Multimodal Chinese Text Entry with Speech and Keypad on Mobile Devices

Yingying Jiang[1] **Xugang Wang**[1,2] **Feng Tian**[1]
[1]Intelligence Engineering Lab & Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences
P.O.Box 8718, Beijing, China
{jyy,wxg,tf}@iel.iscas.ac.cn
+86-10-6266-1577

Xiang Ao[2] **Guozhong Dai**[1] **Hongan Wang**[1]
[2]Ministry of information Industry Software and Integrated Circuit Promotion Center
{ax, dgz, wha}@iel.iscas.ac.cn
+86-10-6266-1577

ABSTRACT
Chinese text entry is challenging on mobile devices which rely on keypad input. Entering one character may require many key presses. This paper proposes a multimodal text entry technique for Chinese. In this method, Chinese user can enter Chinese text by simultaneously using the simplified phonemic input method named Jianpin with keypad and speech utterance. The key of the technique is a multimodal fusion algorithm, which synchronizes speech and keypad input and fuses redundant information from two modalities to get the best candidate. A preliminary evaluation shows that users appreciate this technique and it could reduce key presses and enhance the input efficiency.

Author Keywords
Chinese text entry, mobile devices, multimodal, speech, keypad.

ACM Classification Keywords
H5.2. Information interfaces and presentation: User Interfaces- Interaction styles

INTRODUCTION
Many mobile devices, such as PDAs and cell phones, rely on keypad input. In these devices, text writing, ranging from entering URLs and search queries, typing commands, to writing emails and messages, is one of the hard problems awaiting solutions [13].

Chinese text input presents unique challenges to the field of human computer interaction [11]. As an ideographic language, Chinese is fundamentally different from English. The minimum functional unit in Chinese is called a character, which is a blend of sound, structure, and meaning [7]. In general, Chinese input methods can be classified as pronunciation-, structure- or arbitrary-based coding schemes [4]. Currently, on mobile devices such as cell phones, the dominant Chinese text input method is a pronunciation-based method, called "T9" [9], which lets users input pinyin[1], the symbolized pronunciation denoted by a - z, of characters to be entered, and then select from a candidate list. However, as the keys on mobile devices are quite small and each of them corresponds to 3 or 4 Latin letters, many key presses.will be required to enter just a few characters.

Speech and handwriting input methods are two alternatives to keypad input on mobile devices. Although more natural and intuitive, they're both recognition-based approaches, where recognition error is an unneglectable issue, which is usually exacerbated by the limited computing capabilities of mobile devices. Research shows multimodal fusion is an effective means of reducing recognition errors [8] and enhancing input efficiency. By fusing multiple modalities, such as speech and keypad input, the text entry on mobile devices may be more efficient with relatively low error rate.

In this paper, we propose a multimodal Chinese text entry technique for mobile devices, which utilizes speech and keypad input. With the technique, users can use "Jianpin", a phonemic input method, with which only pressing the keys corresponding to the initials of the pinyin of each character to be entered, and speak out these characters simultaneously. The speech is recognized as pinyin to sort the keypad input

IUI'08, January 13-16, 2008, Maspalomas, Gran Canaria, Spain

[1] Pinyin is the coding scheme of Chinese character's pronunciation. For example, the pinyin of character "敢" (dare) is "gan", in which "g" is called "initial" corresponding to the consonant in English, and "an" is called "final" corresponding to the vowel in English. Thus, to enter "敢", a user using keypad should press key "4","2" and "6", then select "gan" from "han, gao, gan, han", and then select "敢" from "感, 干, 敢, 赶, 甘, 肝…", all of which pronounced "gan".

candidates so that the desired input is at the top of the candidate list. Thus the selection process is accelerated. The key of a this method is the speech/keypad fusion algorithm, which synchronizes speech input and keypad input and then fuses redundant information from two modalities. Preliminary evaluation shows that this technique is effective.

BACKGROUND

Multimodal input has been explored in [1,3,5,6,10,12]. Recent researches have shown that a system that fuses two or more information sources, which are complementary to each other, can be an effective means of reducing recognition uncertainty, thereby improving robustness [8]. In some applications redundant information in multimodal interaction was useful. Bo [3] demonstrated a system that used the combination of speech and keypad input for mobile devices. The system used speech as the primary input modality and keypad as the confirmation or error correction modality. Speech Dasher [10] used speech and gestures to input text while gesture was used for confirmation or error correction. Speech pen [6] was predictive handwriting to reduce the burden of manual handwriting of the Japanese language by ambient multimodal recognition. Kaiser [5] utilized co-referenced, co-temporal handwriting and speech and could recognize out-of-vocabulary (OOV) terms. Wang and Ao [12,1] tried to correct handwriting recognition errors by speech.

Research showed that about 97% of computer users in China use pinyin or some variations of it for daily input [2]. This is probably because pinyin is in the users' preexistent knowledge and users can use it with little learning. Meanwhile, mobile device users are familiar with the keypads on mobile devices. As a result, the combination of Jianpin input (a simplified Pinyin input method) on keypad and ambiguous speech input, which is used in our proposed method, would be promising.

USAGE SCENARIO

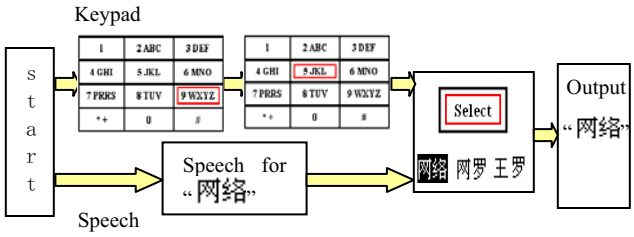

Fig.1 Usage scenario for inputting word "网络"

With T9 on Nokia 5200, when the user wants to input a Chinese word "网络" (network) whose Chinese pronunciation is "wang luo", the user first presses key sequence "9264" and then selects "wang" from candidate list "yang wang zang…". Subsequently, the user selects the character "网" from the candidate list with the same pronunciation "王望往网忘亡汪…". As for "络" with the pronunciation "luo", the user can select it from a character list "络上址吧站恋聊页球号速下路…" that may combine with the character "网" to form a frequently-used word. Totally, 11 key presses are needed in this process. With T9 Jianpin features, the user can also use Jianpin to input the word. After pressing key sequence "95" for letters "w.l", the user has to select "w.l" from "z.j w.j y.j y.k x.l w.l z.l x.j y.l x.k w.k z.k". And then the user selects "网络" from "网络,忘了,为了…". 9 key presses are still needed with Jianpin-based T9.

Figure 1 demonstrates how to use our technique to input Chinese word "网络". User can just press "95" keys in the keypad sequentially for the letters "w" and "l" and simultaneously say "wang luo". The speech recognition candidates with similar pronunciation, include "huang luo, wang luo, huang ruo, wang ruo…". After fusion and internal pinyin to character mapping, "网络" is the first input candidate, and user selects it as the input by pressing the select key on keypad. In this process, only 3 key presses are needed.

Our proposed technique is more effective due to the following reasons. Similar to T9, our keypad input utilizes Jianpin that can greatly reduce the number of key presses. Furthermore, the redundant speech is utilized with keypad simultaneously, which can reduce the number of possible candidates of Jianpin. Meanwhile the keypad input can disambiguate the speech recognition and sort the candidates with the determinate information. So the input efficiency can be improved much.

SPEECH/ KEYPAD FUSION

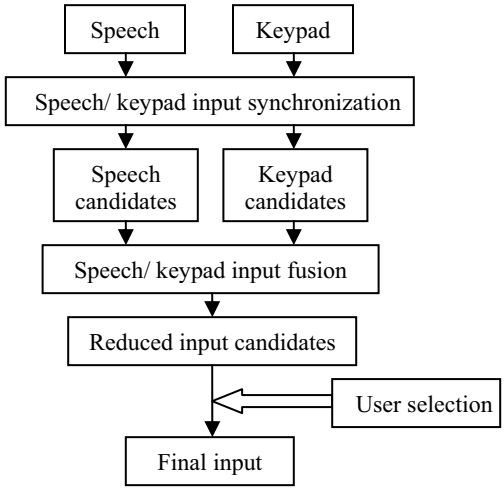

Fig.2 Overview of speech/keypad fusion

The overview of the speech/keypad fusion is shown in Fig. 2. Speech input and keypad input are synchronized at first. Then the speech recognition candidates and the keypad input candidates are fused to get the re-sorted candidate list where the most possible inputs are at the top of the

candidate list. And then the user selects the final input from the candidate list by keypad. In the whole process, input synchronization and input fusion are two important phases.

Speech and Keypad Input Synchronization

The speech input that is temporally close to the keypad input is considered to be the speech corresponding to the keypad input. The confirmation (user's selection of a word as input) indicates the word division of both speech and keypad input.

Figure 3 demonstrates the speech and keypad synchronization process. The horizontal axis represents time. Data input by keypad are shown above the axis. Data input by speech are shown below the axis. Suppose there are M1 key presses before the confirmation for word 1 and there are M2 key presses before the confirmation for word 2. The valid speech time $T(ValidSpeech)$ is the time period that the system receives and processes user's speech input. It can be defined as the time between

$$\max\{T(PreConfirmation), T(FirstKeyPress) - Interval\}$$

and current word confirmation time $T(Confirmation)$, where $T(FirstKeyPress)$ is the time that the first key for the current word is pressed, and $Interval$ is a constant of consumed time for processing speech input by the system. Red lines on the time axis represent the $Interval$ before $T(FirstKeyPress)$. The speech with $Interval$ earlier before the first key press is discarded without processing.

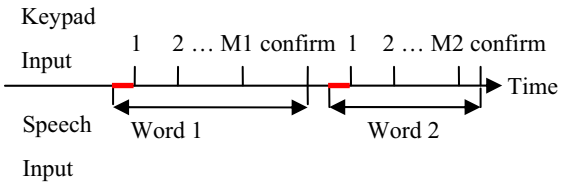

Fig.3 Speech/keypad synchronization

Speech/ Keypad Input Fusion

Keypad input may have several candidates (denoted by K). Speech input should be also recognized as a candidate set (denoted by S). Both K and S are represented in the form of pinyin. The desired input should be selected from K because the keypad input is precise while the speech recognition is more ambiguous. The output is in the form of Chinese word.

We suppose that the keypad candidate set K has M input candidates and the number of the speech recognition candidates in S is N. k_i is the *i-th* candidate of keypad input. s_j is the *j-th* candidate of speech input recognition candidates. The similarity $w(k_i, s_j)$ of k_i and s_j is related to the Levenshtein Distance (edit distance) between k_i and s_j and it is computed as following:

$$w(k_i, s_j) = \frac{1}{edit\_distance(k_i, s_j) + a}$$

, where a is a constant.

Let $p(s_j)$ be the recognition confidence of speech candidate s_j. The probability of choosing keypad input candidate k_i as the input given s_j is $p(k_i \mid s_j)$ that .is proportional to the speech recognition probability $p(s_j)$ and the similarity between k_i and s_j.

$$p(k_i \mid s_j) = p(s_j) \times \frac{w(k_i, s_j)}{Z'(j)}$$

where $i \in \{1, 2...M\}$ and $j \in \{1, 2...N\}$.

$Z'(j)$ ensures that given speech input candidate s_j, the summation of all keypad input candidates probability $p(k_i \mid s_j)$ equals to the speech input candidate probability $p(s_j)$.

$$Z'(j) = \sum_{i=1}^{M} w(k_i, s_j)$$

where $j \in \{1, 2...N\}$.

The probability of choosing keypad candidate k_i can be calculated by the following formula. It is related to the usage frequency of word k_i and the probability of choosing k_i as input given the speech input candidates.

$$p(k_i) = \frac{1}{Z''}\left(b \times freq(k_i) + \sum_{j=1}^{N} p(k_i \mid s_j)\right)$$

where $i \in \{1, 2...M\}$ and b is the coefficient. Z'' is the factor to make the k_i distribution sums to one. Z'' is defined as follows.

$$Z'' = \sum_{i=1}^{M} p(k_i)$$

The keypad input candidates K are sorted descendingly by the probability $p(k_i)$. The desired input is often among the first few keypad candidates with high probabilities.

PRELIMINARY EVALUATION

We implemented the prototype on a virtual mobile device which was simulated on the desktop computer. We used Microsoft speech SDK5.1 as the speech recognizer. A Wacom touching screen was utilized as the keypad input and output device. A head-worn microphone was used as the speech input device.

We informally evaluated the technique with 4 graduate students in our team who were familiar with keypad input on mobile devices. Each session included training the speech recognizer, performing Chinese text entry tasks and a debrief session. The first task was to input 50 words in Chinese with T9 on Nokia 5200. The second task was to input the same words with the proposed multimodal input technique.

Our participants accomplished the tasks with little assistance. We got positive feedback on the multimodal input technique in spite of some issues about speech such as the security problems. All participants thought that the technique was straightforward and easy to use. In particular, they thought the method was efficient and they liked the idea of using speech and keypad simultaneously to improve the Chinese text input efficiency. They thought the technique would be useful for Chinese text entry in several appropriate situations such as at home.

Figure 4 compares the number of key presses between multimodal input technique and T9 input technique for each user. The total number of key presses with T9 was obviously larger than the multimodal input method. And the participants have different key press numbers for the same word set due to the fact that each participant has his or her own inputting habit.

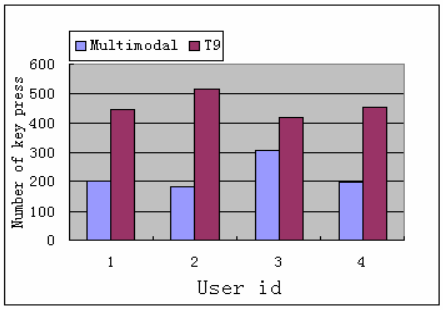

Fig.4 Comparison of key press number for inputting Chinese words between multimodal input technique and T9 input technique

CONCLUSION AND FUTURE WORK

We propose a multimodal Chinese text entry technique for mobile devices.The technique could improve the Chinese text entry efficiency by utilizing the mutual disambiguation of speech and keypad to reduce key presses. Elementary evaluation shows that users appreciate this method and it is efficient.

Although currently the proposed technique utilizes a simplified phonemic input method on keypad, the multimodal fusion algorithm is applicable to a structure-based input method on keypad. In fact, the structure-based input by keypad and speech method will bring greater gain because of the complementary information from two modalities. In addition, sentence input by utilizing a language model would be tested. Meanwhile, we will try to find ways to input text more fluidly with less requirement of users' confirmation.

ACKNOWLEDGEMENT

This research is supported by the National Fundamental Research Project of China (973 Project) (2002CB312103), the National Natural Science Foundation of China under Grant No. 60503054 and No. 60603073, and the National High Technology Development Program of China under Grant No. 2007AA01Z158.

REFERENCES

1. Ao, X., Wang, X.G., Tian, F., Dai, G.Z. and Wang, H.A. Crossmodal error dorrection of continuous handwriting recognition by speech. In *Proc. IUI 2007*. ACM Press (2007), 243-250.
2. Chen, Y., *Chinese Language Processing*, Shanghai Education publishing company, China, 1997.
3. Hsu, B.J., Mahajan, M. and Acero, A. Multimodal Text Entry on Mobile Devices. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) 2005*
4. Hsu, S.C. A flexible Chinese character input scheme. In *Proc. UIST 1991.* ACM Press (1991), 195-200.
5. Kaiser, E.C. Using redundant speech and handwriting for learning new vocabulary and understanding abbreviations. In *Proc ICMI 2006.* ACM Press (2006), 347-356
6. Kurihara, K., Goto, M., Ogata, J. and Igarashi, T. Speech Pen: Predictive Handwriting based on Ambient Multimodal Recognition. In *Proc. CHI 2006.* ACM Press (2006), 851-860.
7. Lin, M., and Sears, A. Graphics matter: a case study of mobile phone keypad design for Chinese input. *Ext. Abstracts 2005*. ACM Press (2005), 1593-1596
8. Oviatt, S. Ten Myths of Multimodal Interaction, *Communications of the ACM*, 42,9(1999), 74 –81
9. T9. http://www.nuance.com/t9/textinput/
10. Vertanen, K. *Efficient Computer Interfaces Using Continuous Gestures, Language Models, and Speech.* M.Phil Thesis, University of Cambridge, 2004
11. Wang, J.T., Zhai, S.M. and Su, H. Chinese input with keyboard and eye-tracking –an anatomical study. In *Proc. CHI 2001.* ACM Press (2001), 349-356.
12. Wang, X.G., Li, J.F., Ao, X., Wang, G. and Dai, G.Z. Multimodal Error Correction for Continuous Handwriting Recognition in Pen-based user Interfaces. In *Proc IUI 2006*. ACM Press (2006), 324-326
13. Zhai, S.M., Kristensson, P.O. and Smith, B.A. In search of effective text input interfaces for off the desktop computing. *Interacting with Computers.* 17, 3 (2005), 229-250.