Seongkook Heo, Poika Isokoski, and Geehyuk Lee. 2015. SplitBoard: A Simple Split Soft Keyboard for Wristwatch-sized Touch Screens. In *Proc. CHI '15*. ACM, 1233–1236.

[4] Michael Longé, Richard Eyraud, and Keith C Hullfish. 2013. Multimodal disambiguation of speech recognition. (2013). https://www.google.com/patents/US7881936 US Patent 7,881,936.

[5] Stephen Oney, Chris Harrison, Amy Ogan, and Jason Wiese. 2013. ZoomBoard: a diminutive qwerty soft keyboard using iterative zooming for ultra-small devices. In *Proc. CHI '13*. ACM, 2799–2802.

[6] Dipesh Pradhan and Nugroho Sujatmiko. 2014. *Can smartwatch help users save time by making processes efficient and easier?* Master's thesis. University of Oslo. 18 Nov.

[7] Yuhui Wan. 2013. A Novel Approach for English Phonetic Alphabet in Wireless Communication. (2013).