

# Will Input Style Affect Mandarin Short Messages in Mobile Device?: a Wizard of Oz Study

Ying Liu

Nokia Research Center, Beijing  
Building no.2, No.5, Dong Huan Zhong  
Lu, Beijing Economic & Technological  
Development Area, Beijing 100176,  
China

Ying.y.liu@nokia.com

Li Jiang

Institute of Psychology, CAS  
Jia no.4, Da Tun Lu, Chaoyang  
District, Beijing 100101, China

jiangli@psych.ac.cn

Xinxing Yang

Nokia Research Center, Beijing  
Building no.2, No.5, Dong Huan Zhong  
Lu, Beijing Economic & Technological  
Development Area, Beijing 100176,  
China

Xinxing.yang@nokia.com

## ABSTRACT

Speech input is a natural text entry method for handheld devices that are used in different contexts. We conducted an experiment to understand effects of input (speaking) style (phrasal vs. sentence input) on Chinese text entry rates and user satisfaction with other two variables: recognition rate (50%, 70% and 90%) and message length (10 vs. 20 characters). Wizard of Oz was applied in the experiment due to lack of a working prototype. The results indicated sentence input was better on text entry rates and preferred by end users rather than phrasal input when recognition rate was high. Recognition rate and message length affected both user performance and satisfaction. Further task analysis on the Mandarin message dictation process indicated that error correction took the most percentage of task completion time, followed by candidate selection and speaking phase. Finally, design guidelines on Mandarin dictation application were discussed.

## Categories and Subject Descriptors

H.1.2 [Information Systems]: User/Machine Systems – *human factors, software psychology.*

## General Terms

Performance, Experimentation, Human Factors

## Keywords

Mandarin, Dictation, Input style, Wizard of Oz.

## 1. INTRODUCTION

The short message service (SMS) was widely accepted by mobile phone users in mainland China and had a tremendous growth in terms of both user penetration rate and user amounts in recent years.

SMS requires intensive user interaction for text entry purposes. Entering Chinese characters is not easy in mobile devices.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Conference '04, Month 1–2, 2004, City, State, Country.  
Copyright 2004 ACM 1-58113-000-0/00/0004...\$5.00.

Chinese is a logossyllable language and has tens of thousands of characters. An intermediate encoding layer is necessary to enable Chinese text entry with keyboards designed for Latin languages, which naturally require training [2, 3]. People are still lack of natural ways of entering Chinese text with mobile devices. In that sense, speech input provides a good option for users who are not able to learn to use the coding systems. Moreover, speech input enables text entry with both hands free, which is critical for mobile devices that are used in various contexts.

Speech recognition technologies confront special challenges with Chinese. First, most Chinese characters are homophonic. Tones are thus very important for people to understand the meaning of utterances. Second, spoken Chinese comprises ten dialect groups and Mandarin is the one with the most speakers in China [7]. Third, there are many different tones for different Chinese dialects and most of the time they are semantically meaningful. For example, Mandarin has five tones. Those challenges make the Chinese speech recognition technology even more complex than for other languages.

Mandarin message dictation applications based on speech recognition enables only character input [1]. Users can enter short messages by speaking to the phone character by character. After a character is uttered, the applications will recognize it and provide recognition options in the form of pinyin marks. Users can select the correct pinyin and Chinese character. However, the discontinuous input style impaired usability of the application. People were expecting advanced and continuous input styles. However, there is still more than one option for continuous Mandarin speech recognition; the recognition can be based on either phrase or sentence. Usually a Chinese phrase includes 2 to 4 characters and has a comparatively complete meaning.

On the other hand, there is no study to explore the advantage and disadvantage of the different speech input styles. Thus we designed the study to understand user performance and preference of different input styles (phrase based or continuous speaking styles). Due to lack of working prototypes, we applied the Wizard of Oz method (WOz) in the study. The method is usually applied in the early phase of system design to gain an understanding of the user's mental model and evaluate interaction design options, especially when there is no working prototype available. Past studies also indicated that text length and recognition rate affect usability of speech recognition systems [6]. Thus we study the effects of input style together with recognition rate and text length.

In the rest of the paper, we first introduce objectives and methods of the study, followed by the results on user performance and subjective evaluations. Finally, we will discuss the research results and present conclusions.

## 2. EXPERIMENT

### 2.1 Objectives

We want to answer following questions with the study:

- What is the effect of input style on text entry rate and satisfaction together with the other two factors of recognition rate and text length?
- What are the interaction effects of the three factors on text entry rate and satisfaction?
- Will the effects of recognition rate and text length be comparable to past results?
- What design implications can we get from the study results on Mandarin text dictation systems?

### 2.2 Method

#### 2.2.1 Design

The experiment was a 3\*3\*2 within-subject design with three independent variables of input style, recognition rate and length of messages. The three input styles were phrase based input with 4 options, phrase based input with 7 options and the sentence input. Three levels of recognition rates were taken into account: 50%, 70% and 90%. Users needed to input short messages of two different lengths: 10 characters or 20 characters. Thus, at least 18 short messages need to be inputted to cover all the testing cases.



**Figure 1 The input styles (from left to right: phrase input with 4 options, phrase input with 7 options and the sentence input)**

Time for entering each short message was recorded and the participants were also asked to give a score to each input case with a 5-point Likert scale questionnaire.

#### 2.2.2 Participants

12 users, half male and half female, took part in the study. All use SMS and the Chinese pinyin method every day. Although all were familiar with the Sybmian S60 user interface, they are novice with the speech input. All were right handed. Each received a gift after the experiment.

#### 2.2.3 Tasks and Materials

The tasks for the participants were to copy 18 short messages by speaking to a mobile phone one by one in pre-defined ways. All messages were shown to the participants during the input process. Before each dictation case, the participants were instructed on the

input styles to be used with the next message. The outcome messages of each dictation were also pre-defined to indicate the recognition rate as if there was a working speech recognition engine behind. The participants were also instructed to revise errors in the outcome messages. The Chinese pinyin input method based on the 12-key keypad was used in the error correction process.

All SMS used in the study were selected from a true short message corpus. All characters in the messages belong to the top 500 most often used Chinese characters. To indicate the different recognition rates in the “recognition” results, the wrongly recognized characters would randomly appear in the message. Table 1 shows examples of the short messages and their presentation form indicating different input styles.

**Table 1 SMS examples and the means of indicating input style**

Input Styles	Message example
Phrase	我们__现在__正在__教室__上课
Sentence	衷心希望大家每天快乐

#### 2.2.4 Apparatus

Nokia E50, with a software program designed specifically for the experiment, was used in the study. Results on task completion time were automatically logged by the program for data analysis.

#### 2.2.5 Procedure

Each participant took part in the experiment individually with one researcher in a quiet lab. The researcher first introduced the study objectives and emphasized that the study focused on the interactions instead of the participants. Then the researcher explained how to use the dictation application. A trial session followed, with which the participant could practice until s/he thought s/he could start the testing sessions. The experiment included two sessions: one for 10-character messages and the other for 20-character messages. Testing order was counter balanced to avoid transfer effect. An Infinite Latin-square experiment technique was applied to offset effects caused by testing orders of input style and recognition rate.

When entering a message, a participant needed to speak to the mobile phone in the pre-defined way and made sure the message was exactly the same as instructed. After entering a message, participants gave a score on the task difficulty with a 5-point Likert scale questionnaire. The process would repeat until the participant finished entering all of the 18 messages.

## 3. RESULTS

We present the results in the following two sessions depending on their measurements: text entry rates and subjective evaluation results. In each session, descriptive results are first presented, followed by statistical testing results.

### 3.1 Text Entry Rates

Figure 2 shows the descriptive input speed results. Recognition rate, input style and message length all affected text entry rate according to the table. The text entry rate increased as the recognition rate increased. Sentence input were better on text entry rate while there is no difference between the two phrasal

input methods. Text entry rates of dictating long messages were higher than those of dictating short messages.

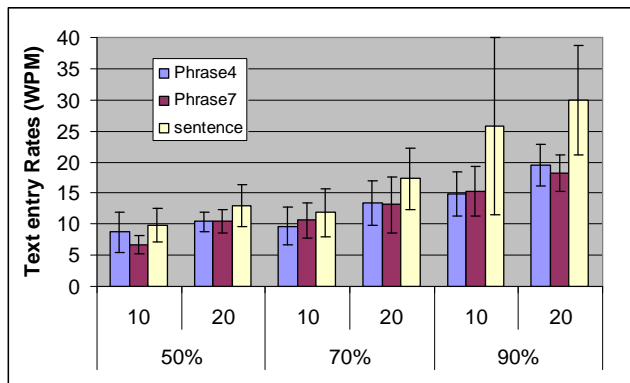


Figure 2 Text entry rates in different conditions

Table 2 shows the ANOVA test results. Main effects of recognition rate ( $F(2, 22) = 87.884, P < .001$ ), input style ( $F(2, 22) = 27.088, P < .001$ ) and message length ( $F(1, 11) = 27.053, P < .001$ ) are all significant. The text entry rates were higher when the recognition rate was higher; users enter longer message; and sentence input was used. The interaction effect between recognition rate and input style was significant ( $F(4, 44) = 46.88, p < .001$ ). When the recognition rate were 50% and 70%, the text entry rates with different input styles were marginally significant at the 0.05 level (50%:  $F(2, 22) = 5.517, p = 0.005$ ; 70%:  $F(2, 22) = 3.525, p = 0.035$ ). When the recognition rate is 90%, text entry rates of the three input styles were significantly different at the 0.001 level ( $F(2, 22) = 16.782, p < 0.001$ ). The trend is clearly shown in Figure 2.

Table 2 ANOVA test results on text entry rates

Source	F	p
Recognition Rate	87.884**	<0.001
Input Style	27.088**	<0.001
Message Length	27.053**	<0.001
Recognition Rate * Input Style	7.494**	<0.001
Recognition Rate * Message Length	0.249	.780
Input Style * Message Length	0.281	.755
Recognition Rate * Input Style * Message Length	0.342	.850

### 3.2 Subjective Scores

Results of subjective scores are shown in Figure 3. The participants gave 4 or more to tasks whose recognition rate was 90%. However, when the recognition rates were lower, people gave higher scores to the tasks of dictating 20-character messages. The differences on subjective scores among the three input styles were not consistent or clear in pattern.

Table 3 shows the ANOVA test results, which indicates that the main effects of recognition rate ( $F(2, 22) = 135.644, P < .001$ ) and SMS length ( $F(2, 22) = 38.047, P < .001$ ) were significant. Recognition rate had significant interaction with SMS length ( $F(2, 22) = 9.683, P < .001$ ) and input style ( $F(4, 44) = 3.196, P < .05$ ).

No other significant effect was found. When the recognition rates were 50% and 70%, participants gave similar scores to the three input styles. When the recognition rates increased to 90%, they gave marginally higher scores to sentence input ( $F(2, 22) = 2.546, p = 0.086$ ). When the recognition rates were 70% and 50%, users gave significantly higher scores to tasks of entering 20 characters (50%:  $F(1, 11) = 35.609, p < 0.001$ ; 70%:  $F(1, 11) = 30.075, p < 0.001$ ). When the recognition rate reaches 90%, users inclined to give similar scores to both types of tasks ( $F(1, 11) = 1, p = 0.321$ ).

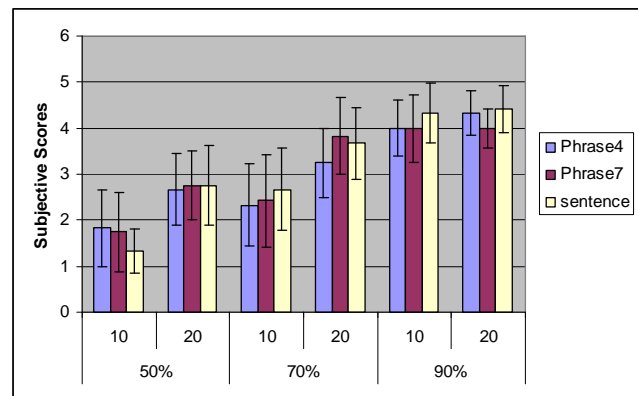


Figure 3 Subjective Scores (the higher the better)

Table 3 ANOVA test results on subjective scores

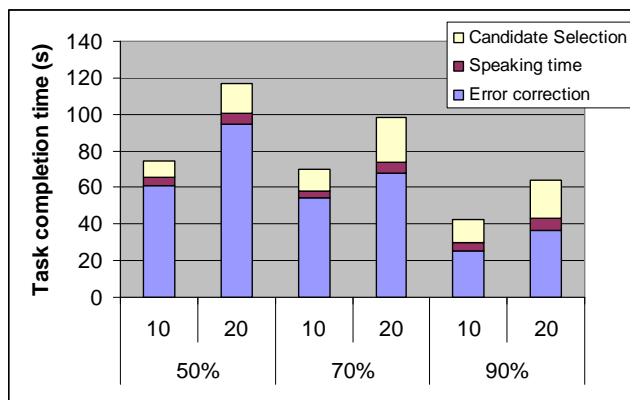
Source	F	p
Recognition Rate	135.644**	<0.001
Input Style	0.933	0.408
Message Length	38.047**	<0.001
Recognition Rate * Input Style	3.196*	.022
Recognition Rate * Message Length	9.683*	.001
Input Style * Message Length	0.376	.691
Recognition Rate * Input Style * Message Length	1.883	.130

## 4. DISCUSSIONS AND CONCLUSIONS

The results showed that text length, recognition rate and input style all affected text entry rate of Mandarin dictation. The text entry rates for Mandarin dictation range from several WPM to about 30WPM, which are comparable with the speeds of other methods [2, 3, 4].

We tried to explain the results based on a task analysis of the dictation process. The phrasal dictation task includes three sub-processes: speaking, error correction and candidate selection from phrase options, while the sentence dictation consists only the first two phases. We recorded and computed participants' speaking time when they uttered a whole sentence. The average speaking times for the 20-character messages and the 10-character messages were respectively 6.20 seconds ( $SD = 0.377$ ) and 4.17 seconds ( $SD = 0.175$ ) and the input speeds were respectively 3.22 and 2.4 syllable per second. The speaking time only took about 4% to 6% of the total task completion time. Thus we can assume that the speaking time in different recognition rates were the same.

Moreover, with the same recognition rate, time used to correct errors in messages of the same length could also be assumed as the same. Thus, when the message length and the recognition rate were the same, the only difference between the phrasal input and the sentence input was that the latter doesn't include the candidate selection process. If we minus the time for sentence input from the time for phrasal input, the results would be the time for the candidate selection process. Furthermore, we can minus the speaking time and the candidate selection time from the total time for phrasal input to get the time used for error correction. Figure 4 presents time costs for each sub-phase with the phrasal dictation with 4 options. According to the results, the error correction process took the most percentage of the total task completion time; followed by candidate selection and the speaking process.



**Figure 4 Time cost of each sub-phase for phrase input with 4 options**

First, the sentence input was significantly quicker than the phrasal inputs because the former saved time of the candidate selection process, which took on average 23.07% of the total task completion time. Moreover, error correction took the most percentage of the total task completion time and was the most critical process affecting user performance. Thus, higher recognition rate resulted in higher text entry rate. The results also indicated when the recognition rate reached 90%, the input speed of sentence input was much better than that of phrasal input whereas it was just marginally better when recognition rates were 50% and 70%. This was because when the recognition rate was low, error correction took the most user efforts and it offset the advantages of sentence input on candidate selection and speaking speed. However, when the recognition rate increased to 90%, user efforts and time on error correction decreased and advantages of sentence input on candidate selection and speaking speed were comparatively enlarged.

The results also indicated that the input speed of entering 20-character messages was quicker than that of entering 10-character messages. The result partly supported theoretical models provided by Price and Sears [6]. One of the key assumptions was that the input speed of text dictation would increase as message length became longer. This might be because Mandarin speaking speed of long material was faster than that of short materials. Moreover, participants had more practice with longer messages, which might result in an improvement in performance.

Subjective scores are usually not consistent with objective measures for speech recognition [5]. We found input style doesn't

solely affect the subjective scores like the other two factors do. However, input style interactively affected the subjective scores with recognition rate. When the recognition rate was 90%, participants preferred the sentence input. When the recognition rates were 50% and 70%, the participants didn't consistently show their preference. This was probably because when the recognition rate was low, people still paid more attention on the error correction process instead of speaking style. Message length also interactively affected the subjective scores with recognition rate. When the recognition rates were low at 50% and 70%, the participants gave significantly higher scores to the task of entering 20-character messages than 10-character ones. When the recognition rate was 90%, the subjective scores on both types of tasks were similar. This was probably because when the participants gave the scores, they still can remember the errors occurred in the 10-character messages. However, it was quite difficult to remember the errors which happened in the 20-character messages because the capacity of short memory was only 5 to 9 items. However, when the recognition rate was 90%, the errors were few and easy to remember in both cases.

Current error correction with keyboard text entry methods are not proper for message dictation application since they costs higher percentage of time and require interactions with hands. If we refer to the daily oral communications between people, repeating or further explain are common. Can we apply the same approach? For a Mandarin text dictation system, continuous input might be a better choice since it can increase user performance and satisfaction on the same recognition level.

## 5. REFERENCES

- [1] Alhonen, J., Yang, C., Ding, G., Liu, Y., Olsen, J., Wang, X. and Yang, X. 2007. Mandarin short message dictation on Symbian series 60 mobile phones, In Proceedings of Mobility, 431-438.
- [2] Lin, M. and Sears, A. 2007. Constructing Chinese characters: keypad design for mobile phones, Behaviour & Information Technology, 26 (2), 2007, 165 – 178.
- [3] Liu, Y. and Raiha, K-J. 2008. RotaTxt: Chinese Pinyin input with a rotator, In proceeding of Mobile HCI, 2008, 225-233.
- [4] Liu, Y. and Wang, Q. 2007. Chinese Pinyin phrasal input on mobile phone: usability and developing trends, In proceedings of Mobility, 2007, 540-546.
- [5] Melto, A., Turunen, M., Kainulainen, A., Hakulinen, J., Heimonen, T. and Antila, V. 2008. Evaluation of Predictive Text and Speech Inputs in a Multimodal Mobile Route Guidance Application, In Proceedings of Mobile HCI, 2008, 355-358.
- [6] Price, K.J. and Sears, A. 2005. Speech-based text entry for mobile handheld devices: an analysis of efficacy and error correction techniques for server-based solutions, International Journal of Human-Computer Interaction, 19(3), 279-304.
- [7] Wang, R. H., National Performance Assessment of Speech Recognition Systems for Chinese. *Proceedings of Oriented COCOSA Workshop*, 1999.