Browsing The Chinese Web Pages Using Mandarin Speech

Hsin-min Wang,Yu-hsueh Chou, and Berlin Chen
Institute of Information Science, Academia Sinica,
Taipei 11529, Taiwan, Republic of China
E-mail: whm@iis.sinica.edu.tw

Abstract

A speech interface that allows easy access to information on the WWW has the potential to make the browser more friendly and powerful. This paper presents a working Mandarin speech interface for using unconstrained Mandarin speech to control the WWW browser for conveniently browsing the Chinese Web pages. The interface currently provides speakable commands, bookmarks, and links. The experimental results show that our approach specially considering the characteristics of the Chinese language is very effective, and very high accuracy can be achieved.

1 Introduction

The popularity of the WWW and the huge amount of information available on it have attracted many newcomers who have never used computers before. For these people, instead of using keyboard and mouse, a speech interface that allows easy access to information on the WWW has the potential to make the browser more friendly and powerful. During the past years, the Chinese community has also shown its great enthusiasm for the WWW. We are seeing a sudden increase in traffic and the number of Web pages written in Chinese has increased very fast. Because the Chinese language is not alphabetic and input of Chinese characters into computers is very difficult, for the general Chinese population which had much lower access to computers compared to other countries using alphabetic languages, such a speech interface to access the computer and the internet is actually very highly desired. Accordingly, this paper presents a Mandarin speech interface for using unconstrained Mandarin speech to control the WWW browser for conveniently browsing the Chinese Web pages. The interface currently provides:

(1) speakable commands: Allow users to speak browser control commands, e.g. “回到前一頁” (“back”), “列印” (“print”), etc.
(2) speakable bookmarks: Allow users to go to a favorite page by speaking the user-defined key phrase or the name of that page, e.g. “我的公司” (“My company”) or “中央研究院” (“Academia Sinica”).
(3) speakable links: Allow users to speak any visible link names on the current page and invisible link names on the previously visited pages to go to a link. Here, the speakable links on the previously visited pages enable users to penetrate any link on those pages without the need to exhaustedly kick the back bottom to find out that link.

These functions are actually similar to those of the systems proposed by Charles et al in [1][2] for English and Japanese, except that, in their systems, only visible links on the current page are speakable. However, our approach specially considering the characteristics of the Chinese language is totally different from their approaches.

To enable users to access the Chinese Web pages by voice, there are at least two major issues that need to be solved; problems occurring from converting the Chinese sentence into a phonetic string and from speech recognition. Chinese sentences are composed with strings of characters without blanks to mark words. To make the commands, bookmarks, and links speakable, the first step is to identify the words (i.e., segment the character strings of the sentences into word strings) by matching the sentences with a lexicon, such that the pronunciation can be obtained. To allow users the flexibility to just speak a few words of links instead of to speak the full link names, these words as well as the full link names should be put together to form a dynamic keyword lexicon for speech recognition. It is believed that a very large lexicon with sufficient amount of lexical entries which covers all of the Chinese words is critical for correct word identification. However, such a large lexicon is never existing nor will be composed not only because there are unlimited number of compounds and proper names but also because new words keep on growing very fast every day. The compounds, proper names, and new words, which are frequently used as link names or part of link names, and on the same time are very likely to be used as key terms to represent a specific link by users, will be segmented into short words or characters. This will inevitably degrade the accuracy in speech recognition. In our approach, at later stage after word identification, a post-processor is specially designed to regroup short words or characters into compounds and proper names. To allow users the flexibility to speak naturally without the need to follow a rigid speaking format, a vocabulary-flexible keyword spotter integrated with utterance

verification techniques is adopted as the speech recognition engine.

2 A Brief Analysis of the Chinese Web Pages

In Chinese language, every character is a morpheme with its own meaning and pronounced as a monosyllable while a word is composed of one to several characters (or syllables). The total number of Chinese words is believed to be unknown, but the total number of phonologically allowed syllables in Mandarin Chinese is only about 416 if the tones associated with the syllables are disregarded. The combination of these 416 syllables actually gives unlimited number of words. This monosyllabic structure of the Chinese language actually becomes the key for the proposed approach and makes our approach obviously different from approaches previously proposed for other languages.

Before going to the details of our approach, a brief analysis of the Chinese Web pages is discussed in this section first. The analysis is based on Yam's Taiwan Pop100 most frequently visited Chinese Web pages in December 1997 [3]. All local hyperlinks from these 100 pages are collected, while those links that connect to different Web sites are simply discarded for simplicity. The total number of pages used for analysis is 11,898 while the total number of links on these pages is about 137,278, i.e., the average number of links on each page is 11.54. The number of links on each page is summarized in Table 1. It was found that this number ranges from 0 to more than 100. About 87% of pages contain less than 20 links, but still more than 5% of pages contain more than 50 links. It's very interesting that about 23% of pages do not contain any link while 26% of pages contain only 1 or 2 links. The small number of links on each Chinese page indicates that in addition to the visible links on the current page, the invisible links on the previously visited pages can be handled very well by the keyword spotter, and thus can also be speakable. Moreover, it was found that the length of each link ranges from 1 to more than 30 characters. The analysis is summarized in Table 2. It was found that 4-character links are most frequently used. Though most of links (83.49%) contain 10 or less than 10 characters, about 16.51% of links contain more than 10 characters and some links even contain more than 20 or 30 characters. On average, each link contains 6.9 characters. To allow users the flexibility to speak naturally without any constraint, the speech input is allowed to be any combination of words or characters that comprise the corresponding bookmark or link. Actually, most users won't try to input the full name of a link, especially for those links with long names. Most of time, users only speak one or two to three keywords randomly selected from the full link names and hope the speech interface to response exactly. This is why we choose vocabulary-flexible keyword spotting as the speech recognition engine here, and why in addition to performing word identification to obtain the pronunciation of links, keyword extraction is used to further extract the potential keywords.

Number of links	Number of pages	%	Number of links	Number of pages	%
0	2787	23.42	51~60	271	2.28
1	756	6.35	61~70	98	0.82
2	2329	19.57	71~80	39	0.33
3~10	2562	21.53	81~90	22	0.19
11~20	1948	16.37	91~100	146	1.23
21~30	454	3.82	101~	93	0.78
31~40	235	1.98	Total	11898	100.00
41~50	158	1.33			

Table 1: The number of links on each Chinese Web page.

Length	%	Length	%		%
1	0.82	8	4.04	15	1.22
2	8.25	9	2.86	16~20	4.18
3	10.99	10	3.07	21~25	1.95
4	29.99	11	1.97	26~30	1.01
5	6.67	12	2.02	>30	0.81
6	10.94	13	1.82		4.18
7	5.86	14	1.53	>30	0.81

Table 2: The length of links on the Chinese Web pages.

3 Overview of the Proposed Approach

Figure 1 shows the interaction between the speech interface, browser, and the WWW. To allow anyone to use the speech interface naturally without training, here we use a real-time, speaker-independent, vocabulary-flexible keyword spotter as the speech recognition engine. To solve the problem of dynamic link names on each page, the recognizer uses the vocabulary-independent context-dependent INITIAL/FINAL models [4] to allow recognition of any vocabulary word without training on that specific word.

First of all, all commands are automatically defined as keywords and thus added to the dynamic keyword lexicon. For bookmarks, user-defined key phrases are also defined as keywords and thus added to the dynamic keyword lexicon, while the names of bookmarks must be further processed to allow users to go to a favorite page by speaking the name of that page. Word identification [5] is first applied to obtain the pronunciation according to a general Chinese lexicon [6], then keyword extraction is performed based on the words contained in these bookmarks to extract the potential keywords, and finally these keywords are also added to the dynamic keyword lexicon. While for the new bookmarks and visible link names, exactly the same word identification and keyword extraction procedures can be applied. Note that, the word identification and keyword extraction procedures must be

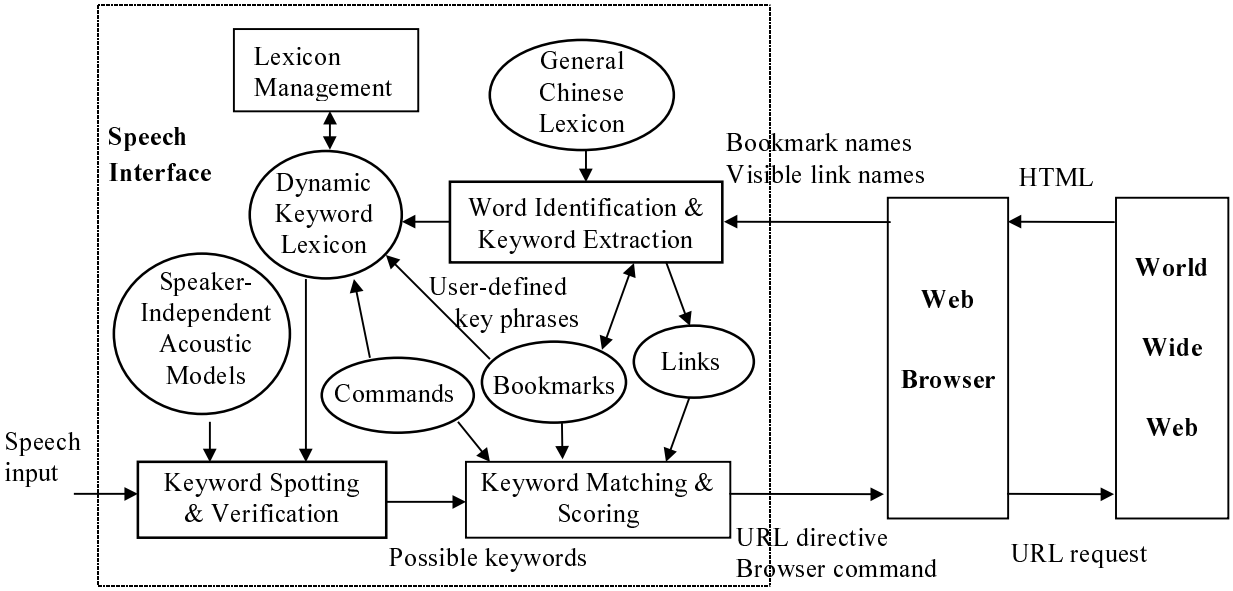

Figure 1: Interaction between the speech interface, browser, and the WWW.

performed on-line in real-time once a new bookmark is added to the bookmark lists or the browser has switched into a new page. To guarantee high keyword spotting accuracy, the size of the dynamic keyword lexicon must be limited. The lexicon management module is thus designed to keep the lexicon size under a pre-determined number based on the first-in-first-out criterion. In this study, the dynamic keyword lexicon is designed to be under 3,000 keywords. Therefore, in addition to the keywords needed for commands, bookmarks, and visible links on the current page, the dynamic keyword lexicon also records the keywords for invisible links of many previously visited pages. The users thus can penetrate any link on the previously visited pages without the need to exhaustedly kick the back bottom to find out that link.

Given a speech input, the vocabulary-flexible keyword spotter thus first generates a set of possible keyword candidates, then the utterance verification mechanism which acts as a post-processor is used to further verify these candidate keywords. Finally, the keyword matching and scoring module chooses the command, bookmark, or link that most fits these keywords, and the speech interface sends a browser control command or a URL directive to the Web browser. Here, the speech input can be any combination of words or characters that comprise the corresponding bookmark or link, and the users need not try to input the full name of a link or a bookmark. Furthermore, the high degree of flexibility of this interface even allows users to input speech like "go to *link* (or *bookmark*) *name*", "I want …" and so on. Some of the key modules will be described in detail in the following sections.

4 Word Identification and Keyword Extraction

For each link or bookmark, word identification is used to obtain the pronunciation as well as the words contained, according to a general Chinese lexicon [6]. The lexicon used in this study consists of a total of 14,052 single character words and another 70,000 words that are composed of from 2 to 4 characters. The basic idea is the words that comprise the corresponding link or bookmark are defined as keywords and added to the dynamic keyword lexicon. As shown in Table 2, about 83% of links contain 10 or less than 10 characters. For those people who would like to speak the full link names to achieve the higher speech recognition accuracy, the full link names that contain 10 or less than 10 characters are also defined as keywords. The number of keywords extracted from 1,190 links of 104 Chinese pages is summarized in Table 3. These 104 Chinese pages were randomly selected from the 11,898 pages previously used for analysis in Section 2. In addition to the keywords for these 1,190 links, a total of 81 browser control commands were considered as keywords in this analysis for realistic considerations. The first row of Table 3 shows the results of the above baseline approach. Since the words in the Chinese lexicon at most contain 4 characters, all keywords with more than 4 characters represent the control commands and the full link names. A large number of single character keywords were found in the dynamic keyword lexicon. It is because new words, compounds and proper names that are very frequently used as link names or part of link names usually are not included in the general lexicon and thus will be segmented into a string of single character words. Such a large number of single character keywords inevitably increase the ambiguity and thus degrade the accuracy in keyword spotting. The keyword extraction approach to further regroup these single character keywords is therefore highly desired. A simple but very effective

approach is to directly regroup these single character keywords into a multi-character keyword. Figure 2 shows an example, where "王大明" is a proper name, and thus the character string "王" "大" "明" will be considered as a 3-character keyword "王大明". When the above simple approach is applied, the second row of Table 3 indicates that the number of single character keywords is significantly reduced from 747 to 180, compared to the first row of Table 3. Most of them were merged into 2-character or 3-character keywords, and some of them were merged into even longer keywords. The most important advantage achieved from the significant reduction of the number of single character keywords is the reduction of ambiguity in keyword spotting. Furthermore, as shown in Table 2, about 30% of links are with 4 characters, while the other 20% are with only 1 to 3 characters. For those links with only 4 or less than 4 characters, it is not necessary to segment them into shorter keywords since most users are likely to input such short full link names. Based on this consideration, the third row of Table 3 shows that the number of single character keywords, 2-character keywords, and 3-character keywords can be slightly reduced while the number of 4-character keywords and other longer keywords keeps unchanged, compared to the first row of Table 3. The total number of keywords is reduced from 2,812 to 2,658. The reduction of the number of short keywords indicates the reduction of ambiguity in keyword spotting. As shown in the last row of Table 3, with all considerations mentioned above, the total number of keywords is reduced from 2,812 to 2,536, with significant reduction in the number of short keywords but slight increase in the number of long keywords. With the above keyword extraction procedure, the ambiguity in keyword spotting can be significantly reduced while the high degree of flexibility can be retained.

Original link name: 程式設計師王大明網頁
Word identification: 程式 設計師 王 大 明 網頁
Keyword extraction: 程式 設計師 王大明 網頁 程式設計師王大明網頁

Figure 2: An example of word processing for the link "The homepage of a programmer, 王大明".

Length of keywords	1	2	3	4	5	6	7	8	9	10	Total
Baseline	747	1085	170	207	110	122	125	121	72	53	2812
Regroup single character words	180	1318	263	249	138	132	130	123	73	55	2661
No word segmentation on short links	665	1014	169	207	110	122	125	121	72	53	2658
Both	165	1209	262	249	138	132	130	123	73	55	2536

Table 3: The number of keywords with different lengths in an example dynamic keyword lexicon.

5 Vocabulary-Flexible Keyword Spotting [4]

Here, the acoustic models for voice recognition are designed to be independent of the keyword vocabulary based on the monosyllabic structure of the Chinese language such that very high degree vocabulary flexibility can be achieved for the time-varying page contents. In Chinese language, every character is a morpheme with its own meaning and pronounced as a monosyllable, while a word is composed of one to several characters (or syllables). The total number of phonologically allowed syllables in Mandarin Chinese is only about 416 if the tones associated with the syllables are disregarded. When the syllables are further decomposed into sub-syllabic units, an even much smaller number of units will be needed. The sub-syllabic units adopted here are INITIAL's and FINAL's, in which the INITIAL is the initial consonant of the syllable while the FINAL is the vowel (or diphthong) part of the syllable but including an optional medial or nasal ending. The combination of these 416 syllables or smaller number of sub-syllabic units actually gives unlimited number of words.

The search framework of keyword spotting is based on a lexical network concatenated with a left filler model and a right filler model, built specifically according to the dynamic keyword lexicon obtained from the word processing procedure as described above. Each arc of the lexical network represents a subword unit, thus the network is able to handle arbitrarily assigned vocabulary set. Here, the syllable is chosen as the subword unit and each syllable is composed of two sub-syllabic models. The left and right filler models that consist of a silence model, a general acoustic filler model, and a set of syllable filler models are used to absorb the non-keyword part of the input utterances. Then, the keyword spotting process is based on a two-pass search strategy. In the first pass, the left and right filler models are decoded left-to-right and right-to-left respectively, with the Viterbi scores stored at every time index. At the same time, a compact syllable lattice is generated from the lexical network based on constraint grammar considering the keyword structure, with the respective lattice node scores evaluated left-to-right including the left filler model scores. The node scores are then taken as the heuristic

scores. In the second pass, a backward time-asynchronous A* search on the right filler model and the lexical network is performed right-to-left, with the aid of the heuristic scores obtained previously. In the A* search process, in every iteration a partial path having the maximal evaluation function is popped, extended, and stored in a stack, and the stack is sorted by the evaluation functions of all these extended partial paths. This process is terminated when the N-best complete paths (candidate keywords) are obtained sequentially.

For each spotted keyword, the sub-syllable level verification is then performed. Here, the sub-syllable level verification score of a specific sub-syllabic unit s is defined as a log likelihood ratio (*LLR*),

$$LLR_s = \log \frac{P(O|\lambda_0)}{P(O|\lambda_1)} \quad (1)$$

where O is the observed speech segment obtained from Viterbi decoding, λ_0 is the corresponding sub-syllabic model, and λ_1 is the competing model. Then, the log likelihood ratio is transformed to a range between 0 and 1 by a Sigmoid function ζ, and the confidence measure (*CM*) of keyword k is represented as,

$$CM_k = \frac{1}{N} \sum_n \zeta(LLR_n) \quad (2)$$

where N is the total number of sub-syllabic units contained in keyword k. Only the keyword candidates with the confidence measure upper than a predetermined threshold τ will be accepted while those with lower confidence measure will be rejected.

6 Keyword Matching and Scoring

Based on the possible keyword candidates obtained from the keyword spotter, the keyword matching and scoring module then calculates the similarity between the speech utterance and commands, bookmarks, or links, and select the one with the highest score as the output. The similarity score is defined as,

$$S_i = \frac{\sum_J CM_j}{W_i} \quad (3)$$

where S_i is the similarity score of the *i*-th sentence (each command, bookmark, or link can be thought as a sentence), W_i is the number of keywords that the *i*-th sentence contains, J is the number of keywords that were matched from the spotted keyword candidates, and CM_j is the confidence measure of the matched keyword *j*. Thus, the command, bookmark, or link with a higher percentage of keywords that are matched from the spotted keyword candidates and on the same time these keywords are with higher confidence measure will have higher priority to be chosen as the output.

In the following, several weighting factors will be introduced to further improve the accuracy of the keyword matching and scoring procedure. The similarity score defined previously in equation (3) is thus modified as,

$$S_i = \frac{\sum_J CM_j}{W_i} \times \frac{\sum_J l_j}{L_i} \times \frac{\sum_J d_j}{D} \times f\left(\frac{\sum_J l_j}{J}\right) \quad (4)$$

where l_j is the length (number of characters) of the matched keyword *j* while L_i is the length of the *i*-th sentence, and d_j is the duration of the match keyword *j* while D is the duration of the speech utterance. The first weighting factor makes the sentence with a higher percentage of matched characters have higher priority to be chosen, while the second weighting factor makes the sentence with a higher percentage of matched duration have higher priority to be chosen. As mentioned above, each keyword can contain 1 to 10 characters. Experience has shown that a long keyword is more reliable than a short one. The basic idea for the third weighting factor is that the sentence with a longer average length of matched keywords should have higher priority to be chosen. Thus, *f(x)* can be any monotonically increasing functions. In this study, *f(x)* is a simple linear function with *f(10)=1* and *f(1)=0.5*.

	Male		Female	
	Number of testers	Number of utterances	Number of testers	Number of utterances
Cmd	8	183	9	130
FullLk	8	304	9	306
CmdIrWd	8	114	9	121
FullLkIrWd	7	134	9	166
KeyLk	7	115	9	189

Table 4: The number of testing utterances.

7 Experiments

This section presents the experimental results. The following experiments were based on 104 Chinese pages previously used for analysis in Section 4. 8 male users and 9 female users were asked to utter the browser control commands (Cmd), full link names (FullLk), control commands with irrelevant words (CmdIrWd), full link names with irrelevant words (FullLkIrWd), and keywords of link names (KeyLk). The number of testing utterances is summarized in Table 4 and an example of possible speech input for a link is shown in Figure 3. The experimental results are summarized in Tables 5 and 6 for male and female users, respectively. The baseline tests used only word identification to obtain the pronunciation and were based on the similarity measure defined in equation (3), the results marked by "With KE" were obtained with the special keyword extraction approach discussed in Section 4 applied, while the results

Original link name: 程式設計師王大明網頁

Possible speech input:
Full link name: 程式設計師王大明網頁
Keywords: 王大明, 王大明網頁, 設計師王大明, ……
Full link name with irrelevant words: 我想去程式設計師王大明網頁

Figure 3: An example of possible speech input for the link "The homepage of a programmer, 王大明".

marked by "With KE&IS" were obtained with the keyword extraction approach applied and based on the similarity measure defined in equation (4). The baseline results are actually not good, with only 63.99% and 57.78% of average accuracy for male and female users respectively. But they can be improved to 73.64% and 71.49% respectively with the special designed keyword extraction approach applied and further improved to 90.47% and 88.59% respectively based on the improved similarity measurement. Both the special designed keyword extraction approach and the improved scoring approach are shown to be very effective in improving the accuracy of using all the 5 different ways to browse the Chinese Web pages. Moreover, it can be found that very high accuracy can be achieved if the input utterances contain only the pure control commands, i.e., the Cmd case in Tables 5 and 6, or full link names without any irrelevant word, i.e, the FullLk case in Tables 5 and 6. On the other hand, slightly worse results were achieved for the other three cases in which more flexible speech input was used. Though, the accuracy reduction is the necessary price paid to achieve the higher degree of flexibility, in any case browsing the Web pages using quasi-natural-language speech is the most natural and attractive and thus highly desired. Therefore, better approaches to further improve the performance of such cases are very important. For example, more complicated acoustic models are believed to be helpful in improving the speech recognition accuracy and thus the accuracy of the system.

8 Conclusion

We have described a working interface that enables users to browse the Chinese Web pages using unconstrained Mandarin speech in a variety of ways, e.g. using speech commands to control the browser, speaking any visible link name on the current page and invisible link name on the previously visited pages to go to a link, and speaking the user-defined key phrase or the name of a bookmarked page to go to a favorite page. Once a new Chinese Web page is visited, word processing is performed on-line in real-time to obtain the pronunciation and potential keywords of links on that page. The size of the dynamic keyword lexicon is always kept under a predetermined number based on a first-in-first-out criterion, the keyword spotter thus can perform very well. The experimental results have shown that very high accuracy can be achieved.

	Baseline	With KE	With KE&IS
Cmd	0.5136	0.6502	0.9289
FullLk	0.8585	0.8815	0.9638
CmdIrWd	0.3859	0.4736	0.8070
FullLkIrWd	0.5597	0.6492	0.8283
KeyLk	0.6086	0.8521	0.8956
Average	0.6399	0.7364	0.9047

Table 5: The testing results for male users.

	Baseline	With KE	With KE&IS
Cmd	0.5923	0.7384	0.9538
FullLk	0.6503	0.7516	0.9215
CmdIrWd	0.4049	0.5371	0.8016
FullLkIrWd	0.6084	0.6807	0.9156
KeyLk	0.5343	0.7830	0.8095
Average	0.5778	0.7149	0.8859

Table 6: The testing results for female users.

References

[1] Charles T. Hemphill and Philip R. Thrift, "Surfing the Web by Voice", *Multimedia95*, pp. 215-222.

[2] Kazuhiro Kondo and Charles T. Hemphill, "Surfing the World Wide Web with Japanese", *ICASSP97*, pp. 1151-1154.

[3] Yam home page, "http://www.yam.org.tw".

[4] B. Chen, H. M. Wang, L. F. Chien, and L. S. Lee, "A*-admissible Key-phrase Spotting with Sub-syllable Level Utterance Verification", *ICSLP98*.

[5] Keh-Jiann Chen and Shing-Huan Liu, "Word Identification for Mandarin Chinese Sentences", *COLING92*, pp. 101-107.

[6] CKIP group, "Analysis of Syntactic Categories for Chinese", *CKIP Technical Report*, No. 93-05, Institute of Information Science, Academia Sinica, Taipei, 1993.