Spoken Correction for Chinese Text Entry

Bo-June (Paul) Hsu and James Glass

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, MA 02139, USA
{bohsu,glass}@mit.edu

Abstract. With an average of 17 Chinese characters per phonetic syllable, correcting conversion errors with current phonetic input method editors (IMEs) is often painstaking and time consuming. We explore the application of spoken character description as a correction interface for Chinese text entry, in part motivated by the common practice of describing Chinese characters in names for self-introductions. In this work, we analyze typical character descriptions, extend a commercial IME with a spoken correction interface, and evaluate the resulting system in a user study. Preliminary results suggest that although correcting IME conversion errors with spoken character descriptions may not be more effective than traditional techniques for everyone, nearly all users see the potential benefit of such a system and would recommend it to friends.

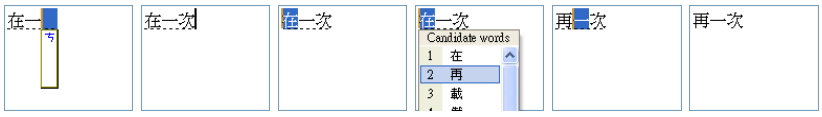
1 Introduction

The number of Chinese-speaking Internet users has quadrupled over the past 5 years to over 132 million today [1]. With more than 120 million instant messaging users and 60 million bloggers projected in China alone by the end of 2006, efficient Chinese text entry plays an ever increasing role in improving the overall user experience for Chinese speakers [2, 3].

Unlike text entry in English, the individual keys on the standard keyboard do not map directly to Chinese characters. Instead, an input method editor (IME) transcribes a sequence of keystrokes into characters that best satisfy the specified constraints. Phonetic IMEs are a popular category of Chinese IMEs that interpret the keystrokes as the pronunciations of the input characters. However, in Traditional Chinese, more than a dozen homonym characters commonly share a single pronunciation. Thus, the IME often leverages a language model (LM) to select the character sequence that maximizes the sentence likelihood [4].

The process of converting phonetic input into the corresponding characters is known as pinyin-to-character, phoneme-to-character, or syllable-to-character conversion [4–6]. Popular phonetic alphabets include zhuyin (注音), also known as bopomofo (ㄅㄆㄇㄈ), and pinyin (拼音). Recent advances in phoneme-to-character conversion have improved the character conversion accuracy to above 95% on newspaper articles [6]. However, the accuracy is reduced on text with mismatched writing styles and is significantly lower on out-of-vocabulary words in the LM. Consequently, efficient text entry requires an effective correction mechanism for users to change the incorrect homonyms to the desired characters.

The Microsoft New Phonetic IME (MSIME) (微軟新注音輸入法) [7] is a popular IME for Traditional Chinese input. To correct a conversion error when using the MSIME, the user first moves the cursor to the incorrect character and then selects the desired character from a candidate list of homonyms with matching pronunciations, as illustrated in Fig. 1. For errors far from the current cursor position, navigating to the target position can be tedious. Since some pronunciations have more than 200 matching characters, the candidate list is often divided into multiple pages. While the desired character often appears within the first page and can be selected with a single keystroke, visually finding the correct character can at times be painstaking given that characters are rendered with a small font and sometimes differ only by their radicals.



(a) (b) (c) (d) (e) (f)

Fig. 1. Illustration of steps involved in correcting the character 在 in 在一次 to 再. After the phonetic sequence is entered in zhuyin (a,b), the user first highlights the conversion error (c). Next, the user selects the desired character from the drop-down candidate list (d) and commits the correction (e). Once all characters in the IME composition window have been corrected, the user commits the composition (f).

In an e-mail survey conducted with 50 Chinese typists, 40% reported skipping past the target character accidentally more than 5% of the time when scanning the candidate list. Due to the frustrating nature of current correction interfaces, 56% admitted that they sometimes do not correct conversion errors, especially in informal text conversations with close friends. With intelligent IMEs that learn from the words and phrases entered by the user [7, 8], leaving conversion errors uncorrected further reinforces the errors and increases the likelihood of the system making similar errors in the future.

In this work, we explore the use of a novel spoken correction approach to address some of the shortcomings in current correction interfaces. Specifically, leveraging users' familiarity with describing the characters in their names when making self-introductions, we support spoken correction via usage, structure, radical, or semantics description of the desired character. For example, to correct the IME composition 在一次, we can say the phrase 再見的再 to specify the desired character 再 from its usage 再見.

In the following sections, we first provide additional background on Chinese text entry and discuss related work. Next, we compute various statistics involving Chinese homonyms and analyze how users disambiguate among them using character descriptions. We then describe the design and implementation of the spoken correction interface and evaluate the system through a user study. Finally, we discuss observations from the user study and areas for future work.

2 Background

2.1 Chinese Text Entry

Popular IMEs for Chinese text entry generally can be categorized as editors that input characters by either compositional structure or pronunciation. While IMEs based on character structures, such as Changjie (倉頡), Boshiamy (嘸蝦米), and Wubi (五筆), often allow for fast entry rates with infrequent conversion errors, they typically require users to learn a set of decomposition rules that take time to master. On the other hand, phonetic IMEs using phonetic alphabets, such as New Phonetic (新注音) and Natural (自然), require minimal learning for most users, as they are taught phonetic spelling in school. Although phonetic methods generally do not involve more keystrokes than structural methods initially, it incurs more conversion errors due to the large number of homonyms per syllable pronunciation. With each correction requiring a visual search for the desired character and additional keystroke for navigation and target character selection, the correction of even a small percentage of characters can account for a significant portion of the overall entry time. Thus, the overall character entry rate of experienced users of phonetic IMEs is typically lower than those using structural input.

2.2 Related Work

Tsai et al. [9] applied spoken descriptions of characters to help resolve homonym ambiguities in Chinese names for a directory assistance application. In addition to generating character usage descriptions from automatically extracted words, phrases, and names, a list of character descriptions for the most common last names was manually collected. With 60,000 descriptions for 4,615 characters, the character description recognizer achieved a success rate of 54.6% at identifying the target character.

In this work, we apply the approach of using character descriptions for disambiguating among homonyms as a correction interface for Chinese text entry using IMEs. We observe that in addition to describing characters by usage phrase (e.g. 再見的再), descriptions using character radical (女字旁的她), compositional structure (土川圳), and character semantics (女生的她) are also fairly typical. In addition, since these descriptions include the target character at the end, the position of the desired character within the current IME composition can often be unambiguously inferred from the character description. Furthermore, because the pronunciations of the characters in the uncommitted IME composition are known, we can limit the recognizer grammar to only accept descriptions for characters with those pronunciations, reducing the grammar perplexity.

Leveraging these observations, we have extended the commercial MSIME with the capability for users to correct errors in the conversion using spoken character descriptions. Preliminary results from user studies suggest that with additional refinements and improvements to recognition accuracy, spoken correction using character descriptions has the potential to improve the correction experience for a significant group of Chinese typists.

3 Analysis

3.1 Homonym Statistics

Due to the obscurity of many characters and the continuous introduction of new characters, the number of Chinese characters varies significantly depending on the particular dictionary or computer character encoding. The CNS11643 standard, for example, defines over 48,000 characters, although many are unpronounceable and the average person only uses around 5,000 characters [10]. The distinction between traditional (繁體) and simplified (簡體) Chinese introduces further complications as many character sets include characters from both styles. In this work, we will only consider the set of characters that can be phonetically entered via the MSIME.

We gathered two text corpora for frequency analysis and system evaluation. The first corpus, *CNA2000*, consists of newswire articles from the Central News Agency of Taiwan in the year 2000 [11]. Specifically, we considered only the headline and core news content for the analysis. For the second corpus, *Blogs*, we extracted text excerpts from 10,000 RSS feeds of randomly selected blogs from a popular blogging website in Taiwan. For both corpora, we segmented the content at punctuations, symbols, and other non-Chinese characters and discarded segments containing character outside our character set. Although blogs better match the informal style of most text entry scenarios, they are also more likely to contain conversion errors that the writer neglected to correct. For simplicity, we will treat both corpora as containing the correct reference text.

To gain insight into the homonym problem in Chinese, we computed, in Table 1, various statistics relating characters to their pronunciations, specified with and without tone. Although there are only 16.8 characters per pronunciation on average, the number of homonym characters with the same pronunciation averaged over the character set is over 38. In the worst but not infrequent case, the candidate list for the pinyin *yi4* has over 207 items. Fortunately, through the application of the language model to order the characters in the candidate list, more than 96% and 95% of the target characters appear on the first page, when correcting conversion errors in a simulated entry of the text from a random subset of the *CNA2000* and *Blogs* datasets, respectively.

Table 1. Statistics on the pronunciations of the 19,991 characters in the character set.

(average / max)	With Tone	Without Tone
# Pronunciations	1387	408
# Characters per Pronunciation	16.8 / 207	54.4 / 383
# Homonyms per Character	38.2 / 206	101.4 / 382
Average Rank of Target Character		
CNA2000	3.0	6.4
Blogs	3.0	6.2

3.2 Character Description

To better understand how character descriptions disambiguate among homonym characters, we asked 30 people to describe the characters in their Chinese name. In a separate study with 10 participants, we requested descriptions for 50 randomly selected characters from among the 250 most frequently confused characters by the IME, displayed next to the incorrect homonyms. Most of the 587 character descriptions collected can be classified into one of the description types listed in Table 2, where we also provided analogous English examples.

Table 2. Types of character descriptions with typical templates, Chinese examples, and approximately analogous examples in English. The target character is in bold.

Description	Typical Template	Example	Approximate English Analogy
Usage	[*usage phrase*]的[*char*]	希望的**希**	**lead** as in lead paint
Structure	[*composition*][*char*]	人白**伯**	**rainbow**, rain plus bow
Radical	[*radical name*]的[*char*]	草字頭的**蔡**	**dialog** with the Greek root log
Semantics	[*meaning*]的[*char*]	數字的**一**	**red** as in the color
Strokes	Character-dependent	三橫一豎**王**	**H** with 2 vertical and 1 horizontal strokes
Compound	Speaker-dependent	微笑的**微**	psych as in psychology
	[*char*] usually omitted	加草字頭 (**薇**)	with an extra E at the end (**psyche**)

When describing by usage, the description is generally a word phrase, idiom, or proper name, consistently in the form [*usage phrase*] 的('s) [*target character*]. While most structural descriptions specify the character by its subcomponents, a few users describe some characters by removing components from more easily describable characters. For example, the character 念 can be described as 唸書的唸, 沒有口字旁. Furthermore, when the desired character differs from the incorrect character by a single component, it is often natural to base the description on the current character. Thus, to change 啊 to 阿, one might say 沒有口的啊 (啊 without 口).

Character descriptions by radical generally can be derived from the radical name and a few simple templates. However, some of the 214 radicals have common aliases, especially when appearing in an alternate form or in a particular position within the character. For example, both 拿 (take) and 打 (hit) share the radical 手 (hand), which can be described using the standard template 手部的[拿,打]. However, because the radical 手 appears in an alternate form in the character 打, 提手旁的打 is another popular description for 打.

Some characters, such as 她 (she) and 九 (nine) are most commonly associated with their semantics, rather than their usages, structure, or radical. For these, special character-dependent descriptions are often used, such as 女生的她 (female's she) and 數字的九 (number's nine). Although descriptions using strokes are also character-dependent, they are specific and do not vary across speakers.

In Table 3, we summarize the observed occurrences of each description type for characters from last names (Last), first names (First), and the most frequently

confused characters (Confused). Overall, descriptions using usage dominate all other description types, except when describing last names. Since the characters in last names differ significantly in distribution from the characters in first names [12], it is not surprising that their description type distributions are also different. However, in addition to the dependency on the specific character, character descriptions also depend on the context. For example, whereas most people would describe the last name 許 by its structure 言午許, in the context of a sentence, many would describe the same character by its usage 許多的許 instead.

Table 3. Occurrences of each character description type from user studies.

Description	Last	First	Confused
Usage	8	53	400
Structure	17	1	8
Radical	3	1	45
Semantics	0	1	25
Strokes	2	1	0
Compound	1	0	2
Others	0	0	19

To measure the variability across speakers in the descriptions of a character, we computed the normalized entropy of the character descriptions for each character spoken by at least 5 participants. For a sample size of N, we define the normalized entropy H_0 as the entropy of the empirical distribution divided by $\ln(N)$. Thus, if all samples have the same value, $H_0 = 0$. If each sample has a different value, $H_0 = 1$. As shown in Table 4, the normalized entropy for most characters are significantly less than 1. Although each character can be described in numerous ways, only a few descriptions are commonly used across users in general. Thus, an effective spoken correction system should not only accommodate the different description types, but also leverage the limited variability of character descriptions across users to improve the speech recognition accuracy.

Table 4. Normalized entropy of character descriptions. For example, of the 7 description instances for the character 集, there are 6 集合的集 and 1 集中的集. Thus, the normalized entropy is $H_0 = -\left(\frac{6}{7}\log\frac{6}{7} + \frac{1}{7}\log\frac{1}{7}\right) / \log 7 = 0.21$.

Norm. Ent.	# Chars	Example
0.0–0.2	9	雨(0.00): 下雨的雨 7
0.2–0.4	7	集(0.21): 集合的集 6, 集中的集 1
0.4–0.6	19	一(0.41): 一二三四的一 5, 一二三的一 4, 一個人的一 1
0.6–0.8	9	不(0.66): 不是的不 3, 不要的不 2, 不可能的不 1, 不好的不 1
0.8–1.0	7	來(0.83): 來去的來 2, 來了的來 1, 來往的來 1, 起來的來 1

4 Design

To investigate the use of spoken character description as a correction interface for Chinese text entry, we extended the MSIME with the capability to correct homonym errors using the usage, structure, radical, or semantics description of the desired character. With the Microsoft Speech API 5.1, we built a custom context-free grammar (CFG) for the spoken character descriptions and used the Microsoft Chinese (Traditional) v6.1 Recognizer as the speech recognizer [13, 14]. The following sections describe the construction of the character description grammar and the design of the correction user interface in more detail.

4.1 Grammar Construction

As observed in Sect. 3.2, character descriptions by usage, compositional structure, and radical generally follow specific templates, allowing for automatic generation. In the user study, the few users who initially deviated from the typical templates showed no difficulty adjusting after being instructed on the expected patterns. Unfortunately, character descriptions by semantics and strokes cannot be generated automatically and required manual data collection. Thus, given the constrained descriptions, we chose to build a language model consisting of a finite state network of data-driven and manually collected character descriptions.

To build usage descriptions, we extracted all word phrases with 2 to 4 characters from the CEDict Chinese-English Dictionary [15], for a total of 23,784 words (辭), idiomatic phrases (成語), and proper names (專有名詞). For each character in each word phrase, we added to the grammar a usage description of the form [*word phrase*]的[*char*].

The Chinese Character Structure Database (漢字構形資料庫) provides the structure information for 7,773 characters in the IME character set [16]. From this, we added simple compositional descriptions of the form [*composition*][*char*]. We leave support for more complex structural descriptions to future work.

Most radicals can be described with a few template expressions. For example, the radical 人 may be described using 人部, 人字部, 人字旁, or 人字邊. However, some radicals also have additional aliases, such as 單人旁 for the radical 人. Thus, to build character descriptions using radicals, we manually identified a set of template expressions appropriate for each radical and supplemented it with a list of radical aliases obtained from the Table of Chinese Radical Names (漢字偏旁名稱表) [17]. Finally, for each character in the IME character set and each corresponding radical name, we added character descriptions of the form [*radical name*]的[*char*] to the grammar.

A single IME composition generally contains only a small subset of the 1,387 pinyin pronunciations. Since users only need to disambiguate among characters whose pronunciation appears within this subset, it suffices to dynamically constrain the language model to only those character descriptions. Thus, when building the CFG, we grouped the character descriptions by the pronunciation of the target character and built a separate rule for each pronunciation. Depending

on the IME composition, we selectively activated the appropriate grammar rules to improve both recognition speed and accuracy.

To further speedup the recognition and reduce the grammar size, we optimized the finite state network by merging all character arcs with the same pronunciation, in effect determining the network at the syllable level. To recover the target character, we encoded it in the grammar as a property tag.

However, when reduced to syllables, not all character descriptions yield unique characters. For example, the phrase 就是的就 actually shares the same phonetic representation as 救世的救 and 舊式的舊. In Table 5, we summarize the statistics on the number of character descriptions with identical pronunciations. Although radicals are often the easiest to describe, they are also the most ambiguous on average. Given the ambiguities associated with even character descriptions, an effective correction user interface will need special handling for this condition.

Table 5. Statistics on character descriptions with the same pronunciation.

(avg / max)	# Descriptions / Pron		Example
	With Tone	Without Tone	
Usage	1.03 / 4	1.11 / 21	就是/救世/舊式 (*jiu4 shi4*)
Structure	1.04 / 6	1.15 / 8	昕/栖/桸/睎/榽/樨 (*xi1*)
Radical	1.34 / 10	1.82 / 16	泄/洩/浥/液/溢/洟/涕/潩/澺/灪 (*yi4*)

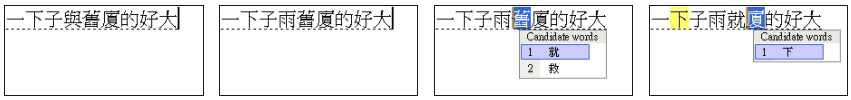
4.2 User Interface Design

To enable transparent switching between the traditional and the new spoken correction interfaces, we reassigned the `Control` key while composing text with the IME to act as a push-to-talk microphone button for the character description speech recognizer. For each correction, the user may choose to select the target character using the arrow keys as before or press the `Control` key to speak a character description. To simplify end-point detection in the initial implementation, we require the microphone button to be depressed while talking.

After the microphone button is depressed, we enable the grammar rules corresponding to user-specified pronunciations in the current IME composition and begin listening for a character description. Upon a successful recognition, we look up the potentially multiple candidate characters matching the description. Typically, the recognized target pronunciation only corresponds to a single syllable position in the IME composition. Thus, if the character description specifies a unique candidate, we immediately replace the character at the matching position with the user-described character, as illustrated in Fig. 2(b). If the character description matches multiple characters, a list containing the candidate characters is displayed at the matching syllable position, as shown in Fig. 2(c). Ideally, the list will be sorted by the language model likelihood. As an approximation, we sort the list according to the ordering of these characters in the original IME

candidate list. Since users are unlikely to describe the currently hypothesized character, it is explicitly moved to the bottom of the list, if included.

Occasionally, the pronunciation corresponds to multiple candidate syllable positions, requiring user intervention prior to making the correction. To allow the user to select the syllable from among these candidate positions, we highlight all candidate positions, display the filtered candidate list containing the matching characters, and restrict the left/right arrow keys to navigate only among these positions, as illustrated in Fig. 2(d). To reduce keystrokes, the candidate list is initially displayed under the position corresponding to the single correction that maximizes the language model likelihood.



(a) (b) (c) (d)

Fig. 2. Illustration of the steps involved in correcting conversion errors. Entering the phonetic sequence for 一下子雨就下的好大 yields the IME hypothesis 一下子與舊廈的好大 (a). The user depresses the microphone button and says 下雨的雨 to correct 與 with 雨. Since this character description uniquely identifies the character 雨 and only corresponds to a single position, the system automatically replaces the error 與 with 雨 (b). To describe 就, the user speaks the usage phrase 就是的就. In this case, because 救世的救 is acoustically identical to this character description, the system shows the candidate list to allow the user to specify the desired character (c). Finally, the user says 下面的下 to replace 廈 with 下. Because two positions in the IME composition contain the syllable *xia4*, the system highlights both candidate positions and selects the one most likely to contain the error (d). In this case, the candidate list appears under 廈 since the first position already contains the specified character 下.

5 User Study

For evaluation, we conducted a user study with 10 students from Taiwan with varying proficiency in Chinese text entry. The study included a questionnaire on the participant's experience with Chinese input, approximately 5 minutes of speech recognition enrollment for acoustic model adaptation, and a collection of 50 spoken character descriptions. Participants were also asked to enter 2 distinct sets of 20 Chinese sentence fragments with the IME, one using traditional keyboard correction, the other using spoken correction with character descriptions. Sentences from both sets were manually selected from the *Blogs* corpus to contain one or more conversion errors. The two sets were randomly alternated for each participant to remove any bias resulting from differences between the two sets.

Table 6 summarizes the results from the study. Overall, the response to spoken correction is positive, with half of the participants expressing interest in

using the system. Through the post-study questionnaire, we learned that of the 5 users expressing a neutral or negative opinion, 3 have memorized deterministic key sequences of common characters for their respective IMEs. Thus, minor improvements to correcting the sporadic errors that they encounter do not justify overcoming the learning curve of a new system and the need to set up a high-quality microphone whenever performing text entry. Interestingly, of these 5 users without definite interest in using the system themselves, 4 would still recommend it to friends. As user J observed, "This system is very useful and convenient for users less familiar with Chinese input... [However], frequent typists will still choose selecting characters [using the keyboard]." Thus, although spoken correction may not be more effective for everyone, nearly all participants saw the potential value of such a system, even with less than 10 minutes of usage.

Table 6. Summary of user study results. Prior to the survey, we asked participants to estimate the average amount of time per week they spend entering Chinese text and indicate the IME they use most frequently. After the study, in which the user had a chance to enter text using both the traditional keyboard correction and spoken correction, we asked users if they would consider using spoken correction in the future and recommend the system to a friend.

User	A	B	C	D	E	F	G	H	I	J
Use Spoken Correction	Y	Y	Y	Y	Y	M	M	M	M	N
Recommend to Friends	Y	Y	Y	Y	M	M	Y	Y	Y	Y
Usage/Week (hr)	1	2	2	2	7	1	3	4	6	2
Typical IME	NP	NP	NP	G	NP	CJ	HI	P	G	NP

NP: New Phonetic 新注音 P: Phonetic (舊)注音 CJ: Changjie 倉頡
G: GOING Natural 自然 HI: Hanin 漢音 (Mac)
Y: Yes N: No M: Maybe

6 Discussions

One concern with spoken correction is the cognitive load associated with identifying an appropriate description for the target character. Unlike the characters in their names, all users experienced some degree of difficulty describing certain characters, such as 之 (possessive particle), that are not associated with common word phrases and are difficult to describe by radical or structure. However, once a description for a difficult character is suggested, the participants did not encounter any difficulty recalling the description the next time the character is observed a few minutes later.

Many factors contribute to the difficulty of describing characters. As observed in Sect. 3.2, users naturally describe characters by usage in a word phrase. However, this may not always be the most effective approach. Although less natural, it is sometimes easier to identify a character by its compositional structure or

radical. For characters from a word phrase in the target sentence, many users have the false notion that because the IME converted the character incorrectly, using the same word phrase to describe the character will not fix the error. As analyzed in [6], more than a third of conversion errors from a bigram-based IME are due to segmentation errors. Thus, explicitly specifying the segmentation boundary through character descriptions can actually correct many of these errors.

Lastly, in the current design, users generally cannot attribute the cause of misrecognitions to acoustic mismatch or unexpected character description, as they have identical behavior. Using the preliminary grammar constructed for the user study, out-of-grammar character descriptions account for 35% of the total spoken corrections. Of the in-grammar descriptions, 16% contained recognition errors. Thus, to improve the spoken correction system, we need to not only improve the grammar coverage, but also mitigate the effect of recognition errors.

7 Conclusion & Future Work

In this work, we introduced a novel correction interface for Chinese text entry using spoken character descriptions. Specifically, we identified common approaches people use to describe characters and constructed an automatically generated character description grammar from various lexical corpora. Finally, we evaluated a preliminary implementation of the spoken correction interface system through a user study that demonstrates the potential benefit of the spoken interface to a considerable subset of Chinese typists.

As shown in Sect. 3.2, most users describe characters using a small subset from among all potential descriptions. Thus, an effective approach to improving the recognition accuracy is to weigh the different character descriptions by their likelihood of utilization. Furthermore, as observed with difficult-to-describe characters, once users identify a successful description for a character, they tend to reuse the same description again for future instances of the character. This suggests that we can further improve the language model performance by emphasizing previously observed character descriptions.

For future work, in addition to incorporating more data to improve grammar coverage, we would like to explore such language model adaptation techniques to reduce the recognition error rate. We also hope to incorporate various feedback from the user study participants to improve the user interface design. Finally, to reduce the effort in evaluating changes to the system, we plan to simulate user input and measure the overall system performance.

In this paper, we focused on applying spoken character descriptions to Chinese keyboard IMEs. However, the approach generalizes to other East Asian languages, such as Japanese and Korean and even to text entry via handwriting and speech, where there are ambiguities in the resulting text. With the rapid growth in text input on mobile devices, we would also like to study the application of spoken correction to text entry interfaces using the keypad.

Acknowledgement We would like to thank Ingrid Bau for the numerous design validation discussions, scheduling of the user studies, and assistance with data transcription. Further credit goes to Chao Wang for the helpful discussions, feedback on the user study design, and detailed review of the paper drafts. The speech recognition software used in this work is provided by Microsoft Corp. This research was supported in part by the Nokia-MIT collaboration.

References

1. Miniwatts Marketing Group: Internet users by languages. Internet World Statistics Website (Mar. 31, 2006) http://www.internetworldstats.com/stats7.htm
2. iResearch Inc.: The Number of IM Users Will Reach 200 million by 2010 in China. iResearch–China Internet Research Center Website (May 17, 2006) http://english.iresearch.com.cn/instant_messenger/detail_news.asp?id=6938
3. Xinhua News Agency: Chinese bloggers to reach 100 million in 2007. China View Website (May 6, 2006) http://news.xinhuanet.com/english/2006-05/06/content_4513589.htm
4. Gao, J., Goodman, J., Li, M., Lee, K.: Toward a Unified Approach to Statistical Language Modeling for Chinese. ACM Transactions on Asian Language Information Processing, Vol. 1, No. 1 (2002) 3–33
5. Hsu, W., Chen, Y.: On Phoneme-To-Character Conversion Systems in Chinese Processing. Journal of Chinese Institute of Engineers 5 (1999) 573–579
6. Tsai, J., Chiang, T., Hsu, W.: Applying Meaningful Word-Pair Identifier to the Chinese Syllable-to-Word Conversion Problem. In Proc. ROCLING (2004)
7. Microsoft Corp.: 微軟新注音輸入法 2003. Microsoft Website (Jun. 15, 2006) http://www.microsoft.com/taiwan/windowsxp/ime/windowsxp.htm
8. IQ Technology Inc.: Natural Chinese Input 8. IQ Technology Website (Jun. 15, 2006) http://www.iq-t.com/en/PRODUCTS/going_01.asp
9. Tsai, C., Wang, N., Huang, P., Shen, J.: Open Vocabulary Chinese Name Recognition with the Help of Character Description and Syllable Spelling Recognition. In Proc. ICASSP (2005)
10. CNS11643 國家中文標準交換碼. CNS11643中文全字庫 Website (Jun. 15, 2006) http://www.cns11643.gov.tw/web/word.jsp#cns11643
11. Graff, D., Chen, K.: Chinese Gigaword. Linguistic Data Consortium (2003)
12. Tsai, C.: Common Chinese Names. Chih-Hao Tsai's Technology Page Website (Dec. 5, 2005) http://technology.chtsai.org/namefreq/
13. Microsoft Corp.: SAPI 5.1. Microsoft Website (Mar. 3, 2003) http://www.microsoft.com/speech/download/old/sapi5.asp
14. Microsoft Corp.: Install and Train Speech Recognition. Microsoft Office Online Website (Jun. 15, 2006) http://office.microsoft.com/en-us/assistance/HP030844541033.aspx
15. Peterson, E.: CEDICT: Chinese-English Dictionary. On-line Chinese Tools Website (Jun. 15, 2006) http://www.mandarintools.com/cedict.html
16. Institute of Linguistics, Academia Sinica: 漢字構形資料庫. 中研院資訊所 Website (Aug. 15, 2005) http://ckip.iis.sinica.edu.tw/CKIP/tool/toolreg_intro.html#HANZI
17. 漢字偏旁名稱表. 楓雪軒 Website (Nov. 17, 2004) http://www.fxx520.com/Article/ShowArticle.asp?ArticleID=365